

MCGRAW-HILL ENCYCLOPEDIA OF SCIENCE & TECHNOLOGY



www.MHEST.com



Sorghum — Supramolecular chemistry

Sorghum

Sorghum includes many widely cultivated grasses having a variety of names. Sorghum is known as guinea corn in West Africa, Kafir corn in South Africa, mtama in East Africa, and durra in the Sudan. In India sorghums are called juar, jowar, or cholam; in China, kaoliang; and in the United States, milo. Cultivated sorghums in the United States are classified as a single species, *Sorghum bicolor*, although there are many varieties and hybrids. The two major types of sorghum are the grain, or nonsaccharine, type, cultivated for grain production and to a lesser extent for forage, and the sweet, or saccharine, type, used for forage production and for making syrup and sugar.

Grain Sorghum

Grain sorghum is grown in the United States chiefly in the Southwest and the Great Plains. It is a warmseason crop which withstands heat and moisture stress better than most other crops, but extremely high temperatures and extended drought may reduce yields. Sorghum responds well to optimum growing conditions, fertility, and management to produce large grain yields. It is extensively grown in Texas, Kansas, Nebraska, Oklahoma, Missouri, Colorado, and South Dakota. This grain production is fed to cattle, poultry, swine, and sheep primarily, with less than 2% going into nonagricultural markets such as starch, dextrins, flour, and industrial preparations. Sorghum is considered nearly equal to corn in feed value.

Origins and description. Sorghums originated in the northeastern quadrant of Africa. They have been grown in Africa and Asia for more than 2000 years. Introduction of a sorghum called chicken corn was made on the southern Atlantic coast in Colonial American times, but it was not successfully cultivated. The variety escaped and became a weed. Practically all grain sorghums of importance, until recent years, introduced into the United States were

tall, late maturing, and generally unadapted. Since its introduction into the United States, the crop has been altered in many ways, these changes coming as a result of naturally occurring genetic mutations combined with hybridization and selection work of plant breeders. The rapid expansion in acreage came with the development of widely adapted varieties and later higher-yielding hybrids. The fact that hybrid grain sorghums with high yield potential could be produced with stems that are short enough for harvesting mechanically (**Fig. 1**) made the crop appealing to many farmers.

Grain sorghum is difficult to distinguish from corn in its early growth stages, but at later stages it becomes strikingly different. Sorghum plants may tiller (put out new shoots), producing several headbearing culms from the basal nodes. Secondary culms may also develop from nodal buds along the



Fig. 1. Dwarf grain sorghum hybrids are grown throughout the sorghum belt of the United States because their short stems make them adaptable to mechanical harvesting.



Fig. 2. Heads of grain sorghum. (a) Typical variety. (b) Typical hybrid.

main stem. The inflorescence (head) varies from a dense to a lax panicle, and the spikelets produce perfect flowers that are subject to both self- and crossfertilization. The amount of natural cross-pollination ranges from 25 to 0% but averages about 5%. Mature grain in different varieties varies in size and color from white to cream, red, and brown. Color pigments are located in the pericarp (outer covering) of the grain or in a layer of cells beneath the pericarp called the testa. In some varieties the testa is absent; when the testa is present, however, the seed color is brown or some variation of brown. The endosperm (starch portion of the seed) is either white or yellow. There are endosperm modifications which cause the starch to be sugary or waxy, or to process a higher lysine content. The texture of the endosperm may vary from completely corneous to full floury. Most sorghums are intermediate in endosperm texture.

Grain sorghums are classified into types designated as milo, kafir, feterita, hegari, durra, shallu, kaoliang, and zerazera. This classification is based on morphological rather than cytological differences, since all types (including forage sorghums, broomcorn, and Sudangrass) have 10 pairs of chromosomes and freely cross.

Varieties and hybrids. A combine-height grain sorghum was developed before World War I but was not accepted by farmers or agriculturalists. Acceptance of combine grain sorghums was stimulated by the drought of the 1930s and by a farm labor shortage during World War II. A number of widely adapted productive varieties were developed during this period. Nonetheless, varieties disappeared rapidly when hybrids were introduced in the mid-1950s (Fig. 2).

The commercial production of seed of sorghum hybrids was made possible by the discovery of cytoplasmic male sterility in the early 1950s. This type of sterility, as used in corn and a few other crops, prevents the development of normal pollen grains and makes possible the formation of hybrid seed by crosspollination. Because the flowers of sorghum are perfect, containing both staminate (male) and pistillate (female) parts, the production of commercial quantities of hybrid seed was not possible without a workable male-sterility mechanism. The first hybrid seed in quantity was sold to farmers in 1957, and within a period of less than 5 years hybrids had replaced most of the varieties previously grown. Sorghum hybrids yield at least 20% more than varieties, and concurrent improvements in cultural practices have combined to boost per-acre yield over 50% since their development. The emphasis on research in fertilization, irrigation, insect and disease control, and other areas has provided information to help account for the remarkable yield and production increases over the years.

Through a process of conversion many of the best varieties from around the world are being changed from tall, late, unadapted types to short, early, very useful cultivars. From this program plant breeders are expanding germ-plasm utilization and are developing parents of hybrids for the entire sorghumproducing world. *See* BREEDING (PLANT).

Planting. Grain sorghum seeds are small and should not be planted too deep since sorghum lacks the soil-penetrating ability of corn. A seeding depth of 1 in. (2.5 cm) is acceptable in moist and friable soil, but 2 in. (5 cm) may be necessary under dry soil conditions. The seeds are planted either in rows wide enough for tractor cultivation or in narrower rows if cultivation is not intended. Row planters for corn, cotton, field beans, and sugarbeets may be used when equipped with the proper seed plates. When wheat farming is practiced, much grain sorghum is planted with a grain drill with alternate or various feeder holes plugged to provide the desired row spacing. Soil temperature largely determines when the seed should be planted, assuming soil moisture conditions are adequate. Being tropical in origin, sorghum should not be planted in the spring until soil temperature is 65-70°F (18-22°C) at the planting depth and until there is little chance of subsequent lower temperatures. Dry-land grain sorghum is planted at a rate of 3-5 lb/acre (3.3-5.5 kg/hectare), and the rate is increased up to 10 lb/acre (11 kg/ha)

when planted under more favorable moisture conditions and irrigation.

Cultivation. Good seedbed preparation is essential for full stands and for weed control. Tilling fields improves soil structure in most cases and often aids in warming the soil. A rotary hoe is effective in controlling weeds when the plants are small. Subsequent cultivations are made as needed with the same equipment used for cultivating corn. Minimum tillage is practiced in many areas where control of weeds with chemicals is a part of the technology of sorghum production. When used correctly, herbicides are a boon to sorghum culture, but when used carelessly, disappointment may result. There are several chemical herbicides registered and approved for weed control in sorghum.

Harvesting. Nearly all grain sorghum is harvested standing in the field with a combine (Fig. 1). Harvest begins in southern Texas in early June and slowly proceeds northward. In the central and northern Great Plains, the crop is usually harvested after frost. The grain threshes freely from the head when the seed moisture content is 20-25% or lower. The grain should not contain more than 12% moisture to ensure safe storage after harvest. Grain dryers are often used when the grain at harvest is not dry enough for optimum storage. Proper storage must be maintained until the grain can be marketed. The industry's economic health lies in the ability to provide grain of the right quality in the right quantity at the proper time and place. Frederick R. Miller

Sweet Sorghum

Commonly known as sorgo, sweet sorghum was introduced into North America from China in 1850,



Fig. 3. Sweet sorghum in Oklahoma. The scale indicates feet. 1 ft = 0.3 m. (USDA)

although its ancestry traces back to Egypt. It is an annual, rather drought-resistant crop. The culms are from 2 to 15 ft (0.6 to 4.6 m) tall, and the hard cortical layer, or shell, encloses a sweet, juicy pith that is interspersed with vascular bundles. At each node both a leaf and a lateral bud alternate on opposite sides; the internodes are alternately grooved on one side. Leaves are smooth with glossy or waxy surfaces and have margins with small, sharp, curved teeth. The leaves fold and roll up during drought. The inflorescence is a panicle of varying size having many primary branches with paired ellipsoidal spikelets containing two florets in each fertile sessile spikelet. The plant is self-pollinated. *See* CORTEX (PLANT); PITH.

Seed is planted in cultivated rows and fertilized similarly to corn. Maturity varies between 90 and 125 days. The juice contains about 12% sugar. The main sorghum-syrup-producing area is in the south-central and southeastern United States (**Fig. 3**).

Leonard D. Baver

Diseases

Sorghums are plagued by a variety of diseases that vary in importance from year to year and among locations because of the environment, plant genotypes, cultural practices, variations in pathogens, or the interaction of any of these factors. These diseases may be classified into five general categories: those that rot the seed or injure seedling roots; those that attack the leaves, making the plants less productive; those that attack or destroy the grain in the heads; those that cause root and stalk rots; and those caused by viruses or viruslike organisms.

Fungi causing seed rotting and seedling diseases may be seed-borne or soil-inhabiting and are most destructive after planting, when the soil is cold and wet. Species of *Fusarium*, *Pythium*, *Helmintbosporium*, and *Penicillium* are the most important fungi involved. Damage may be considerably reduced by planting sound hybrid seed treated with an approved fungicide in soil warm enough to ensure prompt germination.

Leaf diseases are caused by three species of bacteria and at least eight species of fungi. Many of these pathogens are favored by high temperatures and humid conditions, but a few are favored by cool, humid conditions. Disease lesions occurring as discolored spots or streaks may coalesce to involve the entire leaf. Rotation, seed treatment, and the use of resistant varieties are recommended control measures.

Three of the four known smuts of sorghum are found in the United States; covered kernel, loose kernel, and head smut. Kernel smuts, while historically important, are now controlled by routine seed treatments. Head smut (**Fig. 4**) destroys the entire head and continues to cause major crop losses. Resistant hybrids are providing control, although new strains of the fungus pathogen require the periodic development of new resistant hybrids.

Sorghum downy mildew (**Fig. 5**) has spread throughout the southern and central sorghumgrowing regions. Losses result when the disease systemically invades the plant. Diseased plants are barren. Fortunately, excellent resistance has been developed and used.

Maize dwarf mosaic, caused by an aphidtransmitted virus, spread throughout the sorghumgrowing regions during the 1970s. Host tolerance reduces losses caused by this prevalent disease. Yellow sorghum stunt, caused by a mycoplasmalike organism and transmitted by leafhoppers, rarely reaches economically significant proportions.

Several diseases of the roots and stalks are of primary importance. Periconia root and crown rot, which caused extensive damage to milo and darso sorghums, is controlled by resistant varieties. *Pythium graminicola* causes a major root rot during periods of frequent rainfall in dryland sorghums. Charcoal rot, most evident as the plant approaches maturity under extreme conditions of heat or drought, causes shredding of the stalks and extensive lodging. Anthracnose, or red rot, develops in susceptible hybrids during wet years; plants lodge at the base of the peduncle. Development of resistant or tolerant hybrids appears to be the only effective method of control of the stalk rots if irrigation is not possible.

Grain mold is a disease of mature grain caused by species of *Fusarium* and *Curvularia*. These fungi



Fig. 4. Sorghum head smut, a disease that completely destroys the normal head and replaces it with masses of smut spores.



Fig. 5. Sorghum downy mildew, a serious fungus disease that systematically invades sorghum plants, causing stripped leaves and barren stalks.

infect at the flowering stage and rot the seed as it matures, particularly during wet weather at harvest time. *See* PLANT PATHOLOGY. Richard A. Frederiksen

Bibliography. W. F. Bennett et al., *Modern Grain Sorghum Production*, 1990; H. Doggett, *Sorghum*, 1970; J. R. Quinby, *Sorghum Improvement and the Genetics of Growth*, 1974; J. S. Wall and W. M. Ross, *Sorghum Production and Utilization*, 1970; R. J. Williams, R. A. Frederiksen, and G. D. Bengston, *Proceedings of the International Sorghum Dis ease Workshop*, International Crops Research Institute for the Semi-Arid Tropics, Hyderabad, 1980; R. J. Williams, R. A. Frederiksen, and J. C. Girard, *Sorghum and Pearl Millet Disease Identification Handbook*, International Crops Research Institute for the Semi-Arid Tropics, Inform. Bull. 2, Hyderabad, 1978.

Sound

The mechanical excitation of an elastic medium. Originally, sound was considered to be only that which is heard. This admitted questions such as whether or not sound was generated by trees falling where no one could hear. A more mechanistic approach avoids these questions and also allows acoustic disturbances too high in frequency (ultrasonic) to be heard or too low (infrasonic) to be classed as extensions of those events that can be heard.

A source of sound undergoes rapid changes of shape, size, or position that disturb adjacent elements of the surrounding medium, causing them to move about their equilibrium positions. These disturbances in turn are transmitted elastically to neighboring elements. This chain of events propagates to larger and larger distances, constituting a wave traveling through the medium. If the wave contains the appropriate range of frequencies and impinges on the ear, it generates the nerve impulses that are perceived as hearing. *See* HEARING (HUMAN).

Acoustic Pressure

A sound wave compresses and dilates the material elements it passes through, generating associated pressure fluctuations. An appropriate sensor (a microphone, for example) placed in the sound field will record a time-varying deviation from the equilibrium pressure found at that point within the fluid. The changing total pressure P measured will vary about the equilibrium pressure P_0 by a small amount called the acoustic pressure, $p = P - P_0$. The SI unit of pressure is the pascal (Pa), equal to 1 newton per square meter (N/m²). Standard atmospheric pressure (14.7 lb/in.2) is approximately 1 bar = 10^6 dyne/cm² = 10^5 Pa. For a typical sound in air, the amplitude of the acoustic pressure may be about 0.1 Pa (one-millionth of an atmosphere); most sounds cause relatively slight perturbations of the total pressure. See MICROPHONE; PRESSURE; PRES-SURE MEASUREMENT; PRESSURE TRANSDUCER; SOUND PRESSURE.

Plane Waves

One of the more basic sound waves is the traveling plane wave. This is a pressure wave progressing through the medium in one direction, say the +x direction, with infinite extent in the *y* and *z* directions. A two-dimensional analog is ocean surf advancing toward a very long, straight, and even beach. The surf looks like a long, corrugated surface advancing uniformly toward the shore but extending transversely in a series of parallel peaks and troughs. A plane wave has acoustic pressure controlled by the onedimensional wave equation in cartesian coordinates, Eq. (1). An appropriate solution to this equation is Eq. (2), with *f* any function differentiable twice with

$$\frac{\partial^2 p}{\partial x^2} = \frac{1}{c^2} \frac{\partial^2 p}{\partial t^2} \tag{1}$$

$$p = f\left(t - \frac{x}{c}\right) \tag{2}$$

respect to *x* or *t*. This solution has the property that the acoustic pressure *p* has a single value for all pairs of *x* and *t* such that the phase (t - x/c) is constant. At some point x_0 and time t_0 , the acoustic pressure *p* has the value $p_0 = f(t_0 - x_0/c)$. As *t* increases from t_0 to $t_0 + \Delta t$, the value p_0 will move from x_0 to a new position $x_0 + \Delta x$, where Δx and Δt are related by Eq. (3). Solving for Δx in terms of Δt gives $\Delta x/\Delta t = c$,

$$(t_0 + \Delta t) - \frac{(x_0 + \Delta x)}{c} = t_0 - \frac{x_0}{c}$$
(3)

so that the specific value p_0 is translated through space with a speed of propagation *c*. Thus, *c* is the speed of sound of the wave. *See* WAVE (PHYSICS); WAVE EQUATION; WAVE MOTION. **Harmonic waves.** A most important plane wave, both conceptually and mathematically, is the smoothly oscillating monofrequency plane wave described by Eq. (4). The amplitude of this wave is *P*.

$$p = P \cos\left[2\pi f\left(t - \frac{x}{c}\right)\right] \tag{4}$$

The phase (argument of the cosine) increases with time, and at a point in space the cosine will pass through one full cycle for each increase in phase of 2π . The period *T* required for each cycle must therefore be such that $2\pi fT = 2\pi$, or T = 1/f, so that f = 1/T can be identified as the frequency of oscillation of the pressure wave. During this period *T*, each portion of the waveform has advanced through a distance $\lambda = cT$, and this distance λ must be the wavelength. This gives the fundamental relation (5) between

$$f = c \tag{5}$$

the frequency, wavelength, and speed of sound in any medium. For example, in air at room temperature the speed of sound is 343 m/s (1125 ft/s). A sound of frequency 1 kHz (1000 cycles per second) will have a wavelength of $\lambda = c/f = 343/1000$ m = 0.34 m (1.1 ft). Lower frequencies will have longer wavelengths: a sound of 100 Hz in air has a wavelength of 3.4 m (11 ft). For comparison, in fresh water at room temperature the speed of sound is 1480 m/s (4856 ft/s), and the wavelength of 1-kHz sound is nearly 1.5 m (5 ft), almost five times greater than the wavelength for the same frequency in air.

Because of many close analogies between sound and electricity, it is often convenient in practice to define the effective pressure amplitude $P_e = P/\sqrt{2}$ in a wave such as that of Eq. (4). Similarly, the frequency is often represented by the angular frequency $\omega = 2\pi f$ and the wavelength expressed reciprocally by the wave number $k = 2\pi/\lambda$. With these definitions, Eq. (4) could be written as Eq. (6), and Eq. (5) as Eq. (7).

$$p = \sqrt{2}P_e \cos\left(\omega t - kx\right) \tag{6}$$

$$\frac{\omega}{k} = c \tag{7}$$

See ALTERNATING-CURRENT CIRCUIT THEORY.

Transient and continuous waves. Monofrequency waves are the building blocks for more complicated waves that are either continuous or transient in time. For example, a sawtooth continuous wave has acoustic pressure which, during each cycle, begins at a positive value P_{max} , decreases uniformly (linearly with time) to a negative value $-P_{\text{max}}$ at the end of the period *T*, and then jumps instantaneously back to the positive value P_{max} , repeating this cycle indefinitely. It is a direct consequence of the wave equation that this waveform can be described equivalently as a Fourier superposition, or summation, of waves $p_1 + p_2 + p_3 + \cdots$, each of the form of Eq. (8), where

$$p_n = \frac{2}{\pi} \frac{P_{\text{max}}}{n} \sin\left[2\pi \eta f\left(t - \frac{x}{c}\right)\right]$$
(8)

the fundamental frequency f = 1/T gives the repetition rate of the waveform and *n* has integer values 1, 2, 3, The waves p_n constitute the overtones of the waveform. In this case the frequencies nf of the overtones are integer multiples of the fundamental frequency f, and the overtones are termed harmonics. Any signal that is nonrepeating or is nonzero only with some limited duration of time can be written as a summation of waves that are not harmonically related or, in the extreme, an integration of waveforms of all frequencies. As an illustration, a very sharp positive pulse of pressure traveling in the x direction and lasting for an infinitesimally short duration of time is represented by the Dirac delta function $\delta(t - x/c)$. This function, which has unbounded value where t = x/c and is zero elsewhere, can be expressed as an integral, as in Eq. (9). These considerations show that

$$\delta\left(t - \frac{x}{c}\right) = 2\int_0^\infty \cos\left[2\pi f\left(t - \frac{x}{c}\right)\right] df \qquad (9)$$

a study of monofrequency sound waves is sufficient to deal with all sound waves, and that the fundamental concepts of frequency and wavelength permeate all aspects of sound. *See* FOURIER SERIES AND TRANS-FORMS; HARMONIC (PERIODIC PHENOMENA); NONSI-NUSOIDAL WAVEFORM; WAVEFORM.

Standing waves. In many situations, sound is generated in an enclosed space that traps the sound within or between boundaries. For example, if there is a boundary that causes the pressure wave given by Eq. (4) to be completely reflected back on itself, then there is an additional wave given by Eq. (10),

$$p' = P \cos\left[2\pi f\left(t + \frac{x}{c}\right)\right] \tag{10}$$

which represents a monofrequency plane wave traveling in the -x direction. This wave combines with the incident wave, resulting in a standing wave with a pressure dependence p_T given by Eq. (11). This kind

$$p_T = p + p' = 2P \cos\left(\frac{2\pi x}{\lambda}\right) \cos\left(2\pi ft\right)$$
 (11)

of wave would be found within a sounding organ pipe and is analogous to a vibrating guitar string. The pressure waveform p_T is zero at positions $x = \lambda/4$, $3\lambda/4$, $5\lambda/4$, These nodes (pressure nulls) occur every half-wavelength in space. Midway between them are antinodes at which the pressure waveform oscillates in time between its maximum and minimum values of $\pm 2P$. More complicated standing waves can be formed from waves traveling in a number of directions. For example, vibrating panels or drum heads support standing waves in two dimensions, and steady tones in rooms can excite threedimensional standing waves. *See* VIBRATION.

Particle Speed and Displacement

As sound passes through a fluid, the small fluid elements are displaced from their equilibrium rest positions by the fluctuating pressure gradients. The motion of a fluid element is described by the particle velocity \vec{u} with which it moves about its equilibrium position. This is related to the pressure by Newton's second law. With the neglect of some small nonlinear and viscous terms, this law can be written as Eq. (12), where ρ_0 is the equilibrium density of

$$\rho_o \frac{\partial \vec{u}}{\partial t} = -\nabla p \tag{12}$$

the fluid and ∇ is the gradient operator. See CALCU-LUS OF VECTORS; FLUID-FLOW PRINCIPLES; NEWTON'S LAWS OF MOTION.

For a one-dimensional plane wave moving in the +x direction, the acoustic pressure p and particle speed u are proportional and related by $p/u = \rho_0 c$. The product $\rho_0 c$ is a basic measure of the elastic properties of the fluid and is called the characteristic impedance. This is an index of the acoustic "hardness" or "softness" of a fluid or solid. (The term characteristic impedance is restricted to the plane-wave value of $\rho_0 c$; more generally, the term specific acoustic impedance is used.) Some representative values of the speed of sound, the density, and the specific acoustic impedance are given in **Table 1**. Significant differences among these quantities for gases, liquids, and solids are evident. *See* ACOUSTIC IMPEDANCE.

Because fluids cannot support shear (except for small effects related to viscosity), the particle velocity of the fluid elements is parallel to the direction of propagation of the sound wave and the motion is longitudinal. In contrast, solids can transmit shearing or bending motion—reeds, strings, drum heads, tuning forks, and chimes can vibrate transversely.

Description of Sound

The characterization of a sound is based primarily on human psychological responses to it. Because of the nature of human perceptions, the correlations between basically subjective evaluations such as loudness, pitch, and timbre and more physical qualities such as energy, frequency, and frequency spectrum are subtle and not necessarily universal.

Intensity, loudness, and the decibel. The strength of a sound wave is described by its intensity *I*. From basic physical principles, the instantaneous rate at which energy is transmitted by a sound wave through unit area is given by the product of acoustic pressure and the component of particle velocity perpendicular to the area. The time average of this quantity is the acoustic intensity, as in Eq. (13). For a

$$I = \frac{1}{t} \int_0^t p u \, dt \tag{13}$$

plane monofrequency traveling wave, this is given by Eq. (14) in the direction of propagation. If all quanti-

$$I = \frac{1}{2} \frac{P^2}{\rho_0 c} = \frac{P_e^2}{\rho_0 c}$$
(14)

ties are expressed in SI units (pressure amplitude or effective pressure amplitude in Pa, speed of sound in m/s, and density in kg/m³), then the intensity will be in watts per square meter (W/m^2).

Because of the way the strength of a sound is perceived, it has become conventional to specify the

7

Gases		Density (ρ_0), k	kg/m ³	Speed of	Speed of sound (c), m/s		
Air		1.21		343			
Oxygen (0°C; 32	2°F)	1.43		317			
Hydrogen (0°C;	32°F)	0.09			1,270		
Liquids		Density (ρ_0),	kg/m ³	Speed of sound (c), m/s			
Water		998		1,481			
Seawater (13°C	; 55°F)	1,026		1,500			
Ethyl alcohol	, <u>,</u>	790		1,150			
Mercury		13,600		1,450			
Glycerin		1,260		1,980			
D (1) (1)		Speed of sound (c), m/s		Specific acoustic impedance $(\rho_0 c)$, N \cdot s/m ³			
Solids	kg/m ³	bar	bulk	bar	bulk		
Aluminum	2,700	5,150	6,300	13.9	17.0		
Brass	8,500	3,500	4,700	29.8	40.0		
Lead	11,300	1,200	2,050	13.6	23.2		
Steel	7,700	5,050	6,100	39.0	47.0		
Glass	2,300	5,200	5,600	12.0	12.9		
Lucite	1,200	1,800	2,650	2.2	3.2		
Concrete	2,600	_	3,100	_	8.0		

SOURCE: After L. E. Kinsler et al., Fundamentals of Acoustics, 3d ed., Wiley, 1982.

intensity of sound in terms of a logarithmic scale with the (dimensionless) unit of the decibel (dB). An individual with unimpaired hearing has a threshold of perception near 10^{-12} W/m² between about 2 and 4 kHz, the frequency range of greatest sensitivity. As the intensity of a sound of fixed frequency is increased, the subjective evaluation of loudness also increases, but not proportionally. Rather, the listener tends to judge that every successive doubling of the acoustic intensity corresponds to the same increase in loudness. This is conveniently expressed by Eq. (15), where the logarithm is to base 10, *I*

$$L_I = 10 \log\left(\frac{I}{I_{\rm ref}}\right) \tag{15}$$

is the intensity of the sound field in W/m^2 , I_{ref} is 10^{-12} W/m², and L_I is the intensity level in dB re 10^{-12} W/m². On this scale, the weakest sounds that can be perceived have an intensity level of 0 dB, normal conversational levels are around 60 dB, and hearing can be damaged if exposed even for short times to levels above about 120 dB. Every doubling of the intensity increases the intensity level by 3 dB. For sounds between about 500 Hz and 4 kHz, the loudness of the sound is doubled if the intensity level increases about 9 dB. This doubling of loudness corresponds to about an eightfold increase in intensity. For sounds lying higher than 4 kHz or lower than 500 Hz, the sensitivity of the ear is appreciably lessened. Sounds at these frequency extremes must have higher threshold intensity levels before they can be perceived, and doubling of the loudness requires smaller changes in the intensity with the result that at higher levels sounds of equal intensities tend to have more similar loudnesses. It is because of this characteristic that reducing the volume of recorded music causes it to sound thin or tinny, lacking both highs and lows of frequency. *See* DECIBEL; LOUDNESS.

Since most sound-measuring equipment detects acoustic pressure rather than intensity, it is convenient to define an equivalent scale in terms of the sound pressure level. Under the assumption that Eq. (14) is valid for most commonly encountered sound fields, a reference effective pressure amplitude $P_{\rm ref} = 20$ micropascals (μ Pa) generates the reference intensity of 10^{-12} W/m² in air (at standard temperature and pressure) and a sound pressure level (SPL) can be defined by Eq. (16), where P_e is the

$$SPL = 20 \log \left(\frac{P_e}{P_{ref}}\right)$$
(16)

effective pressure amplitude in μ Pa. The intensity level and sound-pressure level are usually taken as identical, but this is not always true (these levels may not be equivalent for standing waves, for example). For underwater sounds, the sound pressure level is also expressed by Eq. (16), but the reference effective pressure is defined as 1 μ Pa.

Frequency and pitch. How "high" sound of a particular frequency appears to be is described by the sense of pitch. A few minutes with a frequency generator and a loudspeaker show that pitch is closely related to the frequency. Higher pitch corresponds to higher frequency, with small influences depending on loudness, duration, and the complexity of the waveform. For the pure tones (monofrequency sounds) encountered mainly in the laboratory, pitch and frequency are not found to be proportional. Doubling the frequency less than doubles the pitch. For the more complex waveforms usually encountered, however, the presence of harmonics favors a proportional relationship between pitch and frequency. *See* PITCH.

Consonance and dissonance. Two tones generated together cannot be distinguished from each other if their frequencies are the same. If their frequencies f_1 and (slightly higher) f_2 are nearly but not exactly identical, the ear will perceive a slow beating, hearing a single tone of slowly and regularly varying amplitude. The combination yields an equivalent signal given by Eq. (17), which is heard as a

$$\cos\left(2\pi f_1 t\right) + \cos\left(2\pi f_2 t\right)$$

$$= 2\cos\left(2\pi\frac{f_2 - f_1}{2}t\right)\cos\left(2\pi\frac{f_2 + f_2}{2}t\right) \quad (17)$$

single tone having a frequency that is the average of the frequencies of the two individual tones and an amplitude that varies slowly according to the difference of the two frequencies. As f_2 increases more, the beating will quicken until it first becomes coarse and unpleasant (dissonant) and then resolves into two separate tones of different pitches. With still further increase, a sense of beating and dissonance will reappear, leading into a blending or consonance that then breaks again into beats and dissonance, and the whole cycle of events repeats. These islands of consonance surrounded by beating and dissonance are attained whenever the ratio of the two frequencies becomes that of small integers, $f_2/f_1 = 1/1$, 2/1, 3/2, $4/3, \ldots$ The larger the integers in the ratio, the more subtle the effects become. See BEAT.

Frequency spectrum and timbre. Sounds can be characterized by many subjective terms such as clean, nasal, edgy, brassy, or hollow. Each term attempts to describe the nature of a complex waveform that may be of very short or long duration and that consists of a superposition or combination of a number of pure tones. The sound of a person's whistling or of a flute played softly often has a pure, clean, but somewhat dull quality. These sounds consist mainly of a pure tone with few or no harmonics. As described above, complex repetitive waveforms are made up of a fundamental tone and a number of harmonics whose frequencies are integer multiples of the fundamental frequency. Blown or bowed instruments such as flute, bowed violin, oboe, and trumpet provide good examples. Other sounds are transient, or nonrepetitive, and usually consist of a fundamental plus a number of overtones whose frequencies are not integer multiples of the lowest. Piano, tympani, cymbals, and plucked violin generate these kinds of sounds.

An important factor is the way in which a sound commences. When a chime is struck, there is a clear sharp onset much like a hammer hitting an anvil; the higher overtones are quite short in duration, however, and quickly die out, leaving only a few nearly harmonic lower overtones that give a clear sensation of pitch. Plucking a guitar string near the bridge with a fingernail yields a similar effect. A gong gives a very complex impression because there are many nonconsonant overtones present, dying away at different rates, so that the sound seems to shift in pitch and timbre, and may appear nonmusical or merely noise to some. It is the abundance of harmonics and overtones, the distribution of intensity among them, and how they preferentially die away in time that provide the subjective evaluation of the nature or timbre of the sound. *See* MUSICAL ACOUSTICS.

Propagation of Sound

Plane waves are a considerable simplification of an actual sound field. The sound radiated from a source (such as a loudspeaker, a hand clap, or a voice) must spread outward much like the widening circles from a pebble thrown into a lake.

Spherical waves. A simple model of this more realistic case is a spherical source vibrating uniformly in all directions with a single frequency of motion. The sound field must be spherically symmetric with an amplitude that decreases with increasing distance r from the source, and the fluid elements must have particle velocities that are directed radially. A solution of the wave equation with spherical symmetry describing this kind of motion for an outgoing monofrequency traveling wave is given by Eq. (18),

$$p = \frac{A}{r}\cos\left(\omega t - kr\right) \tag{18}$$

where $\omega = 2\pi f$ is the angular frequency and $k = 2\pi/\lambda$ the wave number. The pressure amplitude is A/r and diminishes inversely with distance from the source. If the spatially dependent pressure amplitude is defined by Eq. (19), then the intensity is still

$$P(r) = \frac{A}{r} \tag{19}$$

given by Eq. (14), but with *P* and *P_e* interpreted as A/r and $(A/r)/\sqrt{2}$, respectively. Thus, the intensity falls off with distance as $1/r^2$. This is consistent with conservation of energy. The acoustic power sent through a sphere of radius *r* surrounding the source is the intensity of the wave multiplied by the surface area through which it passes. This yields $4\pi r^2 I = 4\pi A^2/(2\rho_0 c)$, which is independent of *r*.

While the particle velocity for this wave is a relatively complicated function of r, at distances $r \ll \lambda$ the ratio p/u approaches $\rho_0 c$, the same as for a traveling plane wave. Further, in this limit the surfaces of constant phase, for which (t - r/c) has constant value, become more and more planar. Thus, at sufficient distances from the source a spherical wave becomes indistinguishable from a plane wave when viewed over regions of the space whose dimensions are small with respect to r. This asymptotic behavior allows use of the simple plane-wave relationships for many situations.

Directional waves. Not all sources radiate their sound uniformly in all directions. When someone is speaking in an unconfined space, for example an open field, a listener circling the speaker hears the voice most well defined when the speaker is facing the listener. The voice loses definition when the

speaker is facing away from the listener. Higher frequencies tend to be more pronounced in front of the speaker, whereas lower frequencies are perceived more or less uniformly around the speaker.

Dipole source. The fields radiated by sources of more complicated shape and size can be calculated by considering the complicated source as being made up of a collection of small spherical sources, each radiating a pressure wave like that given by Eq. (18), and then adding the pressure fields together. A simple example is the dipole source, consisting of two small sources spaced very closely together and radiating 180° out of phase, so that one is shrinking as the other is expanding, and vice versa. The two sources will nearly cancel because as one is generating a positive pressure, the other is generating a negative pressure. Because the two sources are slightly separated, however, the fields will not exactly cancel. If θ is defined as the angle included between the desired direction in space and the line joining the centers of the two (out-of-phase) sources, and d is the distance between the two sources, then at large distances r away from the pair (Fig. 1a) the total acoustic pressure field



Fig. 1. Pressure field of a dipole source. (a) Coordinates θ and *r* used to describe the field. The distance between the component sources is *d*. (*b*) Nodal surfaces of the field when the distance *d* is sufficiently large with respect to the wavelength λ .

is given approximately by Eq. (20). This equation is

$$p = \frac{Akd}{r}(\cos\theta)\,\sin\left(\omega t - kr\right) \tag{20}$$

valid when the wavelength of sound is much larger than the distance separating the sources (kd < 1) and r is much larger than d. The amplitude of the pressure P(r) is given by $(Akd/r)|\cos \theta|$. In any fixed direction, the amplitude of the pressure falls off as 1/r, but at a fixed distance the pressure amplitude is modulated in direction by $|\cos \theta|$. In the direction $\theta = 0$ (or $\theta = \pi$ radians), the two sources lie one behind the other and the cancellation is least complete. If $\theta = \pi/2$, then the two sources lie side by side and the cancellation is total. There is a nodal surface, defined by the plane $\theta = \pi/2$, exactly midway between the two sources and perpendicular to the line joining them.

If the distance *d* is not significantly less than the wavelength λ , then Eq. (20) is not accurate, and a more complicated expression, Eq. (21), must be

$$p = 2\frac{A}{r}\sin\left(\frac{1}{2}kd\cos\theta\right)\sin\left(\omega t - kr\right)$$
(21)

used. There is still a nodal surface at $\theta = \pi/2$, but if *kd* is large enough there may be additional nodal surfaces, each a cone concentric with the line joining the sources, in directions given by angles θ_n satisfying *kd* cos $\theta_n = 2n\pi$ with n = 1, 2, ... (Fig. 1*b*).

Generalized source. Generalizing from the above shows that the pressure field radiated by a source of arbitrary shape will have an amplitude that at large distance can be written as the product of a function of r and a function of direction, as in Eq. (22).

$$P(r, \theta, \phi) = P_{ax}(r)H(\theta, \phi)$$
(22)

Here, $P_{ax}(r)$ falls off as 1/r, and $H(\theta, \phi)$ is a function only of direction and has a maximum value of 1. At fixed distance r, the pressure amplitude has maximum value given by $P_{ax}(r)$, termed the axial pressure amplitude. The direction for which $H(\theta, \phi)$ has its maximum magnitude of 1 defines the acoustic axis of the source. The acoustic axis may be a plane or a series of surfaces of constant θ or ϕ , but often the acoustic axis is a single line. The ratio of the intensity found at distance r in some arbitrary direction given by θ and ϕ to the value found at the same distance on the acoustic axis is simply $H^2(\theta, \phi)$, and the expression of this as an intensity level, called the beam pattern $b(\theta, \phi)$, is given by Eq. (23). The decibel value

$$b(\theta, \phi) = 20 \log H(\theta, \phi) \tag{23}$$

of the beam pattern is the reduction in intensity level in that direction compared to the value on the acoustic axis (at the same distance from the source).

Baffled piston source. A loudspeaker or a hole in a partition can be represented as a baffled piston. This is a flat, circular source of radius *a* mounted on a large, flat, rigid surface, commonly called a baffle. All portions of the source move together with uniform particle velocity normal to the baffle. It is convenient to



Fig. 2. Coordinates r and θ used to describe the pressure field of a baffled piston. The radius of the piston is a.

measure the distance *r* from the center of the piston and the direction θ with respect to the perpendicular to the piston and baffle (**Fig. 2**). The acoustic axis is the line $\theta = 0$, and the principal sound radiated by the piston propagates along this axis, falling off as 1/rat larger distances and spreading within a conical nodal surface surrounding the axis. This spreading beam of sound is the major lobe of the radiated pressure field. The angle θ_1 defining this nodal surface is given by $ka \sin \theta_1 = 3.83$. This surface will exist only if the frequency is high enough and the wavelength consequently small enough that ka is greater than 3.83. For sufficiently high frequencies, a number of nodal surfaces may exist, defined by angles θ_n satisfying Eq. (24). [Here, the quantities j_{1n} are the values of

$$ka \sin \theta_n = j_{1n} = 3.83, 7.02, 10.17, \dots$$
 (24)

the argument of the first-order Bessel function J_1 that reduce it to zero: $J_1(j_{1n}) = 0$.] Between these nodal surfaces will be minor lobes of lower-intensity levels. Each of these minor lobes will be funnel- or bowlshaped, surrounding the major lobe as a figure of revolution about the $\theta = 0$ acoustic axis. Some examples for a piston with different values of ka are shown in Fig. 3. The first side lobe, when it exists, has an intensity level about 18 dB less than the major lobe. Subjectively, sound heard by someone standing in the direction of the first side lobe is about one-fourth as loud as that heard in the direction of the major lobe at the same distance. For a given piston, it is apparent that higher frequencies (shorter wavelengths) will yield pressure fields that have narrower major lobes and more nodal surfaces and minor lobes. This is one reason why higher-frequency sounds are often radiated by smaller loudspeakers (tweeters) in high-fidelity systems. The smaller loudspeakers radiate the shorter wavelength sounds more uniformly over larger angles, helping to distribute the energy throughout the space. These small speakers are not useful for lower frequencies, however, because they cannot move enough air to generate sufficiently intense low-frequency sound. See BESSEL FUNCTIONS; LOUDSPEAKER.

In addition to the geometrically simple far field found at large distances from a source, there will also often be a rather complicated near field in the immediate vicinity of the source. For the piston source, the transition between near and far fields occurs around



Fig. 3. Directivity patterns for a rigid circular piston mounted on an infinite baffle for different values of ka, where a is the radius of the piston and k is the wave number. (a) ka = 3. (b) ka = 5. (c) ka = 10.

a distance r_1 estimated by $r_1/a \sim a/\lambda$. See DIRECTIV-ITY.

Diffraction. It is possible to hear but not see around the corner of a tall building. However, higherfrequency sound (with shorter wavelength) tends to bend or "spill" less around edges and corners than does sound of lower frequency. The ability of a wave to spread out after traveling through an opening and to bend around obstacles is termed diffraction. This is why it is often difficult to shield a listener from an undesired source of noise, like blocking aircraft or traffic noise from nearby residences. Simply erecting a brick or concrete wall between source and receiver is often an insufficient remedy, because the sounds may diffract around the top of the wall and reach the listeners with sufficient intensity to be distracting or bothersome. The effectiveness of a barrier in blocking sound can be roughly estimated by calculation of an insertion loss. For a long solid wall (brick or concrete) of uniform height the insertion loss can be estimated from Eq. (25), where r is the direct-

$$\Delta L_I = 19 + 10 \log\left(\frac{D-r}{\lambda}\right) \tag{25}$$

path distance (the distance in a straight line from source to receiver), *D* is the new path length from the source just over the top of the wall to the receiver, and ΔL_I is the difference between the intensity level that would be experienced if there were no barrier and the intensity level with the barrier. The equation is accurate when $(D - r)/\lambda \ll 1$. See ACOUSTIC NOISE; DIFFRACTION.

Speed of sound. In fluids, the acoustic pressure and the density variation are related by Eq. (26),

$$p = \Re s \tag{26}$$

where the condensation *s* is $(\rho - \rho_0)/\rho_0$ and \mathcal{R} is a constant of proportionality. This is a threedimensional Hooke's law stating that stress and strain are proportional. The propagation of sound requires such rapid density changes that there is not enough time for thermal energy to flow from compressed (higher-temperature) elements to decompressed (lower-temperature) elements. Because there is little heat flow, the process is nearly adiabatic and \mathcal{R} is the adiabatic bulk modulus (rather than the more familiar isothermal modulus relevant to constant-temperature processes). A derivation of the wave equation for a fluid reveals that the speed of sound is given by Eq. (27). Equations (26) and (27) can be combined to give Eq. (28).

$$c^2 = \frac{\mathcal{R}}{\rho_0} \tag{27}$$

$$p = \rho_0 c^2 s \tag{28}$$

See ADIABATIC PROCESS; HOOKE'S LAW.

In most fluids, the equation of state is not known, and the speed of sound must be measured directly or inferred from experimental determinations of the bulk modulus. In distilled water the measured speed of sound in meters per second can be represented by a simplified equation (29), where T is the temper-

$$c = 1403 + 4.9T - 0.05T^2 + 0.16\,\% \tag{29}$$

ature in degrees Celsius and \mathcal{R} is the equilibrium pressure in bars. For most gases at reasonable pressures and temperatures, however, the equation of state is known to be very close to the perfect gas law, Eq. (30), where *R* is the specific gas constant

$$\mathcal{P}_0 = \rho_0 R T_K \tag{30}$$

and T_K the absolute temperature. (Absolute temperature T_K in kelvin and standard temperature T in degrees Celsius are related by $T_K = T + 273$.) Further, for sound in a gas, the instantaneous pressure and density are related by Eq. (31), where γ is the ratio

$$\frac{\mathcal{P}}{\mathcal{P}_0} = \left(\frac{\rho}{\rho_0}\right)^{\gamma} \tag{31}$$

of specific heats. For air, $\gamma = 1.4$. For small pressure fluctuations, Eq. (31) can be approximated by Eq. (32). Comparing this with Eq. (28), and using the

$$\frac{\varphi}{P_0} \doteq 1 + \gamma s \tag{32}$$

perfect gas law, Eq. (30), gives a theoretical expression for the speed of sound in an ideal gas, Eq. (33).

С

$$=\sqrt{\gamma RT_K} \tag{33}$$

Thus, the speed of sound in an ideal gas depends on temperature, increasing as the square root of the absolute temperature. Because many applications of Eq. (33) are for gases around room temperature, it is convenient to reexpress the formula in terms of degrees Celsius and the speed of sound c_0 at $T = 0^{\circ}$ C (32°F), as in Eq. (34). This equation can be used to

$$c = c_0 \sqrt{1 + \frac{T}{273}}$$
(34)

calculate the speed of sound for a perfect gas at any temperature in terms of the speed of sound in that gas at 0°C (32°F). In air, for example, $c_0 = 331$ m/s (1086 ft/s), and at 20°C (68°F) the speed of sound is calculated to be 343 m/s (1125 ft/s). *See* GAS; HEAT CAPACITY; UNDERWATER SOUND.

In solids, the sound speed depends on the transverse extent of the solid with respect to the wave. If the solid is rodlike, with transverse dimensions less than the acoustic wavelength, the solid can swell and contract transversely with relative ease, which tends to slow the wave down. For this case, the speed of sound is given by Eq. (35) where \mathscr{Y} is Young's mod-

$$c = \sqrt{\frac{\mathscr{Y}}{\rho_0}} \tag{35}$$

ulus. If, in the other extreme, the solid has transverse dimensions much larger than the acoustic wavelength, then a longitudinal wave will travel with a greater speed of sound given by Eq. (36), where \mathcal{G} is

$$c = \sqrt{\frac{\mathscr{B} + 4\mathscr{G}/3}{\rho_0}} \tag{36}$$

the shear modulus. These two different propagation speeds are often called bar and bulk sound speeds, respectively. *See* ELASTICITY; YOUNG'S MODULUS.

Ray and refraction. Since the speed of sound varies with the local temperature (and pressure, in other than perfect gases), the speed of a sound wave can be a function of position. Different portions of a sound wave may travel with different speeds of sound.

If a highly directional piston projects a sound wave horizontally into the air, and if the value of ka is large enough, the major lobe of the sound radiated by the piston is confined to directions within a very small angle θ_1 about the (horizontal) acoustic axis. The beam of sound has surfaces of constant phase, each resembling a disk, perpendicular to and propagating out along the acoustic axis with the speed of sound of the air. If, however, the air temperature decreases gradually with height above the ground, then the upper portions of the beam travel in air with a lesser sound speed than the lower portions. The top of a surface of constant phase gradually lags behind the bottom, and the disk tilts a little so that its normal points slightly upward. Since the direction of propagation is normal to the surface of constant phase, this surface gradually propagates upward, rising above the horizontal as it travels outward away from the piston. Thus, as a consequence of the speed of sound decreasing with altitude, the beam of sound is refracted upward in an arc with one end horizontal at the piston and the other tilting progressively upward as distance from the piston increases.

Each small element of a surface of constant phase traces a line in space, defining a ray along which acoustic energy travels. The sound beam can then be viewed as a ray bundle, like a sheaf of wheat, with the rays distributed over the cross-sectional area of the surface of constant phase (**Fig. 4**). As the major lobe spreads with distance, this area increases and the rays are less densely concentrated. The number of rays per unit area transverse to the propagation path measures the energy density of the sound at that point.



Fig. 4. Representation of an upwardly refracted sound beam as a ray bundle.

It is possible to use the concept of rays to study the propagation of a sound field. The ray paths define the trajectories over which acoustic energy is transported by the traveling wave, and the flux density of the rays measures the intensity to be found at each point in space. This approach, an alternative way to study the propagation of sound, is approximate in nature but has the advantage of being very easy to visualize.

An equation describing refraction can be obtained from the three-dimensional wave equation (37), with *c* now a function of space, by trying a solution of the form of Eq. (38). The pressure amplitude P(x,y,z)

$$\nabla^2 p = \frac{1}{c^2} \frac{\partial^2 p}{\partial t^2} \tag{37}$$

$$p = P(x, y, z) \cos \left[\omega \left(t - \frac{\psi(x, y, z)}{c_0} \right) \right]$$
(38)

depends on location, and c_0 is a convenient constant reference value of sound speed. The quantity ψ has the dimensions of a length, and constant ψ determines a surface of constant phase. The local direction of propagation is perpendicular to this surface and is given by $\nabla \psi$, the gradient of ψ . For the plane traveling wave of Eq. (4), the surfaces, given by x = constant, are planes perpendicular to the x axis. The gradient of x points in the +x direction, the direction of propagation of the phase surface. For the spherical wave of Eq. (6), the surfaces are given by r = constant, and each is a sphere concentric with the source at r = 0. The gradient of r is direct radially outward at each point. Substitution of Eq. (38) into the wave equation and some approximation results in the Eikonal equation (39) for the direction of prop-

$$|\nabla\psi|^2 \doteq \left(\frac{c_0}{c}\right)^2 \tag{39}$$

agation. The Eikonal equation is valid if the fractional changes in P and c in any direction over distances of a wavelength are very small. In general, solving this equation requires integral and differential calculus and results in rather difficult formulas.

For the case of a one-dimensional speed of sound profile c = c(z), where *z* represents altitude, the solution can be simply stated: If direction is defined by an angle θ of elevation ($\theta = 0$ points horizontally and $\theta = \pi/2$ points vertically), then the local direction of propagation of a ray must obey Snell's law, Eq. (40). If the elevation angle of the ray has

$$\frac{\cos\theta}{c} = \frac{\cos\theta_0}{c_0} \tag{40}$$

value θ_0 at an altitude where the speed of sound is c_0 , then as the ray propagates to a new altitude where the speed of sound is c(z), its directional angle will change to a new value $\theta(z)$ according to Eq. (40). This allows the ray to be traced point by point as it propagates through the fluid, each new directional angle at each new point being obtained from Snell's law. If the speed of sound varies linearly with altitude, that is, $c(z) = c_0 - gz$, where the gradient $g = \partial c/\partial z$ of the profile is a constant, then the ray path will be

the arc of a circle whose center lies at the altitude where the speed of sound would vanish. [This would never happen since c(z) will stop decreasing linearly long before this altitude is reached. It is, however, a useful artifice for drawing the trajectory of the ray.] **Figure 5** provides an example of such a ray path. *See* ATMOSPHERIC ACOUSTICS; REFRACTION OF WAVES.

Reflection and transmission. If a sound wave traveling in one fluid strikes a boundary between the first fluid and a second, then there may be reflection and transmission of sound. For most cases, it is sufficient to consider the waves to be planar. The first fluid contains the incident wave of intensity I_i and reflected wave of intensity I_r , the second fluid, from which the sound is reflected, contains the transmitted wave of intensity I_t . The directions of the incident, reflected, and transmitted plane sound waves may be specified by the grazing angles θ_i , θ_r , and θ_t (measured between the respective directions of propagation and the plane of the reflecting surface).

The intensity reflection coefficient I_r/I_i and the intensity transmission coefficient I_t/I_i are found by imposing boundary conditions that require continuity of pressure and continuity of the normal component of the particle velocity at the interface between the two fluids. These boundary conditions then require that the grazing angle of the incident plane wave equals that of the reflected plane wave, Eq. (41). For

$$\theta_i = \theta_r \tag{41}$$

the transmitted wave, the grazing angle θ_t must be given by Eq. (42), where c_1 and c_2 are the values of the

$$\cos \theta_t = \left(\frac{c_2}{c_1}\right) \cos \theta_i \tag{42}$$

speed of sound in the first and second fluids, respectively. Both of these results are consistent with Snell's law. There are two possible situations depending on whether the speeds of sound c_1 and c_2 and the angle of incidence are such that $\cos \theta_t < 1$, which presents no conceptual problem, or $\cos \theta_t \ll 1$, which does.

If $\cos \theta_t < 1$, then the transmitted wave is a plane wave sent into the second fluid. The beam of sound is refracted according to Snell's law, being inclined toward the normal to the boundary if c_2 is less than c_1 and being deflected away from the normal if c_2 exceeds c_1 . The reflected and transmitted intensities are given by Eqs. (43) and (44), where ρ_1 and ρ_2 are

$$\frac{I_r}{I_i} = \left(\frac{\frac{\rho_1 c_2}{\rho_2 c_2} - \frac{\sin \theta_t}{\sin \theta_i}}{\frac{\rho_1 c_1}{\rho_2 c_2} + \frac{\sin \theta_t}{\sin \theta_i}}\right)^2 \tag{43}$$

$$\frac{I_t}{I_i} = \frac{\left(\frac{4\rho_2 c_2}{\rho_1 c_1}\right)}{\left(\frac{\rho_1 c_1}{\rho_2 c_2} + \frac{\sin \theta_t}{\sin \theta_i}\right)^2}$$
(44)

the densities of the first and second fluids, respectively. Clearly, the intensities depend strongly on the angle of incidence. If there is a large mismatch in the



Fig. 5. Ray path for the refraction of sound in a speed-of-sound profile. (a) Speed of sound *c* as a function of altitude *z*. (b) Path of the ray. r = horizontal distance from source; $\theta =$ angle of elevation of the ray.

characteristic acoustic resistances of the two fluids, regardless of which is the larger, the reflection coefficient approaches unity. If $\rho_2 c_2 < \rho_1 c_1$, the second medium appears rigid, and the pressure amplitude at the boundary (the amplitude of the sum of the incident and reflected waves evaluated at the interface) is as great as possible. In the other extreme, $\rho_2 c_2 \ll \rho_1 c_1$, the second medium offers virtually no opposition to the motion of the first, so there is no reactive force and therefore no pressure at the interface. In this case the boundary is called pressure release.

If $\cos \theta_t \ll 1$, then the transmission angle has no physical meaning. In this case the intensity reflection coefficient has unit value, and the transmitted wave travels parallel to the boundary as if it were attached to it with an amplitude that decays exponentially with increasing distance from the boundary. This case is the acoustic analog of the situation in optics when there is complete internal reflection. *See* REFLECTION AND TRANSMISSION COEFFICIENTS; REFLECTION OF ELECTROMAGNETIC RADIATION; REFLECTION OF SOUND.

Absorption. When sound propagates through a medium, there are a number of mechanisms by which the acoustic energy is converted to heat and the sound wave weakened until it is entirely dissipated. This absorption of acoustic energy is characterized by a spatial absorption coefficient α for traveling waves. For a monofrequency traveling plane wave, Eq. (4) must be modified to Eq. (45). [Simi-

$$p = Pe^{-\alpha x} \cos\left[2\pi f\left(t - \frac{x}{c}\right)\right] \tag{45}$$

larly, a monofrequency traveling spherical wave like that of Eq. (18) would be multiplied by a factor $exp(-\alpha r)$.] The pressure amplitude diminishes exponentially with increasing distance and the sound pressure level decreases linearly with *x*, according

TABLE 2. Values of α/f^2 for selected gases and liquids*				
	Classical thermoviscous value, $\alpha/f^2 \times 10^{11}$, Np \cdot s ² /m	Observed value, $\alpha/f^2 \times 10^{11}$ Np \cdot s ² /m		
Gases				
Argon	1.85	1.87		
Helium	0.53	0.54		
Nitrogen	1.35	1.64		
Oxygen	1.61	1.92		
Liquids				
Acetone	7.0	30.0		
Benzene		750		
Ethyl alcohol		50		
Glycerine	3000	3000		
Mercury	6.0	8.0		
Water (fresh)	8.1	25.0		
*Temperature = 2 lbf/in. ² 1 Np · s ² /m	$0^{\circ}C = 68^{\circ}F$. Pressure = 1 atm = 0.3048 Np \cdot s ² /ft.	= 101.3 kPa = 14.7		

SOURCE: After L. E. Kinsler et al., *Fundamentals of Acoustics*, 3d ed., Wiley, 1982; and J. J. Markham, R. T. Beyer, and R. B. Lindsay, Absorption of sound in fluids, *Rev. Mod. Phys.*, 23:353–411, 1951.

to	Eq.	(46)). The	dimension	of α	is	the	inverse	meter
----	-----	------	--------	-----------	-------------	----	-----	---------	-------

$$SPL(x) = SPL(0) - 8.7 \alpha x$$
 (46)

(m⁻¹). For historical reasons, and to avoid ambiguity with other quantities like the wave number *k*, absorption is given in units of nepers per meter (Np/m), where the neper is a dimensionless label. Because of the use of sound pressure level in acoustics, spatial absorption can also be specified in terms of decibels per meter (dB/m). From Eq. (46), the absorption coefficient in dB/m is 8.7 α . For standing waves, the sound field dies away with time according to a factor exp($-\beta t$). The temporal absorption coefficient β , given in nepers per second (Np/s), is related to α by $\beta = \alpha c$. Some values of spatial absorption coefficients are given in **Table 2**. *See* NEPER.

Classical mechanisms. One mechanism of absorption is shear viscosity, the analog of friction for fluids. Molecules traveling between fluid elements of differing speeds will transfer momentum between the elements through intermolecular collisions, diffusing the differences in collective motion into random (thermal) motion of the individual molecules. This loss of acoustic energy into heat is described by the absorption coefficient for viscosity αv . In fluids, this coefficient increases with the square of the frequency (up to about 10^{10} Hz). *See* VISCOSITY.

Another mechanism arises from the thermal conductivity of the fluid. In a perfectly insulating fluid (zero thermal conductivity), cold regions could not conduct thermal energy away from hot regions. In real materials, small amounts of thermal energy can flow between these hotter and colder regions, tending to equalize the temperatures a little and converting acoustic energy into heat. This energy conversion is described by a thermal absorption coefficient α_t , which also increases with the square of the frequency. The losses from shear viscosity and thermal conductivity add together to give the classical thermoviscous absorption coefficient $\alpha_c = \alpha \upsilon + \alpha_l$. See CONDUCTION (HEAT).

Structural relaxation. In certain liquids, including water, acetone, and alcohols, the observed absorption is significantly larger than that predicted by the classical absorption coefficient but has the same frequency dependence. A postulated mechanism is a structural relaxation, involving transitions between two preferred clusterings of nearest neighbors, one favored at lower density and the other at higher. When the sound wave travels through the fluid, the pressure and temperature fluctuations cause transitions between these two structural states. The net effect is to introduce a bulk viscosity coefficient additive to the shear viscosity. In water, the effect of bulk viscosity results in a measured absorption coefficient about three times the classical value.

Thermal molecular relaxation. In some gases with polyatomic molecules, like air, thermal molecular relaxation is an important mechanism of absorption. When two gas molecules collide, in addition to simply rebounding from each other like two billiard balls, one or the other of the molecules may be excited into internal vibrations, thus removing energy from the collective acoustic motion. Once excited, the molecule may release this energy back into kinetic energy of motion during another collision, but this release is not coordinated with the acoustic wave and so appears at random, contributing to the thermal motion of the molecules. It takes a characteristic amount of time, called the relaxation time τ , to accomplish the conversion of energy into and from these internal vibrations. If τ is sufficiently small compared to the period T of the acoustic wave, then the mechanism is fully activated and the resulting absorption increases with the square of the frequency. At higher frequencies such that $T < \tau$, the absorption from this effect becomes a constant independent of the frequency. Thus, while at lower frequencies molecular thermal relaxation can be an important source of absorption in certain gases, at higher frequencies the effect is completely dominated by the classical thermoviscous absorption. In air, collisions involving carbon dioxide with either nitrogen or oxygen, and oxygen with water vapor, are important in exciting internal energy states. The total absorption coefficient is a strong function of humidity and frequency throughout the audible range of frequencies. The effects are strong enough that the reverberation characteristics of a poorly ventilated concert hall can change during the course of a concert as the moisture exuded by the audience alters the humidity. Figure 6 shows curves for the absorption coefficient in decibels per meter as a function of frequency for three different humidities and the classical thermoviscous value.

Chemical relaxations. In certain liquids, there can be chemical relaxations involving the concentrations of various chemical species and how they change with temperature. A sound wave passing through the liquid causes local temperature changes, and these in turn can cause changes in the ionization of dissolved salts. The basic mechanisms cause losses quite similar in effect to the molecular thermal relaxation



Fig. 6. Absorption of sound in air at 20°C (68° F) as a function of frequency for three humidities and the classical thermoviscous value. 1 dB/m = 0.3 dB/ft. (After H. E. Bass, H.-J. Bauer, and L. B. Evans, Atmospheric absorption of sound: Analytical expressions, J. Acous. Soc. Amer., 52:821–825, 1972)

described above. In seawater, chemical relaxations contribute appreciable additional absorption at frequencies below about 100 kHz for magnesium sulfate (MgSO₄) and below about 1 kHz for boric acid.



Fig. 7. Absorption of sound in fresh water and in seawater of 35 parts per thousand salinity at atmospheric pressure and 5°C (41°F). 1 dB/m = 0.3 dB/ft. (*After F. H. Fisher and V. P. Simmons*, Sound absorption in sea water, J. Acous. Soc. Amer., 62:558–664, 1977)

Figure 7 gives a representative curve of the frequency dependence of the absorption coefficient in decibels per meter for seawater. *See* SEAWATER. Alan B. Coppens

Bibliography. L. L. Beranek, *Acoustics*, 1986; R. E. Berg and D. G. Stork, *The Physics of Sound*, 2d ed., 1995; D. E. Hall, *Basic Acoustics*, 1987, reprint 1993; L. E. Kinsler et al., *Fundamentals of Acoustics*, 4th ed., 2000; P. M. Morse and K. U. Ingard, *Theoretical Acoustics*, 1986; M. Rossi, *Acoustics and Electroacoustics*, 1988; T. D. Rossing and N. H. Fletcher, *Principles of Vibration and Sound*, 1994.

Sound absorption

The process by which the intensity of sound is diminished by the conversion of the energy of the sound wave into heat. The absorption of sound is an important case of sound attenuation. This article will address absorption mechanisms in varying media but will concentrate on absorption in ideal gases (such as air), since these effects are the most commonly observed.

Attenuation. Regardless of the material through which sound passes, its intensity, measured by the average flow of energy in the wave per unit time per unit area perpendicular to the direction of propagation, decreases with distance from the source. This decrease is called attenuation. In the simple case of a point source of sound radiating into an ideal medium (having no boundaries, turbulent fluctuations, and the like), the intensity decreases inversely as the square of the distance from the source. This relationship exists because the spherical area through which the energy propagates per unit time increases as the square of the propagation distance. This attenuation or loss may be called geometrical attenuation.

In addition to this attenuation due to spreading, there is effective attenuation caused by scattering within the medium. Sound can be reflected and refracted when incident on media of different physical properties, and can be diffracted and scattered as it bends around obstacles. These processes lead to effective attenuation, for example, in a turbulent atmosphere; this is easily observed in practice and can be measured, but is difficult to calculate theoretically with precision. *See* DIFFRACTION; REFLECTION OF SOUND; REFRACTION OF WAVES.

In actual material media, geometrical attenuation and effective attenuation are supplemented by absorption due to the interaction between the sound wave and the physical properties of the propagation medium itself. This interaction dissipates the sound energy by transforming it into heat and hence decreases the intensity of the wave. In all practical cases, the attenuation due to such absorption is exponential in character. Thus, if I_0 denotes the intensity of the sound at unit distance from the source, then *I*, the intensity at distance *r* in the same units, has the form of Eq. (1), where *e* represents the base

I

$$=I_0 e^{-\alpha r} \tag{1}$$

of napierian logarithms, 2.71828 ... and α is called the intensity absorption coefficient. The α can be expressed by Eq. (2). Hence α has the dimensions

$$\alpha = \frac{-1}{r} \ln \left(\frac{I}{I_0} \right) \tag{2}$$

of reciprocal length or distance. However, it is customary to express the logarithm, which is a pure dimensionless number, in nepers, and hence α is in nepers per unit length. Since 10 log₁₀ (*I*/*I*₀) is the number of decibels between the intensity *I* and *I*₀, the absorption coefficient α can also be expressed in decibels per unit length or distance. *See* DECIBEL; NEPER; SOUND INTENSITY.

Classical sources. The four classical mechanisms of sound absorption in material media are shear viscosity, heat conduction, heat radiation, and diffusion. These attenuation mechanisms are generally grouped together and referred to as classical attenuation or thermoviscous attenuation. They will be discussed first as they apply to fluids, namely, liquids and gases.

Viscosity. In the flow of a fluid such as air or water, viscosity is the tendency for the fluid flowing in a layer parallel to the direction of motion to retard the motion in adjoining layers. The viscosity of a liquid (indicated qualitatively by the stickiness or sluggishness of its flow) decreases as the temperature rises, whereas that of gases increases with the temperature. *See* VISCOSITY.

Appropriate modifications of the standard wave equation for plane sound waves to take account of the effect of viscosity of a fluid lead to Eq. (3) for

$$\alpha_v = \frac{8\pi^2 f^2 \mu}{3\rho_0 c^3} \tag{3}$$

the absorption coefficient as defined in Eq. (1), but specifically due to viscosity, where μ is the viscosity of the fluid, *f* is the frequency of the sound wave in hertz, *c* is the wave's velocity, and ρ_0 is the equilibrium density of the fluid. The expression for α_v holds only for frequencies that are not too high. For example, in the case of air, it fails at frequencies on the order of 10⁹ Hz, which involve ultrasonics. *See* ULTRASONICS.

Heat conduction. The role of heat conduction can be qualitatively understood as follows. The compression associated with the high-pressure portion of a propagating sound wave produces a local increase in temperature, which sets up a temperature gradient through which heat will flow by conduction before the succeeding rarefaction cools the medium. This is an irreversible phenomenon, leading to energy dissipation into heat. The value of the sound absorption coefficient due to heat conduction is given by Eq. (4), where $\gamma = C_{\rho}/C_{\nu}$, the ratio of the molar spe-

$$\alpha_b = \frac{2\pi^2 f^2 \kappa}{\rho_0 c^3 C_v} \left(\frac{\gamma - 1}{\gamma}\right) \tag{4}$$

cific heat at constant pressure C_{ρ} to that at constant volume C_{ν} , ρ_0 is the equilibrium density of the fluid, and κ is the thermal conductivity. Equation (4) is

again valid only for frequencies that are not too high. *See* CONDUCTION (HEAT); HEAT CAPACITY.

Substitution of the appropriate numerical values for the physical parameters shows that for gases the value of α_b is somewhat less than α_v for the same frequency. For liquids, except for liquid metals, $\alpha_v \gg \alpha_b$ for the same frequency. It is usually considered appropriate to add α_v and α_b to get the total absorption due to these two processes.

Heat radiation. Thermal radiation results whenever one portion of a fluid acquires a temperature different from that of an adjacent portion. This effect is an electromagnetic-wave phenomenon different in character from heat transfer by conduction. Calculation shows that thermal radiation plays a very small role in sound absorption except possibly at very high frequencies in very low pressure gases. *See* HEAT RADIATION.

Diffusion. Diffusion operates mainly in the case of fluid mixtures, for example, air that is a mixture mainly of nitrogen and oxygen, in which the two gases have a slight tendency to diffuse into each other. For ordinary gases in which the components differ little in molecular weight, the effect of diffusion on absorption is negligible compared with that of viscosity and heat conduction. However, in gaseous mixtures in which the components differ markedly in molecular weight—for example, a mixture of the rare gases helium and krypton—the effect of diffusion can be much greater than that due to viscosity and heat conduction combined. *See* DIFFUSION.

Measurement of absorption in fluids. Sound absorption in fluids can be measured in a variety of ways, referred to as mechanical, optical, electrical, and thermal methods. All these methods reduce essentially to a measurement of sound intensity as a function of distance from the source.

Mechanical. The mechanical method, confined to the estimate of absorption in liquids, uses radiation pressure, proportional to the sound intensity, to "weigh" the sound in a vertical beam. *See* ACOUSTIC RADIATION PRESSURE.

Optical. The optical method, which is also mainly used for liquids, is based on the diffraction of a beam of light by the sound beam, the latter being treated essentially as a diffraction grating. The greater the intensity of the sound, the more diffraction bands are produced. *See* DIFFRACTION.

Electrical. The electrical methods, the most generally used for both gases and liquids, employ an acoustic transducer like a piezoelectric crystal to receive the sound beam as it traverses the medium whose absorption coefficient is to be evaluated, and to measure the variation in the sound intensity as the distance from the source is varied. An alternative is the use of an acoustic interferometer, which functions essentially in the same way as an optical interferometer. Still another scheme is based on the use of the reverberation time technique. In this method the medium is confined in a closed spherical vessel with highly reflective walls. The reverberation time, that is, the time taken by a sound pulse initially produced at the center of the vessel by a transducer (which can also act as a sound receiver) to decay to one-millionth of its original intensity provides a measure of the intensity loss and leads to an evaluation of the absorption coefficient. This method has been successfully applied to liquids and gases. *See* INTERFEROMETRY; MICROPHONE; REVERBERATION.

Thermal. The thermal method undertakes to estimate the loss in intensity of a sound wave by direct measurement of the heat produced in its passage. This demands very careful experimental technique, and the difficulties associated with it have provoked controversy about the usefulness of the method.

Experimental values for gases. Measurements carried out by these methods show that the experimental values for the sound absorption coefficient (as a function of frequency) of a monatomic gas like helium agree rather well with the values calculated on the basis of viscosity and heat conduction, except at very low pressures and very high frequencies. On the other hand, for more complicated gases, like air (though in general α varies roughly directly as the square of the frequency, as predicted by classical theory), the actual numerical values are greater (and in many cases very much greater) than the theoretically predicted ones. For example, for air at 20°C (68°F) and normal atmospheric pressure, the theoretical value of $(\alpha_v + \alpha_b)/f^2$ is 1.27×10^{-12} neper s²/cm, whereas the experimental value of α/f^2 for air of low humidity under the same conditions is approximately 8×10^{-12} in the same units.

Experimental values for liquids. The situation with respect to liquids is about the same. Mercury and the liquefied gases like argon, oxygen, and nitrogen have absorption coefficients in moderately good agreement (that is, within 10–15%) with the classically calculated values. But ordinary liquids like water and aqueous salt solutions show values much in excess of those classically calculated. The observed excess in the case of organic liquids like carbon disulfide, for example, is very large indeed, of the order of several hundred times. A satisfactory explanation for this excess absorption has been found in molecular and related relaxation processes.

Molecular relaxation in gases. Although sound propagation in a gaseous medium is phenomenologically a wave process, it actually takes place through the collisions of the molecules composing the gas. According to the kinetic theory, the molecules of a gas are in ceaseless random motion. When the gas is locally compacted by the compression portion of a propagating sound wave, the average translational kinetic energy of the molecules is locally increased. Energy has been introduced into the group of molecules. This energy increase is conveyed through collisions to the adjoining group of molecules, which in turn passes it on through the gas, and so on, and this continuous transfer represents the sound-wave energy.

A sound wave thus involves the momentary local compression of the medium, which by increasing the local density, according to kinetic theory, also leads to an increase in the average translational kinetic energy and hence to an increase in local pressure. In the case of a gas made up of polyatomic molecules such as oxygen and nitrogen, the increase in pressure cannot take place instantaneously, since some of the extra translational kinetic energy is transferred to the internal rotational and vibrational molecularenergy states. If one imagines the diatomic oxygen (O_2) molecule as a dumbbell-shaped weight, with a springy connecter between the weights, some of the collision energy acts to spin the molecule and some acts to vibrate the weights toward and away from each other. The dissipation of these energy states is referred to as rotational relaxation and vibrational relaxation, respectively.

After the local density has begun to decrease due to the rarefaction portion of the propagating wave, the pressure continues to build up as energy is transferred back to the translation form from the internal rotational and vibration states. The excess density and excess pressure thus get out of phase, as in the standard hysteresis phenomenon, which leads to an effective transformation of acoustic-wave energy into heat, corresponding to sound absorption. This process is termed a relaxation process, and the time taken for the density at any point to return to within the fraction 1 - (1/e) [approximately 63%] of its equilibrium value is termed the relaxation time, denoted by τ . See HYSTERESIS; MOLECULAR STRUCTURE AND SPECTRA.

Calculation of the effect. The results of the theoretical calculation of the effect of molecular relaxation will be discussed for the special case in which there are only two energy states of each molecule: (1) that in which it possesses translational kinetic energy only, and (2) that in which it possesses translational kinetic energy plus a single internal energy level. Here the relaxation time is given by Eq. (5), where κ_{12}^0 is the av-

$$\tau = \frac{1}{\kappa_{12}^0 + \kappa_{21}^0}$$
(5)

erage number of transitions per second from energy state 1 to energy state 2, and κ_{21}^0 is the average number of transitions per second from state 2 to state 1. The superscripts mean that the transition rates refer to the equilibrium condition of the gas as unaffected by the sound wave. Analysis then gives the sound absorption coefficient α due to this type of relaxation process in the form of Eq. (6), where *f* is the

$$\alpha = K \frac{4\pi^2 f^2 \tau}{1 + 4\pi^2 f^2 \tau^2} \tag{6}$$

frequency and *K* is a constant containing the molar gas constant *R*, the sound velocity, and the specific heats of the gas. If $2\pi f\tau \ll 1$, α varies as the square of the frequency, as in the case of viscous and heat conduction absorption. However, if $2\pi f\tau \gg 1$, the attenuation becomes a constant with respect to frequency and the effect is said to be frozen out. In the more general case, if $\alpha/2\pi f$ is plotted as a function of frequency, it has a maximum for $f = 1/2\pi\tau$, the plot taking the form of an approximate bell-shaped curve. Such a plot, carried out experimentally, provides a method for evaluating the relaxation time τ .

The theoretical evaluation of τ from Eq. (5) involves essentially a calculation in quantum statistics. The agreement of theory with experiment for many gases showing excess absorption over that attributable to viscosity and heat conduction is satisfactory. The relaxation process described here is usually termed thermal relaxation, since the specific heats of the gas enter prominently into the process.

Examples. Examples of gases for which experimental values are in reasonable agreement with theory include hydrogen, for which the excess absorption over classical values arises primarily from rotational relaxation, with a relaxation time on the order of 10^{-8} s, and oxygen, for which the excess low-frequency absorption depends primarily on vibrational relaxation, with a relaxation time on the order of 3×10^{-3} s at room temperature. In more complicated polyatomic gases, it is rare for a single relaxation process to be solely effective, and rather elaborate analysis is required to take account of multiple relaxations.

Accompanying relaxation absorption in gases is sound dispersion, which is the variation of sound velocity with frequency. The dispersion leads to the distortion of waveforms and features prominently in areas such as sonic boom propagation. *See* NONLIN-EAR ACOUSTICS; SONIC BOOM.

Relaxation processes in liquids. The excess absorption of sound in liquids above the classically predicted values has also been attributed to relaxation processes. The assumption of thermal relaxation has been successful for certain nonpolar liquids such as benzene. It has not worked, however, for polar liquids such as water. For the latter, so-called structural relaxation has been successfully invoked. This assumes that liquid water can exist in two different states of mutual molecular orientation and that the passage of sound affects the average rate of energy transfer back and forth between the two states. The state of lower energy is assumed to be similar to the structure of ice, in which each water molecule is surrounded on the average (in spite of the random motion of the molecules) by its four nearest neighbors arranged at the corners of a regular tetrahedron with the given molecule in the center. The higherenergy state corresponds to a more closely packed configuration of the molecules, with an arrangement like that of a face-centered cubic crystal. In water in equilibrium, most of the molecules are in the lowerenergy state, but the passage of sound causes a transfer of some molecules into the higher state and upsets the equilibrium. There is the usual lag in the return of the molecules to the lower state, leading to a relaxation time, and this produces absorption similar to that in gases. This theory has been successful in accounting for the excess sound absorption in water as a function of temperature in the range from 10 to 80° C (50 to 176° F). In this range, the appropriate relaxation time has been found to be of the order of 10^{-12} s.



Fig. 1. Absorption of sound in air at 20°C (68°F) and various values of relative humidity. (From L. E. Kinsler et al., Fundamentals of Acoustics, 3d ed., John Wiley, 1982; used by permission of John Wiley & Sons, Inc.)

Excess absorption in aqueous solutions has been treated successfully by chemical relaxation. A chemical reaction in the liquid such as a dissociationassociation process can provide two energy states, and the transfer of energy between these states will be affected by the compressive stress involved in the passage of a sound wave. The dissociation reaction of magnesium sulfate has been used to explain excess sound absorption in seawater.

In the atmosphere. The amount of sound that air absorbs increases with audio frequency and decreases with air density, but also depends on temperature and humidity. Sound absorption in air depends heavily on relative humidity, as seen in **Fig. 1**, which displays the predicted classical thermoviscous absorption as well as the actual absorption in air for varying relative humidity.

The reason for the strong dependence on relative humidity is molecular relaxation. One can note the presence of two transition regimes in most of the actual absorption curves, representing the relaxation effects of N_2 and O_2 , the dominant constituents of the atmosphere. *See* ATMOSPHERIC ACOUSTICS.

In seawater. Sound absorption in water is generally much less than in air. It also rises with frequency, and it strongly depends on the amount of dissolved materials (in particular, salts in seawater), due to chemical relaxation (**Table 1**). *See* UNDERWATER SOUND.

In solids. The theory of sound attenuation in solids is complicated because of the presence of many mechanisms responsible for it. These include heat conductivity, scattering due to anisotropic material properties, scattering due to grain boundaries, magnetic domain losses in ferromagnetic materials,

	Absor	otion coefficient,	dB/m†
	10 kHz	100 kHz	1 MHz
Fresh water	0.0001	0.006	0.2
Seawater	0.0008	0.04	0.3

interstitial diffusion of atoms, and dislocation relaxation processes in metals. In addition, in metals at very low temperature the interaction between the lattice vibrations (phonons) due to sound propagation and the valence of electrons plays an important role, particularly in the superconducting domain. *See* CRYSTAL DEFECTS; DIFFUSION; FERROMAGNETISM; SU-PERCONDUCTIVITY.

At boundaries. When sound waves passing through air or water encounter another material or strike a surface, part of the incident wave energy is reflected, part is absorbed within the material, and still another part is transmitted. The general property of interest in acoustical design is the absorption coefficient, a, of the surface: the ratio of the energy that is absorbed and transmitted, to the energy incident on that boundary. For a surface in a room, this coefficient measures the fraction of the energy that is not returned to the room, independent of the mechanism of extraction. For example, an open window has a = 1, since none of the sound comes back into the room, even though none of it may be immediately absorbed by the air. On the other hand, the intensity absorption coefficient α , discussed previously, refers specifically to energy that has been transformed from acoustic energy to thermal energy within the medium during propagation.

There is no simple way of relating a, the sound absorption by a boundary, to α , the absorption of sound on passing through a continuous solid, liquid, or gas. The values of a used in acoustical design are typically empirically determined, being derived from laboratory or field tests of specimens of the particular material of interest.

Practical absorbing materials for airborne and water-borne sound are based on the mechanisms of viscosity and heat conduction. Most useful sound absorbers bring these mechanisms into play by being porous, perforated, or slotted yet having a significant resistance to the flow of air or water through their openings. The sound-wave movement through such openings dissipates energy by viscosity and thermal conduction with the walls. In the case of airborne sound, examples are mineral wool, fiberglass batts, acoustical ceiling tiles, and domestic drapes, all of which are very effective above 1 kHz. *See* ARCHITEC-TURAL ACOUSTICS; MUFFLER.

Material ratings. Full-sized material samples are typically tested in a laboratory reverberation room. The

random-incidence sound absorption (absorption of sound waves coming from all directions) of acoustical materials is measured at frequencies between 125 and 4000 Hz, which corresponds to a significant fraction of the normal audio range encountered in speech and music. The average of the absorption at 250, 500, 1000, and 2000 Hz is called the noise reduction coefficient (NRC). Small material samples are tested in a device known as an impedance tube, which relates the incident sound wave to the reflected sound wave from the test specimen. In this case, the normal-incidence sound absorption (absorption of sound impinging along a line perpendicular to the absorber) is measured.

Applications to room design. Room walls, floors, and ceiling surfaces reflect and absorb sound differently in accordance with the acoustic properties of their construction material, that is, depending on whether they are concrete, glass, brick, or carpet, for example. This is depicted in **Fig. 2**, which shows how sound can arrive at a receiver from many paths with varying degrees of absorption along the route.

Sound generation in the average nonindustrial room occurs via speech, music, equipment or appliance noise, and occupant movements. The subsequent sound-level buildup or reverberation may be undesirably high or low depending on the intended use of the room. If the reverberation is too great, the room is typically referred to as a live room; if it is too low, the room is referred to as a flat room. Reverberation level can be modified by treatment of the room's interior surfaces by adding materials with greater or lesser sound absorptive qualities. The acoustic design of a room depends on its intended use, whether as a conference room, music room, lecture room, or open-plan office. For example, if the intended use is as a church, a greater reverberation may be desired for the typical music played. If the intended use is as an office space, less reverberation may be desired for speech communication. Conventional rugs, hanging drapes, stuffed furniture, and carpeting are typical sound absorption materials used in many applications. However, additional treatment is often necessary. See ACOUSTIC NOISE; SOUND FIELD ENHANCEMENT.

Once the appropriate materials are selected,



Fig. 2. Depiction of direct and reflected sound in a room with varying wall properties.

materials*						
	Frequency, Hz					
Material	125	250	500	1000	2000	4000
Acoustic paneling	0.15	0.3	0.75	0.85	0.75	0.4
Brick	0.05	0.04	0.02	0.04	0.05	0.05
Brick, painted	0.01	0.01	0.02	0.08	0.02	0.03
Concrete block	0.36	0.44	0.31	0.29	0.39	0.25
Curtains	0.05	0.12	0.15	0.27	0.37	0.50
Floor, concrete	0.02	0.02	0.02	0.04	0.05	0.05
Floor, wood	0.15	0.2	0.1	0.1	0.1	0.1
Floor, carpeted	0.1	0.15	0.25	0.3	0.3	0.3
Glass fiber						
insulation, 2 in.	0.17	0.55	0.8	0.9	0.85	0.8
Glass, window	0.35	0.25	0.18	0.12	0.07	0.04
Gypsum board						
on studs	0.29	0.10	0.05	0.04	0.07	0.09
Plaster	0.03	0.03	0.02	0.03	0.04	0.05
Plywood paneling	0.28	0.22	0.17	0.09	0.10	0.11
Water surface	0.008	0.008	0.013	0.015	0.020	0.02
One person	0.18	0.4	0.46	0.46	0.51	0.46

usually with the help of a professional acoustical consultant, the area of treatment, A, is calculated by means of the Sabine formula [Eqs. (7)] to determine

$$T = 0.049 \frac{V}{aA}$$
 (English) (7*a*)

$$T = 0.161 \frac{V}{aA} \qquad \text{(metric)} \qquad (7b)$$

the reverberation time T required for the intended use. In the Sabine formula, V is the room volume in cubic feet or cubic meters, a is the absorption coefficient of the acoustical material, and A is the area in square feet or square meters of the room. If the existing reverberation time for a room is unsatisfactory, acoustical materials with varying values of acan be used to replace portions of the room surface A to modify the reverberation time. The value of a for common acoustical materials depends on frequency; typical values are indicated in **Table 2**.

Industrial applications. In industrial settings it is the overall sound level rather than the quality and intelligibility of the sound which is typically the driving concern. Industrial noise is often at levels sufficient to cause damage to hearing if it is not reduced or if hearing protection (such as headphones, which absorb sound before it enters the ear) is not used. Noise reduction is often achieved by adding soundabsorbing materials and enclosures around the noise source. Qualitatively, the treatment of a noisy industrial setting is quite similar to the treatment of a noisy office space, but special care must be paid to material selection due to the harsher operating environment of a typical factory.

Henry E. Bass; Angelo J. Campanella; James P. Chambers; R. Bruce Lindsay

Bibliography. R. T. Beyer and S. V. Letcher, *Physical Ultrasonics*, 1969; M. J. Crocker (ed.), *Handbook of Acoustics*, 1998; L. E. Kinsler et al., *Fundamentals of Acoustics*, 3d ed., 1982; A. D. Pierce, *Acoustics: An Introduction to Its Physical Principles and Ap-*

plications, 2d ed., 1989; C. E. Wilson, Noise Control, 1989.

Sound field enhancement

A system for enhancing the acoustical properties of both indoor and outdoor spaces, particularly for unamplified speech, song, and music. Systems are subdivided into those that primarily change the natural reverberation in the room, increasing its level and decay time (reverberation enhancement systems); and those that essentially replace the natural reverberation (sound field synthesis systems). Both systems may use amplifiers, electroacoustic elements, and signal processing to add sound field components to change the natural acoustics. Sound field enhancement is used to produce variable acoustics, to produce a particular acoustics which is not attainable by passive means, or because a venue has one or all of the following deficiencies: (1) unsuitable ratio between direct, early reflected, and reverberant sound; (2) unsatisfactory early reflection pattern; and (3) short reverberation times. See AMPLIFIER.

The purpose of sound field enhancement systems is not to provide higher speech intelligibility or clarity—in contrast to sound reinforcement systems—but to adjust venue characteristics to best suit the program material and venue, and in such a way optimize the subjectively perceived sound quality and enjoyment. *See* SOUND-REINFORCEMENT SYSTEM.

The use of sound field enhancement ranges from systems for homes (surround sound systems and home theater sound systems) to systems for rooms seating thousands of listeners (concert halls, performing arts centers, opera houses, and churches). Outdoor spaces may have elaborate designs to approach the sound quality of indoor spaces (advanced concert pavilions, outdoor concert venues). Systems may be used to improve conditions for both performers and listeners. *See* ARCHITECTURAL ACOUSTICS; RE-VERBERATION.

Basic elements. A sound field enhancement system may consist of one or more microphones, systems for amplification and electronic signal processing, and one or more loudspeakers. The signal processing equipment may include level controls, equalizers, delays and reverberators, systems for prevention of howl and ringing, as well as frequency dividing networks and crossover networks. *See* EQUAL-IZER; LOUDSPEAKER; MICROPHONE.

Prerequisites. Many indoor venues have acoustical characteristics unsuitable for communication such as speech, song, or music. Outdoor venues seldom have satisfactory acoustical conditions for unamplified acoustic communication. Satisfactory acoustic conditions include for speech, good speech intelligibility and sound quality; and for music, appropriate loudness, clarity, reverberation envelopment, spaciousness, and other sound quality characteristics for the music to be performed and enjoyed. Rooms with insufficient reverberation are often perceived



Fig. 1. Classification scheme for sound field enhancement systems.

to be acoustically "dry" or "dead." Insufficient reverberation usually is due to lack of room volume for the existing sound absorption areas in the room. The reverberation time is essentially proportional to room volume and inversely proportional to sound absorption, as indicated by Sabine's formula. In these cases, sound field enhancement systems may be used to improve acoustical conditions, by increasing the sound level and reverberation time. Absence of noise is also necessary for good acoustic conditions. *See* ACOUS-TIC NOISE; SOUND ABSORPTION.

System design. A simple reverberation enhancement system uses a microphone to pick up the sound signal to be reverberated, amplified, and reradiated in the auditorium. In most cases, reverberation is added to the original signal by electronic digital signal processing. An advanced sound field enhancement system uses several, possibly directional microphones to pick up sound selectively from the stage area. Signal processing and many loudspeakers are used to obtain the desired sound field for the audience and performers. Most performing art centers need variable acoustic conditions that can be changed instantaneously using electronic controls.

Categories. Sound field enhancement systems can be classified according to different criteria. One possible scheme is shown in Fig. 1. Sound field synthesis systems are designed to provide the necessary sound field components, such as early reflections from side walls and ceiling (the latter are usually called envelopmental sound), and reverberation by using digital filters, which can be designed to imitate a specific desired room response. Reverberation enhancement systems are designed to primarily increase the reverberation time and sound level of the reverberant field, while having negligible influence on the direct sound and the early reflected sound. Reverberation enhancement systems can be further subdivided into regenerative systems, based on the acceptance of unavoidable positive feedback between loudspeakers and microphones to increase the reverberation time; and nonregenerative systems, based on the use of electronic reverberators. Regenerative systems can be designed so as to work over a wide or narrow frequency range; nonregenerative systems are usually wide-frequency-range systems. *See* SOUND; REFLEC-TION OF SOUND.

Routing. A sound field synthesis system can generally be modeled as shown in **Fig. 2**. The system uses interconnected microphones and loudspeakers to form a more or less complete matrix of channels, where signal processing varies between channels. The signals from the microphones to the loudspeakers can be sent either directly between a specific microphone and loudspeaker pair, or in some form of matrix system feeding the signal from a microphone to several loudspeaker channels. A matrixed system can be designed as a phased array loudspeaker system. Another possibility is for a matrixed system to act as multiple individual systems, with uncorrelated or correlated outputs. Digital signal processing may be used in the circuits for various reasons.



Fig. 2. A sound field enhancement system can generally be modeled as a matrix. The system uses microphones (M) and loudspeakers (L).

Direct sound and room reverberation. The concept of room impulse response is useful in understanding sound field enhancement. The room impulse response describes the sound pressure in a room as a result of a very short sound pulse. Hearing uses binaural analysis to separate various types of information such as direction of incidence, direct sound from reverberation, and so on. The binaural room impulse responses for various source and receiver positions characterize the acoustics of the venue. *See* ACOUS-TICS; BINAURAL SOUND SYSTEM; SOUND PRESSURE.

It is common to subdivide the (binaural) room impulse response in the range above 250 Hz into three temporal sections. For music in large venues, these sections are often direct (0-20 milliseconds), envelopmental (20-150 ms), and reverberant sound (beyond 150 ms), all times referred to after the arrival of the direct sound from the source to a listener's location. The time ranges are approximate, somewhat individual, and depend on the frequency range considered. The desirable decibel levels of the three sound field components (direct, envelopmental, and reverberant sound) depend on the use of the venue-for example, whether it is to be used for an organ recital, a symphony, an opera, or a speech. Conditions for optimizing these levels can be attained instantaneously by electronic means using a sound field enhancement system, which can be used to change the reflection patterns as well as the relative decibel level between the different sound field components. The spatial characteristics of the acoustics, such as spaciousness and envelopment, are mainly determined by the directional properties of the envelopmental sound and the reverberant sound.

Direct sound. Direct sound is the combination of both the true direct sound (which has traveled directly from source to listener) and the various early reflections which follow within the first 20 ms. Direct sound usually determines the perceived direction of the sound source. In active systems, unwanted shift of the apparent source direction may



Fig. 3. Basic signal paths of a sound field enhancement system showing an acoustic feedback channel from a loudspeaker back to a microphone.

occur due to excessively strong additional sound components from unwanted directions. Halls for classical music may need the direct sound to be weakened relative to the envelopmental and reverberant sound, if the clarity is excessive. Amplification of direct sound is needed when (1) the sound source is weak, (2) sound is attenuated while passing over an audience area, (3) the venue is very large, and (4) the performance space is outdoors.

Envelopmental sound. The envelopmental sound has been reflected against walls and ceilings relatively few times, particularly in typical halls seating 1000– 3000 people. The envelopmental sound usually is a major part of the energy in the sound reaching the listener, and is therefore critical to the balance of the three sound field components. In many wide halls, lacking sufficiently reflecting side walls and ceiling, or in halls with absorptive balconies on the side walls, the envelopmental sound may be too weak, and it will then be necessary to add envelopmental sound. One of the major problems in enhancing envelopmental sound is obtaining equal coverage over the audience area.

Reverberant sound. The impression of reverberation on classical music is determined by the decay characteristics of the envelopmental sound and reverberant sound, as well as by the degree of correlation between these sounds at the listener's ears. The degree of interaural cross-correlation influences the perceived envelopment and spaciousness. Both the decay rate and relative level of the reverberant sound are critical to the overall impression of reverberation. However, on "running" music it is usually only the envelopmental sound and the early part of the reverberation of a room which is heard, because new music signals continue to arrive to be analyzed by hearing. Therefore, the perceived reverberation characteristics for running music and those for a final chord may be different, the latter reverberation being audible longer.

Feedback. In any real environment there will be unavoidable positive acoustic feedback from the loudspeaker(s) back to the microphone(s) (**Fig. 3**.) The feedback gives rise to regeneration, which may be controlled in various ways. Uncontrolled regeneration, that is, system oscillation (usually called howl), is obtained at and above a critical channel gain setting. Other effects of regeneration are ringing (that is, temporal artifacts such as flutter echo, primarily in wide-bandwidth systems) and coloration (that is, spectral artifacts such as frequency response peaks). From a control viewpoint, the feedback is characterized by its loop gain, which depends on the microphones, loudspeakers, amplification, signal processing, and the room itself. *See* GAIN; OSCILLATION.

Stability and coloration. Because of the acuity of hearing, coloration is noticed at much lower loop gains than oscillation because sound at some frequencies is exaggerated. An example of frequency response curves at some different gain settings of a feedback system involving a room is shown in **Fig. 4**.

The regeneration also prolongs the decay time of the signal because the sound energy from the



Fig. 4. Example of frequency response curves for a room with a feedback system. The upper curve shows response for loop gain setting close to oscillation; the lower curves shows response for gain settings of -5, -10, -15, -20 dB relative to this level.

loudspeakers compensates the sound absorbed by audience, chairs, and so on, leading to an increased reverberation time. In simple systems, regeneration effects are subjectively annoying since only sound at a few frequencies is subject to increase of reverberation time and level. In a system consisting of many channels, the gain can be adjusted so that each channel has low gain and coloration, and therefore the combined effect of such a system will yield both prolonged reverberation, over the entire operating frequency range of the channels, and negligible coloration.

Regenerative systems use this feedback as a means of achieving the desired sound field, whereas nonregenerative systems try to achieve the desired conditions with as little feedback as possible in order to have maximum control. Sound field synthesis systems are usually nonregenerative. To change acoustic conditions outdoors, only sound field synthesis systems are used.

Nonregenerative systems. Sound field synthesis can often be used to good advantage, provided that the environment is acoustically dry, and regeneration has been minimized by use of directional microphones and loudspeakers. It can also be used when the signal is obtained from a source other than microphones, for example in a surround sound system or a home theater system, in which case there is no possibility of regeneration. In these systems, both envelopmental and reverberant sound is simulated using signal processing and is radiated via side and rear loudspeakers.

Direct sound enhancement. Systems for amplification of direct sound may be configured as single-channel

or multichannel systems. The sound level may be increased relative to that of the unamplified direct sound by using the masking effect of the direct sound. Any substantial positive acoustic feedback will give rise to coloration. Therefore, these systems require directive or close-mounted microphones, as well as highly directive loudspeakers radiating primarily at sound absorptive surfaces in the room, such as audience areas, to minimize feedback. Loudspeakers should be near the stage so that the audible directional impression will agree with the visual impression.

Envelopmental sound enhancement. Improved envelopmental sound can also be implemented along the same lines as direct sound amplification. The delayed microphone (or other) signals should arrive in a suitable time pattern at all places in the audience area. Digital filters are used to provide single or multiple delays. Because of the limited number of channels and the need for wide sound coverage and substantial envelopmental sound levels, systems have to be protected from positive feedback, which could otherwise give rise to coloration. *See* DIGITAL FILTER.

Envelopmental sound enhancement can be achieved by both coherent and noncoherent designs. In coherent designs, arrays of close-mounted loudspeakers are used.

An example of a coherent system is the wave field synthesis approach, the most advanced of the nonregenerative systems suggested so far. One can consider it as a holographic system because it uses an array of microphones covering the source, and an extensive, dense array of loudspeakers surrounding the listening area. It is possible to approximately simulate the reflected waves of a space much larger than the one in which the system is installed. For the system to work as intended, the transducers must be very close and individually steered from the matrix. *See* TRANSDUCER.

It is also necessary to have some form of adaptive control over the microphones so that the microphone array can effectively track the sources (to avoid noise problems). In early implementation attempts, it was necessary to severely compromise these principles. A noncoherent approach is used by a beacon system, where integrated, small (wallmounted) microphone-loudspeaker channels pick up sound and retransmit it with suitable delay and reverberation characteristics. Adaptive systems can be used to control and eliminate the internal feedback in beacon systems. If a number of beacons are used in a moderately sound absorbing or sound diffusing surface, one can create an illusion of a more remotely positioned sound reflecting surface or object. See ACTIVE SOUND CONTROL.

Reverberant sound enhancement. Reverberation is obtained along the same lines as the envelopmental sound enhancement systems except that many more and quite evenly spaced loudspeakers are necessary to achieve good coverage over the audience area. The signal is sent through long finite or infinite impulse response digital filters to be reverberated. Many techniques are available to design infinite impulse response filters with little coloration. In practice, however, finite impulse response filters offer superior sound quality, and can be based on calculated or measured room impulse responses. Usually several filters, having insignificant correlation, will be used.

In acoustically dry rooms, the reverberation responses of the filters dominate the reverberation characteristics heard by the audience. In practice, the unavoidable scattered sound by various sound reflecting surfaces (for example, chair back rails) will fill in and smooth the reverberant sound field from the loudspeaker system.

Regenerative systems. The primary reason for using regenerative systems is that some regeneration (positive feedback) will always be present in any sound field enhancement system. Thus, it might be better to try to turn the coloration due to regeneration into something useful instead of trying to eliminate the coloration by other, perhaps more audible methods. Regenerative systems are used only for enhancement of reverberant sound. Microphones are thus placed far from both sound sources and system loudspeakers.

Noncoherent systems. In noncoherent regenerative system, microphones and loudspeakers are not positioned in regular arrays, as is often the case in sound field synthesis systems, but are spread out over the room (hung from ceilings or wall-mounted).

In the simplest systems, each microphone-signal processor-loudspeaker channel in the system is allowed to act independently. The microphones and loudspeakers are interconnected to form a matrix of channels where only the diagonal contains nonzero coefficients. Each channel includes a level control, equalizer, and power amplifier (and possibly a system for minimizing regeneration). This is the approach of the classical reverberation enhancement systems, both narrow and wide frequency band. More complex systems may use microphones and loudspeakers to form a more or less complete matrix of channels of pair combinations, having nonzero transfer functions in all channels. *See* POWER AMPLIFIER.

Coherent systems. Coherent regenerative systems use the same basic approach as more complex noncoherent systems, but channels are interconnected using various signal processing to form many coherent channels, reminiscent of some sound field synthesis systems.

Control of regeneration. There are multiple ways of controlling the loop gain in order to reduce the problems of instability and coloration. The loop gains of the electroacoustic channels vary considerably with frequency and location. The expected value of the peak to mean ratio of the sound pressure level response of a room, as a function of frequency, is not very dependent on the room reverberation time for wide-frequency-range sounds. One can show that the instability condition in a system channel will be fulfilled approximately 10 dB before wideband instability. On the average, the frequency intervals between peaks are inversely proportional to reverberation time.

Directional transducers. The loop gain can be reduced, for example, by using highly directive microphone and loudspeaker systems. Because of sound reflections from walls and so forth, it is usually not possible to achieve a large reduction in the loop gain in this way. Even if loudspeaker systems for envelopmental sound are designed to radiate primarily at the (soundabsorbing) audience areas, their side lobes and the reflection of sound from room surfaces set a limit to the reduction of the loop gain. Microphone systems suffer the same limitations whether highly directive microphones or adaptive steered arrays are used. In any case, the directive properties of the sound sources cause problems. Most music sound sources are quite directional, and even speech is so directive that it is probably necessary to use personal microphones in order to avoid detrimental effects due to directivity when performers are moving.

Adaptive filters. It is impossible to predict the exact location in frequency of the loop gain peaks because of changes in room temperature, movement by persons, and so on. Commercial adaptive electronic notch (band rejection) filters can be used which automatically detect and remove unwanted loop gain peaks. This is necessary because after a peak has been controlled, once the channel gain is increased a new peak will quickly appear.

Modulation by linear periodic time-variant filters. Under typical conditions, the mean distance between adjacent sound pressure level maxima (in a narrow frequency band) is approximately a wavelength. Consequently, higher loop gain or less coloration can be obtained

by moving the sound source or the loudspeakers, or by adding a small shift to the signal frequency. Moving loudspeakers and sources is impractical. Instead, time variance is achieved by use of digital signal processing to modulate the loop gain. Several modulation methods are possible such as amplitude, phase, delay, and frequency modulation. All time variance introduces sidebands, that is, extra frequency content. *See* AMPLITUDE MODULATION; FREQUENCY MODULATION; MODULATION; PHASE MODULATION.

The modulation methods vary as to audibility. Generally, slow modulation (low modulation frequency) is heard as beats, whereas fast modulation (high modulation frequency) is heard as roughness. The sensitivity of hearing to modulation is highest for modulation frequencies around 2-4 Hz, which is unfortunate since this frequency range is often desirable from a technical viewpoint to average the room frequency response. To minimize unwanted audible effects, the modulation frequency should be kept as low as possible, preferably below 1 Hz. Low-pass, filtered random signals can be used for modulation. Most commercial reverberation enhancement systems use some form of modulation. A challenge for future room acoustics research is to design new modulation methods, inaudible yet effective at preventing regeneration. See SOUND-REPRODUCING SYSTEMS. Mendel Kleiner

Bibliography. W. Ahnert and F. Steffen, *Beschallung-stechnik, Grundlagen und Praxis*, S. Hirzel, Leipzig, 1993; M. Barron, *Auditorium Acoustics*, E & FN Spon, London, 1993; L. Beranek, *Concert and Opera Halls*, Acoustical Society of America, Woodbury, NY, 1996; H. Kuttruff, *Room Acoustics*, 4th ed., Spon Press, London, 2000; M. Kleiner and P. Svensson, A review of active systems in room acoustics and electroacoustics, *ACTIVE95 Proceedings of the International Symposium on Active Control of Sound and Vibration*, Newport Beach, CA, July 1995, included in *Proceedings of the Sabine Centennial Symposium on CD ROM* by Acoustical Society of America Publications, P.O. Box 1020, Sewickley, PA 15143-9998.

Sound intensity

A fundamental acoustic quantity which describes the rate of flow of acoustic energy through a unit of area perpendicular to the flow direction. The unit of sound intensity is watt per square meter. The intensity is calculated at a field point (x) as a product of acoustic pressure p and particle velocity u. Generally, both p and u are functions of time, and therefore an instantaneous intensity vector is defined by Eq. (1). The time-variable instantaneous intensity,

$$\vec{I}_i(x,t) = p(x,t) \cdot \vec{u}(x,t) \tag{1}$$

 $\vec{I}_i(x, t)$, which has the same direction as $\vec{u}(x, t)$, is a nonmoving static vector representing the instanta-

neous power flow through a point (*x*). *See* POWER; SOUND PRESSURE.

Frequency spectrum. Many acoustic sources are stable at least over some time interval so that both the sound pressure and the particle velocity in the field of such a source can be represented in terms of their frequency spectra. The static sound intensity vector $\vec{I}(x, \omega)$ represents the time-independent sound power, which propagates through the point (x) in the direction of the particle velocity. The vector $\vec{I}(x, \omega)$ can be calculated at a frequency ω (or frequency band $\Delta \omega$) from Eq. (2), where φ is the phase

$$\vec{I}(x,\omega) = p(x,\omega) \cdot \vec{u}(x,\omega) \cdot \cos(\varphi)$$
(2)

shift between the real quantities p and u. It is practical to define both p and u as complex quantities so that the intensity vector \vec{I} can be calculated from Eq. (3), where * denotes a complex conjugate.

$$\vec{I}(x,\omega) = \frac{1}{2} Re[p(x,\omega) \cdot \vec{u}^*(x,\omega)]$$
(3)

Because intensity represents power, \vec{I} is always a real quantity. *See* COMPLEX NUMBERS AND COMPLEX VARI-ABLES; PHASE (PERIODIC PHENOMENA).

Measurement. The applications of sound intensity were fully developed after a reliable technique for intensity measurement was perfected. Sound intensity measurement requires measuring both the sound pressure and the particle velocity. Very precise microphones for sound-pressure measurements are available. Because there are no microphones for a direct particle velocity measurement, an indirect technique based on Euler's equation (4), which links

$$\vec{u}(x,\omega) = -\frac{1}{jk\rho c} \frac{\partial p(x,\omega)}{\partial x}$$
(4)

u with a pressure gradient, is used. Here, ρ is the density of the fluid medium, *c* is the sound speed, *k* is the wave number ($k = \omega/c$), and $j = \sqrt{-1}$. See MICROPHONE.

Measuring the sound pressure at close points with two precision microphones approximates the pressure gradient. To keep the measurement errors at an acceptable level, both microphones must have identical sensitivity and be phase-matched to 0.05°. The calculation of sound intensity is usually obtained by feeding the microphone output into a dual-channel fast-Fourier-transform (FFT) analyzer. The intensity is proportional to the imaginary part of the crossspectrum of the input voltages. This measurement arrangement measures the intensity component in the direction of an axis connecting the two microphones. To determine the total intensity vector, three perpendicular components must be obtained by measuring along three mutually perpendicular axes. Microphone probes, consisting of six microphones, to implement such a measurement have been constructed. Both international and national standards define the requirements on the instrumentation to secure the needed measurement precision. See FOURIER SERIES AND TRANSFORMS.

The usefulness of the measurement technique based on two microphones goes beyond the measurement of sound intensity. The sound pressure can be obtained by interpolating the pressure measured by the two microphones. Similarly, both the potential and kinetic energy densities as well as the particle velocity vector can be obtained.

Applications. One essential application of the intensity technique is the measurement of sound power radiated from sources. Such a measurement can be done with a precision unobtainable from pressure measurements. The radiated sound power is obtained from intensity measurements over an area circumscribing the source. The knowledge of the radiated power makes it possible to classify, label, and compare the noise emissions from various pieces of equipment and products and to provide a reliable input into environmental design. The intensity technique is perfect for transmission and insertion-loss measurements. Because the measurements can be done in the near field of acoustical radiators, a reliable overall functional analysis and identification of sound radiating sources can be performed. General sound-field studies and mapping are essential for many applications. See SOUND. Jiri Tichv

Bibliography. F. J. Fahy, *Sound Intensity*, 2d ed., E & FN Spon (an imprint of Chapman & Hall), London, 1995.

Sound pressure

The incremental variation in the static pressure of a medium when a sound wave is propagated through it. Sound refers to small-amplitude, propagating pressure perturbations in a compressible medium. These pressure disturbances are related to the corresponding density perturbation via the material equation of state, and the manner in which these disturbances propagate is governed by a wave equation. Since a pressure variation with time is easily observed, the science of sound is concerned with small fluctuating pressures and their spectral characteristics. The unit of pressure commonly used in acoustics is the micropascal (1 μ Pa = 1 μ N/m² = 10⁻⁵ dyne/cm² = $10^{-5} \mu$ bar). One micropascal is approximately 10^{-11} times the normal atmospheric pressure. See PRES-SURE; PRESSURE MEASUREMENT; WAVE MOTION.

The instantaneous sound pressure at a point can be harmonic, transient, or a random collection of waves. This pressure is usually measured with an instrument that is sensitive to a particular band of frequencies. If the pressure is measured in a specified frequency band over a short period of time, the mean value of the fluctuating pressure is found to be zero, but the mean value of the square of the pressure, $\langle p^2 \rangle$, is a positive quantity. The square root of $\langle p^2 \rangle$ is the root-mean-square pressure, $p_{\rm rms}$, and is proportional to the time average of the sound intensity, $\langle I \rangle$. The measured pressures will depend on the bandwidth and filter response of the measurement instrument. The peak sound pressure for any specified time interval is the maximum absolute value of the instantaneous sound pressure in that interval and is equal to $\sqrt{2} p_{\rm rms}$ for sinusoidal signals. Acoustic pressures are commonly studied as a function of frequency, where the frequency can range over the sonic band, which lies between 20 and 20,000 Hz and is bounded by the infrasonic and ultrasonic frequency bands. The Fourier transform and its digital implementation, the fast Fourier transform (FFT), are widely used to characterize the spectrum of the $\langle p^2 \rangle$ versus frequency. Thus $\langle p^2 \rangle$ in a given band, Δf , is often normalized by this bandwidth to yield the spectral density, $\langle p^2 \rangle / \Delta f$, with units of $[\mu Pa]^2$ /Hz. *See* FOURIER SERIES AND TRANSFORMS; SOUND INTENSITY.

A concept widely used in acoustics is "level," which refers to the logarithm of the ratio of any two field quantities. When the ratio is proportional to a power ratio, the unit for measuring the logarithm of the ratio is called a bel, and the unit for measuring this logarithm multiplied by 10 is called a decibel (dB). The sound intensity, I, which describes the rate of flow of acoustic energy (acoustic power flow) per unit area, is given by the mean square pressure, $\langle p^2 \rangle$, divided by the acoustic impedance, defined as the product of the medium density, ρ , and compressional wave speed, C; that is, the intensity $I = \langle p^2 \rangle / \rho C$. A ratio proportional to power can be formed if a reference quantity is specified. The reference pressure for water is usually chosen to be 1 μ Pa with $\rho C \approx$ 1.545×10^{6} Pa \cdot s/m, and the reference pressure for air is 20 μ Pa with $\rho C \approx 413$ Pa \cdot s/m. The effective sound pressure level (SPL, symbolized L_p) is a useful quantity since the power transmitted by a sound wave can be related to $\langle p^2 \rangle$ for a specified medium; however, when comparing levels in seawater to those in air, the acoustic impedance difference between the two media must be considered. With these considerations in mind, the SPL [dB re p_{ref}] is given by Eq. (1).

SPL [dB re p_{ref}] =
$$L_p = 10 \log_{10}(\langle I \rangle / \langle I_{ref} \rangle)$$

= $10 \log_{10} \left(\langle p^2 \rangle / \langle p_{ref}^2 \rangle \right)$ (1)

See DECIBEL.

When the sound pressure level is restricted to a frequency band, the term "band sound pressure level" (BSPL, symbolized L_{pb}) is used. Air acoustics standards use a weighted filter response such as an A or C filter with either fast (*F*, 125-ms) or slow (*S*, 1000-ms) exponential time averaging. In these instances the levels are designated by L_A [dB re 20 µPa] or L_C [dB re 20 µPa]. When the mean square pressure is normalized by the bandwidth, the quantity becomes a density and is referred to as a spectral density or a spectrum level with the sound pressure spectrum level (SPSL, symbolized L_{ps}) defined by Eq. (2).

SPSL [dB re
$$\langle p^2 \rangle_{ref}/Hz$$
] = L_{ps}

$$= 10 \log_{10} \left[\left(\langle p^2 \rangle / \Delta f \right) / \left(\langle p^2 \rangle / \Delta f \right)_{\text{ref}} \right] \quad (2)$$

See NOISE MEASUREMENT; SOUND. William M. Carey

Sound recording

The technique of entering sound, especially music, on a storage medium for playback at a subsequent time. The storage medium most widely used is magnetic tape. *See* MAGNETIC RECORDING.

Monophonic sound recording. In the simplest form of sound recording, a single microphone picks up the composite sound of a musical ensemble, and the microphone's output is recorded on a reel of 1/4-in. (6.35-mm) magnetic tape. This single-track, or monophonic, recording suffers from a lack of dimension, since in playback the entire ensemble will be heard from a single point source: the loudspeaker.

Stereophonic sound recording. An improved recording, with left-to-right perspective and an illusion of depth, may be realized by using two microphones which are spaced or angled appropriately and whose outputs are routed to separate tracks on the tape recorder. These two tracks are played back over two loudspeakers, one each on the listener's left and right. Under ideal conditions this stereophonic system will produce an impressive simulation of the actual ensemble, and the listener may identify the apparent location of each integral part of the ensemble. *See* STEREOPHONIC SOUND.

As early as 1931, Alan Blumlein applied for a patent on several stereophonic microphone placement techniques. Despite the advances in recording technology that have been made since then, the methods described in his patent may still be found in use today. One example uses two microphones constructed within a single housing. The microphones' diaphragms are at right angles to each other, and when the device is placed at a proper location, the diaphragms point toward the left and right sides of the ensemble.

The microphone picks up essentially what listeners would hear if they were in the same room. Instruments at the rear are heard imperceptibly later than those placed closer to the microphone. A complex mixture of direct and reflected sound is heard, and although a listener would scarcely be conscious of these subtle relationships, the brain uses this information to analyze the size of the ensemble and the acoustic conditions of the playing room. *See* HEAR-ING (HUMAN).

The successful use of this technique demands an acoustically excellent concert hall or recording studio, and an ensemble that is musically well balanced. The recording engineer has little or no technical control over the recorded sound, beyond moving the microphone device to a more favorable location. *See* SOUND-RECORDING STUDIO.

Binaural sound recording. In this system, two microphones are placed on either side of a small acoustic baffle in an effort to duplicate the human listening condition. The recording is played back over headphones, so that the microphones are, in effect, an extension of the listener's hearing mechanism. Although the reproduced program is most faithful to the original, headphone listening restricts the listener's freedom to move about and excludes other

ers from hearing the program. Loudspeakers are, of course, more widely used, yet the distance between them, as well as their distance from the listener, somewhat defeats the binaural effect, and other stereophonic microphone placements are apt to prove more satisfactory.

Multitrack sound recording. To give the recording engineer more technical control over the recording medium, many recordings are now made using a multiple-microphone technique. In place of the stereo microphone, one or more microphones are located close to each instrument or group of instruments. *See* MICROPHONE.

In the control room the engineer mixes the outputs of all microphones to achieve the desired musical balance. As a logical extension of this technique, the microphone outputs may not be mixed at the time of the recording, but may be routed to 24 or more tracks on a tape recorder, for mixing at a later date.

When many microphones are so used, each instrument in the ensemble is recorded—for all practical purposes—simultaneously. The complex time delay and acoustic relationships within the room are lost in the recording process, and the listener hears the entire ensemble from one perspective only, as though he or she were as close to each instrument as is each microphone. Electronic signal processing devices may not be entirely successful in restoring the missing information, and the listener hears a recording that may be technically well executed, yet lacking the apparent depth and musical cohesiveness of the original.

However, the multitrack technique becomes advantageous when it is impractical or impossible to record the entire ensemble at once stereophonically. Especially in the case of new music, it may be impractical to spend the amount of recording time required to achieve an ideal balance. By recording the various sections on separate tracks, the engineer may experiment with balances later on, after the musicians have left.

In contemporary popular music, each member of a small group may wish to play or sing more than one part. For example, a basic rhythm section may be recorded, with each instrument routed to a separate track on the multitrack tape recorder. To keep the instruments acoustically isolated from each other, the engineer may place the microphones as close up as possible. For further isolation, acoustic baffles, or goboes, may be placed between instruments, and particularly subdued instruments may be placed within isolation booths.

As a consequence of this technique, each microphone picks up only the direct sound of the instrument in front of it. The echo and reverberation within the room, and the sound of the other instruments, are not heard. Therefore, by processing the output of any one microphone or tape recorder track, the engineer may make major changes in the sound of a musical instrument without affecting any of the others.

After the basic rhythm section has been recorded,



Fig. 1. Adding tracks to an unused recorder tape.

the group may wish to add some additional parts. Or at a later date an arranger may write a string or brass accompaniment, after listening to the rhythm tracks. In either case, the musicians listen over headphones to the material already recorded and play along in accompaniment. The accompaniment is recorded on one or more of the previously unused tracks (**Fig. 1**).

Since the tape recorder's playback head is some distance from the record head, there is a slight time delay after the tape passes the record head until it reaches the playback head. If the musicians listened to the previously recorded program via the playback head, this time delay would place the new material on the tape out of synchronization with the original program. Consequently, the record head is used for monitoring whenever new material is being added. Thus it performs a dual function simultaneously: playback of the previously recorded program, and recording of the new material on previously unused tracks.

Multitrack recording console. A modern recording console gives the engineer the flexibility required to meet the demands of multitrack sound recording. At the console, microphones may be quickly routed to any track on the tape recorder, and just as quickly reassigned to other tracks as the recording progresses. Since the musicians in the studio may not be able to hear each other due to the acoustic isolation between instruments, the console provides a studio headphone monitoring system, which the engineer may regulate to give the musicians the balance re-

quired. Previously recorded material is also routed through the console to the musicians' headphones when the performers are to play along in accompaniment. Although any of this material can be monitored over a loudspeaker in the studio, headphones are preferred so that the microphones do not again pick up the previously recorded program.

In the control room the engineer may wish to hear a somewhat different balance, and the console permits this flexibility. For testing purposes, the engineer may briefly listen to the output of just one microphone, without affecting the recorded program. During playback the balance of the entire program may be adjusted without affecting the recording setup.

Multitrack mixdown session. When all the recording work is completed, the engineer plays back the multitrack tape, mixing and balancing the many separate tracks to produce the desired stereophonic program (**Fig. 2**). During this mixdown session the outputs of the multitrack tape recorder are routed through the console, just as the microphones were during the original recording session. In effect, the multitrack tape now takes the place of the studio musicians, and the various tracks—each with its distinctive musical information—are mixed to create the final master tape. This master tape is used as the first step in the production of phonograph records.

During the mixdown session a wide variety of processing devices may be used to modify the signals that were recorded earlier. An equalizer changes the apparent frequency response of a musical instrument, compressors and limiters restrict the dynamic range, and expanders attenuate or remove low-level signals.

The natural ambience of a large room may be simulated with delay lines and reverberation chambers. One or two electronically delayed echo signals may be mixed with the direct program, while the



Fig. 2. Multitrack sound recording and mixdown session.

reverberation chamber simulates the multiplicity of reflections typical of a large concert hall. Signals that are delayed just slightly, perhaps 25 milliseconds or less, may be mixed with the direct sound to create the illusion of a larger group. The delay is not heard as such, but simulates the slight imprecision of attack that is characteristic of, say, a large group of violins. Excessive use of any of these signal-processing devices may create special effects that the musicians might find impossible to produce on their instruments.

In the case of a monophonic or stereophonic sound recording, the engineer is unable to isolate and modify, or remove entirely, one or more parts of the ensemble. However, the recording may convey an excellent sense of depth, with the complex relationships between instruments well preserved.

At the cost of losing some of the ensemble's spatial relationships within the room, the multitrack sound-recording technique offers the recording engineer an unprecedented degree of technical control over the entire recording and record production process. *See* SOUND-REPRODUCING SYSTEMS. John M. Woram

Bibliography. B. Bartlett, *Practical Recording Techniques*, 2d ed., 1998; J. Borwick, *Sound Recording Practice*, 4th ed., 1996; J. Eargle, *Handbook of Recording Engineering*, 2d ed., 1992; D. M. Huber, *Modern Recording Techniques*, 5th ed., 2001; J. Woram, *Sound Recording Handbook*, 1989; A. J. Zaza, *Mechanics of Sound Recording*, 1991.

Sound-recording studio

An enclosure treated acoustically to provide controlled recording of music or speech. Professional sound-recording studios range from small home studios (such as a converted garage or basement) to large facilities using state-of-the-art equipment. They consist of two rooms, the main studio or stage (performance area) and the audio control room. The stage includes one or more microphones, is acoustically isolated to ensure that external sounds do not interfere with the recording, and can be broken up into small booths to separate individual performers. The audio control room contains the mixing console and recording devices.

Motion picture sound stages. These large enclosures, with a volume of 1,000,000 ft³ (28,000 m³), are used for photography (to allow long shots), video recording, and sound recording. As well as keeping out extraneous noises, the interior walls and roof are treated with sound absorption material covering up to 85% of the total interior surface to achieve a short reverberation time (RT). This assures intelligible speech recording, as well as the realistic simulation of outdoor scenes. The sound mixing person (mixer) is on the floor of the stage. *See* CINEMATOG-RAPHY; OPTICAL RECORDING; REVERBERATION.

Television stages. These are smaller than motion picture sound stages and require separate audio and video control rooms. Emphasis is on dialog; some television stages allow audience participation, such



Fig. 1. Concrete housing for air-conditioning equipment to attenuate both airborne and solid-borne sound transmission to a nearby studio.

as panel shows and short comedies where audience clapping is an essential part of the program. The microphone operators are on the stage with the camera operators. Audio mixing of a video program takes place in the audio control room. Still another enclosure is required for the lighting console. Stages are often built in pairs joined by a passageway, allowing construction on one stage while recordings take place in the other. Sound insulation is provided between these stages so that hammering in one room will not disturb recordings on the other. Sound and video recording and lighting features are carefully rehearsed. Light fixture dimming is controlled electrically, often with high electrical current pulses that necessitate that microphone cables be specially shielded.

HVAC systems. The presence of several hundred people in a television stage requires a high-volume heating, ventilation, and air-conditioning (HVAC) system of low noise level, which is achieved by large ducts purveying air with a velocity less than 500 ft (150 m) per minute. HVAC condensers, chillers, and cooling towers are remote from the television stage in an isolated enclosure at ground level or underground (**Fig. 1**). Mechanical equipment is supported on vibration isolators. Enclosure walls are lined with sound absorption material. Heat removal ventilation may include noise reduction openings. *See* AIR COOL-ING.

Studio floors. Motion picture and television stages have concrete floors for durability, load-bearing capacity, and absence of squeaks. They are isolated from the walls to minimize the transmission of exterior mechanical vibrations. Floors may include drain pipes for rain scenes on the stage, and pits may be built into the floor to allow video recording of simulated below-ground scenes without building up the set.

Figure 2 shows a typical foundation detail of two adjoining television stages separated by a distance of at least 2 ft (0.6 m). Sand and decomposed granite form a cushion for the concrete floors, which are structurally isolated at the perimeters by a 1-in.-thick (2.5-cm) elastic seal. The walls are lined with 2-in.-thick (5-cm) fiberglass sound-absorbent treatment, covered with fireproof facing. The interspace between the stages may be covered with sheet lead. *See* TELEVISION STUDIO.

Music-recording studios. Music-recording studios vary in size up to 100,000 ft³ (2800 m³), the largest



Fig. 2. Adjoining television stages. (a) Vertical cross section. (b) Top view.

being for the recording of symphonies and opera music by large recording companies. However, the majority of music-recording studios are owned and operated by smaller independent producers, or are rented studios or empty theaters. The reverberation time will vary. The recording microphones may be placed near the artists for individual clarity, or hung at strategic locations just above and in front of a stage.

Multitrack recording. Multitrack magnetic recording appeared first as only 3-track audio recordings, then 4-track, then 8-track. Now some of the larger recording companies employ 32 and 64 tracks. Each track contains the sound from only one or two musicians. The conductor is replaced by the recording engineer, who mixes the many music parts in the audio control room. Concomitant with this type of recording is the installation of drum cages, piano chambers, vocal booths, trumpet stalls, and various other cubicles to achieve clean individual section tracks. A number of adjacent enclosures have windows in the partitions for performers to see each other to keep in time. *See* MAGNETIC RECORDING.

Figure 3 illustrates an audio/video recording studio that allows the mixer a direct view of each instrument section. The video mixers have only a partial direct view. They watch their program material over a closed-circuit television circuit anyway for best videography.

Special acoustical devices. Special acoustical features, such as sound traps and bass traps, may be installed in small recording studios to achieve distinction in recording music for records, audio cassettes, video cassettes, and compact disks (CDs).

Built in a pit between adjoining instrument sections, a sound trap (Fig. 4) replaces a barrier that would not allow neighboring performers to see each other. The trap is generally filled with 2-in.-thick (5-cm) fiberglass panels spaced 1 in. (2.5 cm) apart. When a performer sits over a sound trap 3 to 4 ft (1 to 1.3 m) square, much of the sound from the instrument is absorbed into the trap due to the low sound pressure over the trap compared to



Fig. 3. Plan of audio/video recording studio for multitrack recording.



Fig. 4. Sound trap used to minimize reflections from overhead instrument to adjacent instrument microphones. Diffracted sound rays are absorbed into the pit where sound pressure is low.

the adjoining floor. Trap depth *D* should be a quarter wavelength of the lowest frequency to be absorbed. The trap is covered by a metal grid to support the performer. *See* SOUND.

An entire studio can be built from such sound traps and still have sound reflective walls, floor, and ceiling. Both the traps and the reflective surfaces are practically frequency-independent in their absorption and reflection characteristics so that such enclosures do not exhibit the coloration (spectral amplitude variations) from sound reflected from room boundaries.

A bass trap absorbs sound below about 100 Hz. This may be a sound trap deep enough to absorb lowfrequency sound, a resonant absorber (Helmholtz chamber), or a thin panel that resonates on the layer of air trapped between it and a rigid wall near which it is placed. A common drywall on studs will act as a bass trap in the range 80–160 Hz.

Sidewall sections. Another device used in music recording is acoustically variable sidewall sections. These allow parts of a wall in a rectangular room to be turned to be either sound absorbent or reflective without producing discrete or structured echoes (**Fig. 5**).

Control room mixing and audition. A control room is used to monitor a musical program performed in a studio. A modern control room also serves as a venue for the mixer to later adjust a multitrack recording for volume levels, frequency response (equalization), and reverberatory notes, and to expand or compress certain musical passages. The mixer does this by listening to the program reproduced by one or more loudspeakers in the control room and therefore assumes the role of the conductor. The three optimal acoustic treatment theories for such a control room are live-end-dead-end (LEDE), EDEL (LEDE spelled backwards), and MOD (moderate).

For LEDE, a dead or absorbent treatment is placed on surfaces near the loudspeaker. The opposite end of the room is made reflective or diffusive (to scatter sound) [**Fig.** 6*a*]. This arrangement is intended to prevent coloration of the reproduced program by first-order reflections from the sidewalls and ceiling about the loudspeaker. It is, of course, impossible to prevent first-order low-frequency reflections from the window below the loudspeaker, but the originators of the system claim that these reflections are not important for the purpose of mixing sound tracks.

For EDEL, the acoustic treatment has the same allotment of hard and soft surfaces, but with the absorbent end behind the mixer (Fig. 6b). Reflective or diffusive surfaces are placed near the loudspeaker and above the mixing console in the control room, while the rear wall is made absorptive.

For MOD, the acoustic treatment comprises a uniform disposition of the sound-absorbent material all about the control room.

There is little consensus whether LEDE, EDEL, or MOD treatment is best for the mixer. This is understandable, since the reverberation time in such a room is generally very short, approximately 0.1 (log V), where V is the volume of the control room in



Fig. 5. Variable sidewall treatments. (a) Rotatable cylinders. (b) Rotatable triangular prisms. (c) Rotatable cylindrical sections. (d) Folding splays. (e) One stationary and one sliding perforated panel. (f) Retractable curtains. (g) Rotatable flat panels.

cubic feet (1 $m^3 = 35$ ft³). This reverberation time may be only 0.4 second for a large control room of 10,000 ft³ (280 m³). None of these treatments will improve the intelligibility of speech, singing, or music. Rather, each treatment supports subjective preferences that may achieve the individual esthetic goal of the mixer.

ADR room. The automatic dialog replacement (ADR) or postsynchronizing room (**Fig. 7**), traditionally of the motion picture industry, also has wide use in television production. The performer watches himself or herself on the screen and then repeats the original speech with lip synchronism. ADR employs forward-and-backward transportation of a film or magnetic tape sound track so that short scene bites can be repeatedly reproduced visually and aurally for an actor who did not intelligibly record his or her speech during original filming or video recording. This may have been because the performer had



Fig. 6. Control rooms. Top and side views are shown for each type of room. (a) LEDE room with live rear wall and sound-absorbent environment at the frontal sidewalls to prevent lateral reflections from mixing with the direct signals from the loudspeaker. (b) EDEL room with absorbent rear wall and live front end as in conventional listening locales.

laryngitis or because unavoidable noise was present, such as seaside surf or nearby vehicular or aircraft traffic.

To achieve realism in the recording, it is neces-

sary to credibly recreate the ambiance (acoustical liveness) and the soundscape (original background sound) where the original dialog took place. This can be done by placing the actor in an acoustically

Τ hinged reversible vocal booth panels actor microphone audio mixing boardwalk console audio mixing console gravel water sand screen parquet dance floor rice nails film or videotape projection room Foley pits

Fig. 7. ADR (automatic dialog replacement) or postsynchronizing studio with Foley pits containing various materials for generating sound effects.

(a)

It is frequently necessary to add synthetic sound effects to a scene. When played back, the original recording of a slamming door or a pistol shot may not sound natural. ADR rooms are often equipped with Foley pits (Fig. 7) containing materials to imitate sound effects that could not be recorded well during filming or video recording, or devices that provide realistic sound signals for the events pictured on the screen. Such devices may include a boardwalk to allow the recording of footstep sounds on wood when the original scene was performed on a stage with a concrete floor. A parquet dance floor is used to duplicate footsteps shown on the screen, especially in tap dancing, where the dancer listens to the music over headphones, so that his or her recorded footstep signals are not marred by extraneous music. Soundscapes such as birds or surf may be added from a library of environmental sounds. Either mixing console shown in Fig. 7 may be used, depending on the mixer's preference or requirements for recording the sound of a scene.

Optimal studio room shape. Sound recording studios are often highly sound absorbent ("dead"), in which case room shape is unimportant. However, for studios that cannot be made dead, or where some reverberance is desirable, room shape is critical. Small rooms of regular shape (cubes and squares) have redundant resonances (degenerate modes) that will color sound reverberating in that room. Table 1 shows some acoustically recommended dimension ratios to minimize the number of degenerate modes. Such rectangular rooms require treatment for flutter echoes. A recording studio having the harmonic ratio of 1:2:3 will provide the greatest floor area compared to other ratios of rooms of equal volume, accommodating the most performers in a given space. A room with a ratio of 2:3:5 has the highest ceiling compared to the other cited ratios, providing a maximum number of sidewall reflections (diffusion) and hence good stereophonic sound reception, as may be desired for a concert hall. Adherence to any of these ratios is more important for reverberant rooms than it is for studios having the prescribed short reverberation time of 0.1 (log V).

Reverberation time. This is predicted by the Sabine formula RT = 0.049 V/A, where *V* is the room volume in cubic feet and *A* is the sum total of all the acoustic absorption in the room, measured in sabins. (In metric units, $RT = 0.161 V^1/A^1$, where V^1 is the room volume in cubic meters and A^1 is the total absorption in metric sabins.) In large studios and stages, the reverberation time may be greater than predicted. The reported absorption coefficient, determined from tests on small samples, can be greater than unity, but when the same material is placed over the majority of walls and ceiling areas, the resulting reverberation time will be longer, indicating that the material absorption coefficient is then as low as 0.80.

Sound and television stages must have a very short reverberation time to simulate an outdoor acoustical atmosphere. Music recording studios, ADR chambers, and control rooms may have a more live interior. The recording engineer may wish to change the reverberation time. Performers often prefer a live environment for orchestral or ecclesiastical music. Jazz and rock-and-roll performers prefer a shorter reverberation time. Figure 5 shows ways to change the sound reflectivity of the walls in a music recording studio or an ADR room, thus changing the sound liveness near such walls. These devices minimally affect the reverberation time found in the middle of the room unless they are installed on all surfaces of that room. To change the reverberation time in a large studio, it is necessary to move either one of its walls or its ceiling. This is rarely done because of the expense. See ARCHITECTURAL ACOUSTICS.

Windows. Recording studio windows for control rooms and public viewing must present high sound insulation. They may consist of two thick glass panes widely spaced and sealed to the adjoining wall surface. To avoid first-order sound reflections, they may be triple-pane units, with the center pane made of a 3/4-in.-thick (19-mm) to 7/8-in.-thick (22-mm) noise barrier glass. The outside panes are 1/4-in.-thick (6-mm) Lucite or Plexiglas sheets cold-bent through an arc of 60° with a radius equal to the window width.

Noise control. Recording studios' background noise level must be very low. This is particularly important for digital recording, because of its very wide dynamic range of 80–90 dB. Heating, ventilating, and air-conditioning equipment must not be installed on the roof of such studios. Airflow control or augmenting devices (VAV boxes) must not be located at, over,

TABLE 1. Rooms with recommended ratios of dimensions						
	Rooms nor for equal v	malized olume	Rooms no for equa	ormalized I height		
Name of ratio	Normalized ratio	Relative floor area	Normalized ratio	Relative floor area		
Harmonic Knudsen European Volkmann Golden section	1:2:3 1.08:2.03:2.71 1.1:1.84:2.95 1.14:1.82:2.85 1.12:1.82:2.94	6 5.5 5.43 5.19 5.35	1:2:3 1:1.88:2.5 1:1.67:2.67 1:1.6:2.5 1:1.62:2.62	6 4.7 4.45 4 4.24		
	Name of ratio Name of ratio Harmonic Knudsen European Volkmann Golden section Sabine	Rooms of dimensions Rooms norm for equal v Name of ratio Normalized ratio Harmonic 1:2:3 Knudsen 1.08:2.03:2.71 European 1.1:1.84:2.95 Volkmann 1.14:1.82:2.85 Golden section 1.12:1.82:2.94 Sabine 1.17:1.75:2.92	Rooms normalized for equal volume Normalized ratios of dimensions Name of ratio Normalized ratio Relative floor area Harmonic 1:2:3 6 Knudsen 1.08:2.03:2.71 5.5 European 1.1:1.84:2.95 5.43 Volkmann 1.14:1.82:2.85 5.19 Golden section 1.17:1.75:2.92 5.11	commended ratios of dimensions Rooms normalized for equal volume Rooms normalized for equal volume Rooms nor for equal Normalized Name of ratio Normalized ratio Relative floor area Normalized ratio Harmonic 1:2:3 6 1:2:3 Harmonic 1:2:3 5.5 1:1.88:2.5 European 1.11:1.84:2.95 5.43 11:167:2.67 Volkmann 1.14:1.82:2.85 5.19 11:16:2.5 Golden section 1.12:1.82:2.92 5.11 11:16:2.5		



Fig. 8. Room background noise balanced criterion curves.

or near the studio space. Large HVAC mechanical equipment (boilers, chiller pumps, condensers, and cooling towers) should be located on the ground at a distance from the studio (Fig. 1). It may be necessary to employ a sound attenuator to the outdoor ventilation intake and exhaust openings in that enclosure.

Noise criterion (NC) curves. L. Beranek's balanced NC curves (**Fig. 8**) provide a convenient measure of the annoyance of background noise in buildings. Matching the noise level found in a room to any of these curves is done as a tangent criterion, where the noise background level in every frequency band must not exceed the indicated curve. A dBA level computed from the sum of all bands is often numerically 5 decibels or more greater than the corresponding NCB value. **Table 2** shows the recommended noise criterion balanced (NCB) value for occupied spaces.

Use of the limiting-curve criterion alone can lead to dubious results for low-frequency mechanical equipment noise ("rumble"). Low-frequency sound (and vibration) above 60 dB can cause audible rattles in ceiling and studio fixtures. The noise from an air-conditioning installation could exceed an NCB-30 specification, or the alternative RC-30 specification, by 2 or 3 dB in the octave band whose midfrequency is 1000 Hz and thus be deemed unacceptable, while the spectrum of another installation may hug the NCB-30 at all points on the curve and be declared acceptable. Vibration and discomfort can result from low-frequency sound. Thus, in Fig. 8, additional limits (- - -, moderate; — —, severe) are placed on the lowest-frequency noises. To minimize these problems, the alternative RC curve set with suppressed lower and higher frequency levels is used as an HVAC design tool by mechanical engineers.

Background limit. This may be NCB-15 (about 20 dBA) for high-quality sound recording studios or NCB-25 (30 dBA) for television studios (Fig. 8). For an ordinary studio it should not exceed NCB-20 (25 dBA). When many performers are in the room or photography must take place there, the limit should be NCB-25 (30 dBA). *See* NOISE MEASUREMENT.

Exterior noise. The outdoor noise at a prospective site is surveyed over at least one 24-hour period and over several days if possible. **Figure 9** shows the 24-hour noise history of a prospective studio site. Hours of maximum noise exposure, especially at low frequencies (for example, from diesel locomotive and aircraft passages), are evident. The spectrum of these maximum noise events must be known for airborne sound insulation design of studio exterior walls, roof, doors, and windows.

Wall noise isolation design. Concrete walls are good noise barriers. They may be solid formboard-cast, ground-cast and later raised with a crane (tilt-up walls), or concrete block (CMU). One side of a CMU is plastered to seal joint leakage paths for sound. Solid



Fig. 9. Results of a 24-hour noise exposure level survey at a planned studio site. L_{1} , L_{10} , and L_{50} represent the C-weighted noise level that was exceeded 1, 10, and 50% of the observation period. Dots are aircraft flyover noise maximum values.

TABLE 2. Recommended background noise criteria for

various rooms		
Enclosure	Noise criterion balanced (NCB) value	A-weighted sound level, dB
Concert halls Recording studios	15–20 15–20	25–30 25–30
Theaters	20–25	30-35
Homes	25-30	35-40
Offices	30–35	40-45
Stores	35-40	45-50

concrete walls approximately obey the acoustical mass law for the transmission line (TL) in decibels for a sound of frequency, f,

$$TL = 14.5 \log M + 20 \log f - 31 + 20 \log \cos \theta$$

= 14.5 \log M¹ + 20 \log f - 41 + 20 \log \cos \theta

where *M* is the surface density in lb/ft² (M^1 is the surface density in kg/m²), θ is the angle of sound incidence. For instance, at 500 Hz (log *f* = 2.7) and 0° angle of sound incidence (log cos θ = 0), a 6-in.-thick (15-cm) solid concrete wall weighing 70 lb/ft² (log *M* = 1.85) or 340 kg/m² (log M^1 = 2.53) exhibits a transmission loss of 50 dB.

The enclosing floor, walls, and ceiling of studios located in buildings housing other activities require similar analysis and treatment. Ceiling noise isolation is improved by suspending a grid on vibration isolators that will carry several layers of drywall. The air gap so formed is filled with one or more layers of sound-absorbing insulation. *See* ACOUSTIC NOISE.

Ground vibrations. Studios should be isolated from ground vibrations from trains, subways, and industrial plants. Studios located on upper floor levels are subject to vibrations from mechanical equipment anywhere in the building such as HVAC, transformers, and pumps. Ground vibration and its spectral content is measured with an accelerometer that is either buried in the ground or attached to the affected floor, preferably over a 24-hour period. A studio concrete floor is isolated from these vibrations by "floating" it on pads 2-3 in. thick, on coil springs of appropriate deflection or on air springs over a similar air gap. The resonant frequency of this support and gap must be a quarter or less that of the lowest audio frequency expected to be reproduced, for example, a 10-Hz resonance for a 40-Hz lowest reproduced audio frequency. See ACCELEROMETER; MECHANICAL VIBRATION; SOUND RECORDING; VIBRA-TION ISOLATION. Michael Rettinger; Angelo Campanella

Bibliography. G. G. M. Alkin, Sound Recording and Reproduction, 1997; L. Beranek, Balanced noise criterion (NCB) curves, J. Acous. Soc. Amer., 86(2):650– 664, August 1989; L. L. Beranek and I. L. Ver (eds.), Noise and Vibration Control Engineering: Principles and Applications, 1992; D. Gibson and G. Petersen (eds.), The Art of Mixing: A Visual Guide to Recording, Engineering and Production (Mix Magazine Pro Audio Series), 1997; P. Newell, Studio Monitoring Design: A Personal View, 1995; D. N. Nordantonio, Sound Studio Production Techniques, 1990; A. Nisbett, The Sound Studio, 6th ed., 1995; P. White, Creative Recording II: Microphones, Acoustics, Soundproofing, and Monitoring, vol. 2, 1999.

Sound-reinforcement system

A system that amplifies or enhances the sound of a speaking person, singer, or musical instrument. Sound-reinforcement systems are used to increase the clarity of the original sound as well as to add loudness and reverberation. Those that improve loudness and clarity are often called public address systems, and those that improve reverberation characteristics are called reverberation enhancement systems. Systems for increased loudness are used when listeners are located far from the sound source, as in a large audiorium or outdoors. Systems for increased clarity are used to overcome the effects of unwanted noise or excessive reverberation in a room (such as concert halls and churches with reverberation characteristics designed for music). Reverberation enhancement systems are used to add sound field components, such as reverberation to rooms having unsuitable levels of direct, early reflected sound and reverberant sound or having short reverberation times. See ARCHITECTURAL ACOUSTICS; REVERBERA-TION; SOUND; SOUND FIELD ENHANCEMENT.

Basic elements. The basic elements of a soundreinforcement system are (1) one or more microphones, (2) amplification and electronic signal processing systems, and (3) one or more loudspeakers. The signal processing equipment may include level controls, equalizers or tone controls, systems for prevention of howl and ringing, delays and reverberators, as well as frequency dividing and crossover networks.

Figure 1 illustrates the signal flow diagram of a simple system. In the simplest systems, all elements are joined as one physical unit. An example of a simple public address system is the hand-held battery-driven electronic megaphone, consisting of a microphone, a frequency-response-shaping filter, an amplifier, and a horn loudspeaker (**Fig. 2**). Typical use of the megaphone by police and others is to address a crowd of people. In order to obtain high sound levels in combination with high intelligibility, the frequncy response of a megaphone is limited to save battery power. This may result in poor sound quality.



Fig. 1. Basic sound-reinforcement system.



Fig. 2. Electronic megaphone, with all the basic elements of a sound system in one package.


Fig. 3. Polar directivity patterns of microphones.

The "amplified speaker pulpit" is a stationary and only slightly more complex system. It is a soundreinforcement system built into a pulpit. Because the system is stationary (and most system parameters are known to the designer), it can be tailored to give a very good and natural sound quality. Since the loudspeaker is close to the talker, there is no shift in apparent source location.

Most sound-reinforcement systems, however, are assembled from individual elements to optimize and adapt the system performance for the needs of a particular installation.

Microphones. Microphones often use dynamic, condenser, or electret types of transducers, and may have various polar directivity patterns such as omnidirectional, unidirectional, or bidirectional (**Fig. 3**). In some microphones, it is possible to switch the pattern by remote control. Microphones for use on stage floors, boardroom tables, and so on are boundary microphones with special frequency response characteristics, and their use can reduce or eliminate phase cancellation effects from reflections. *See* ELECTRET TRANSDUCER; TRANSDUCER.

Condenser microphones require a preamplifier close to the microphone element. They may be permanently polarized (electret) or require a polarizing voltage. In either case, dc power must be sent to the preamplifier, usually positive on the signal line pair and negative on the shield. *See* ELECTRET; PREAMPLI-FIER.

Digital microphones use an analog transducer, but integrate analog-digital converters and other components to offer improved possibilities for signal identification, cabling, and routing, making them less susceptible to noise and hum. *See* ANALOG-TO-DIGITAL CONVERTER.

Radio links may substitute for wire to connect system components. The most familiar radio link is wireless microphone. Wearing such a microphone on the body permits the user to move around freely. The wireless microphone usually consists of a miniature omnidirectional microphone (either electret condenser or dynamic type) worn close to the neck on the outside of clothing; a compact FM transmitter, worn inside clothing; an antenna, usually a wire worn along a leg; and a receiver located in a fixed position convenient both for connecting to the rest of the system and for obtaining unobstructed radio pickup from the transmitter (**Fig. 4**). Miniature wireless microphones may be worn on the head and hidden by hair. See FREQUENCY-MODULATION RADIO; MI-CROPHONE.

Signal control equipment. Control and signal processing equipment may range from a single volume control in a preamplifier to complex control boards similar to those used in recording studios, with possibly hundreds of input channels, complex equalizers for every input channel, feedback control equipment, reverberation generation equipment and delay units, matrixed outputs, and so forth. Modern large and complex systems feature preset automatic digital control, with the audio signals determining real-time settings. *See* SOUND-RECORDING STUDIO.

Delay units use analog-digital conversion, digital delay, and digital-analog conversion to delay an input signal. Reverberation may be generated by using finite or infinite impulse response filters, corresponding to the addition of many delayed sound components. Finite impulse response filters use an assembly of delays, and can be based on natural (measured) or calculated room impulse responses. In infinite impulse response filters, some of the delay output signals are fed back to the delay inputs and added in special patterns to reduce coloration in the generated reverberation. Power amplifiers may range in output from 10 watts to thousands of watts, and a system may have only one power amplifier; POWER AMPLIFIER.

Because sound-reinforcement systems add energy to the sound field in a room, there is a possibility that some frequencies will be particularly emphasized due to in-phase feedback of sound to the microphone(s). At a critical gain setting, the system gain will be sufficient to let the system start to oscillate (feedback). The gain-before-feedback is a measure of the loudness the system can produce, with reference to the loudness of the live sound source, before the system starts to oscillate. At lower gain settings, close to oscillation, the system will sound "colored" because sound at some frequencies is exaggerated. The feedback also creates reverberation, that is, prolongs the decay time of the signal. These effects are subjectively annoying. Feedback control equipment may be based on techniques such as (1) adaptive



Fig. 4. Elements of a wireless microphone.



Fig. 5. Loudspeakers used in sound-reinforcement systems. (a) Direct-radiator loudspeaker in a cabinet. (b) Multiple direct radiators in a vertical line array or column. (c) Straight-axis horn loudspeaker. (d) Folded or reentrant horn loudspeaker. (e) Two-way horn loudspeaker system with a single high-frequency straight directional horn and a separate low-frequency horn enclosure with one or two bass loudspeakers.

notch filters, that is, very sharp (narrow-band) rejection filters, (2) frequency shifting, or (3) time-varying delay or amplification. All these techniques may introduce audible artifacts into the sound if not properly used, but may add up to approximately 6 dB of gain in a system. *See* GAIN; LOUDNESS.

Loudspeakers. Loudspeakers may be direct radiators or horn-loaded drivers. The direct radiators are usually cone-type loudspeakers, but are sometimes bending-wave-transducer, electrostatic, or other unconventional types. Horn loudspeakers use the reentrant horn for simple low-quality paging; diffraction horns for somewhat better-quality paging; frontloaded directional horns of various types, including the constant directivity horns (which attempt to provide the same coverage characteristics over a wide frequency range and which have largely displaced the older multicellular and radial horns in high-quality applications); and horn bass enclosures, which often employ folds or bends for compactness, or else supplement the horn loading with ports to bring up bass response (Fig. 5).

A particular loudspeaker system may consist of a single-radiator loudspeaker, a series of direct-radiator loudspeakers in either horizontal or radial lines (vertical line-source loudspeaker systems are often called column loudspeakers), combinations of direct radiators for bass radiation and horn loudspeakers for treble radiation, or large enclosures with large lowfrequency loudspeakers, used with high-frequency horns (Fig. 5). Some systems used for high-level music reinforcement divide the frequency range into four parts, with subwoofers for the lowest bass frequencies, a second set of bass loudspeakers for midbass frequencies, midrange constant-directivity horns, and supertweeters for the extreme highfrequency end of the audio spectrum. Each of these loudspeakers may be driven by separate amplifiers, which are fed by a four-way crossover network (frequency dividing network). The use of separate amplifiers for different portions of the sound spectrum permits the lowest distortion and the greatest flexibility for adjustment. *See* LOUDSPEAKER; SOUND-REPRODUCING SYSTEMS.

Loudspeaker placement and selection. Fixed sound systems are of two basic types: central and distributed. The central system employs loudspeakers reasonably close to the source of live sound (but not so close that feedback is an unsolvable problem), while the distributed system has many loudspeakers fairly close to the audience. For central systems to work properly, the performing area must be defined, as in a typical auditorium or concert hall. Spaces where the sound source may be located anywhere, such as in hotel ballrooms where the head table might be along any wall, are usually best served by some type of distributed system.

Central systems. The most common type of central system employs a loudspeaker or loudspeaker system (sometimes called a cluster) directly over the front of the stage, with the directional characteristics of the loudspeaker or system chosen to cover the audience with good uniformity without directing too much energy toward the stage. Such a system is usually the first choice for most sound system designers because it can be installed most simply and usually provides the greatest intelligibility and naturalness (**Fig. 6**).

Variations on the central system include the use of vertical column loudspeakers on each side of the stage, the use of a horizontal line source along the top of the stage, and the use of directional horn loudspeakers on each side of the stage. All these solutions have their own peculiar problems. Any of the



Fig. 6. Central cluster loudspeaker systems. (a) System in a theater. The constant-coverage high-frequency horns are below the supporting frame, and a two-loudspeaker straight-horn low-frequency enclosure is above (*courtesy of Amir Pishdad*). (b) System for a gymnasium. Ten multicellular high-frequency horns are below two double-loudspeaker straight-horn bass enclosures.

split-source methods (with loudspeakers on each side of the stage) can result in alternate zones of cancellation and reinforcement near the centerline of an auditorium. Modern column loudspeakers may offer steerable array technology, using digital signal processing so that the radiation pattern at various frequencies can be tailored to give optimal coverage over a wide area.

Distributed systems. The most common type of distributed system employs loudspeakers mounted in the ceiling, usually directed straight down. Less frequently used variations include loudspeakers in seat backs (the "pew-back" approach, most often used in churches) or in building columns. Horizontal rows of closely spaced loudspeakers have been used from one side of a theater to the other above the proscenium, as well as down the length of the apex of the ceilings of Mormon churches for reasonably uniform coverage in spaces without excessive reverberation. Loudspeakers in side walls are best used for adding reverberance in a room, and for surround sound and specific effects in cinema and theater. Low ceilings and chandelier systems require wide-dispersion loudspeakers, usually about 100-mm-diameter (4-in.) or larger coaxial (small high-frequency cone and larger bass cone) loudspeakers; however, above 10 m (33 ft) height directional horns are often required.

Reduction of reverberation. With proper placement and selection of loudspeakers, sound systems in live rooms can actually reduce the apparent reverberation by concentrating the amplified sound energy into the sound-absorbing audience or seats, with minimum energy directed to wall and ceiling surfaces. The constant-directivity horns have particular

applications in central systems used for this purpose. Such systems can provide high speech intelligibility, with naturalness, in highly reverberant spaces that have poor speech acoustics.

Enhancement of reverberation. In a system consisting of many sound reinforcement channels, one can adjust the gain so that each channel has low gain and coloration. The combined effect of the system will yield a prolonged reverberation time over the entire operating frequency range of the channels. Systems with up to a hundred channels are sometimes used for reverberation enhancement purposes in rooms perceived to be acoustically "dead". The subjective perception of reverberation may also be enhanced by adding reverberation into the signals using a reverberator in the signal-processing system. Unless this reverberation is radiated incoherently by the use of many loudspeakers in different directions from the listener, the reverberation will be insufficiently diffuse and inadequate. Sometimes reverberation is added selectively to a singer's voice to enhance the quality.

Signal processing. In addition to proper loudspeaker placement and selection, the design of a sound-reinforcement system with the desired characteristics requires the use of signal-processing techniques such as signal delay and equalization.

Signal delay. Digital signal delay equipment is now often used with distributed systems. This equipment places the electronic audio signal in a memory device, producing it in time for the appropriate loud-speakers to be synchronized with the live sound from the performance area.

Haas or precedence effect. Use of this effect can contribute greatly to naturalness. If the live signal from a talker or singer is heard first and the amplified signal arrives up to 25-30 milliseconds later and is not more than 10 dB louder, the listener will determine the direction of the composite signal as coming from the talker or singer. That is, the listener will not be aware that the amplified signal is coming from a different direction. This effect can be used with central systems by proper location of the loudspeakers. In some central systems and most distributed systems, use of the effect requires both proper loudspeaker placement and proper selection of delays.

Equalization. The equalization process can also contribute to the naturalness of a sound system and to its gain-before-feedback in many situations. Before equalization, almost every sound system exhibits some frequency response irregularities that are a departure from an ideal frequency response curve. In the past, such irregularities (due to deficiencies in either the room or the loudspeakers) were accepted. Now, broadband or narrowband equalization usually is applied to a sound system to ensure both the smoothest possible response and adequate gain-before-feedback. Various types of equalizers and frequency response analysis equipment are used in the equalization process. The most widely used appear to be octave-band or third-octave-band equalizers that reduce (or sometimes boost) the audio electrical signal in adjacent bands, and audio analyzers that display the relative acoustical signal level in the same bands via a visual readout. Such equipment is often used with a "pink" noise generator, giving a random electrical signal that has the same signal level in each band, during the analysis and adjustment process. In some cases, the equalization is supplemented by the use of (adaptive) notch filters to prevent oscillation. Equipment in digital mode can continually analyze signals to provide real-time adaptive equalization. See EQUALIZER.

Time-energy-frequency analyzers. Both the setting of correct signal delays and proper equalization can benefit from the use of analyzers, which display the energy of a signal as a function of time and frequency. Simple audio analyzers present the total sound energy arriving at the microphone with the sound system excited by a continuous random-noise generator. Modern analyzers, however, excite the system with a particular pulse, a pseudo-random noise, or a pure-tone signal swept in frequency. The analysis equipment uses information about the signal emitted to calculate the response as a function of time and frequency. This technique allows the time alignment of two separate signals to set delays and locate loudspeakers, and allows independent evaluation of the frequency response of the direct signal (the signal from the loudspeaker without reflections) and the various components of room sound (the reflected energy). There are situations where equalization of the direct sound is preferred to equalization of the total sound field. Even with modern equipment, the ear is the final judge of any equalization scheme.

Power requirements. The size of the power amplifiers needed for an indoor sound system depends on the acoustics of the room served, the sound pressure

levels needed (much less for a speech-reinforcement system than for a music system), and the efficiency of the loudspeakers used. Direct-radiator loudspeakers may be only 1% efficient in converting electrical power to acoustical power, while certain types of horn loudspeakers may be as high as 25% efficient. Generally, low-frequency loudspeakers are less efficient than high-frequency components in a multiway system; thus, if electronic crossovers are used, larger amplifiers are needed for low frequencies than for high frequencies. A small church might use only 30-50 W of power amplification for speech reinforcement, while a small night club with less efficient loudspeakers and higher sound levels might require several thousand watts. Temporary outdoor systems set up for rock concerts may run into the hundreds of thousands of watts, as extremely high levels are produced over long distances for hundreds of thousands of listeners. Electronic reverberation and surround systems, particularly for organ and for orchestra, require tremendous amounts of bass power and large areas of bass-loudspeaker radiating surface to sound convincing. This also holds true for all outdoor music reinforcement. See SOUND PRESSURE.

Outdoor systems. Outdoor systems (**Fig. 7**) are similar to indoor systems, without the reflections that



Fig. 7. Cluster of high-frequency horns at the base of the loudspeaker tower at a stadium, ensuring even coverage and the projection of sufficient sound power to produce 100-dB levels 700 ft (210 m) away at home plate. The mouths of the horns are protected by mesh.



Fig. 8. Large control console used for both reinforcement and recording at a theater. The window enables the operator to hear the sound in the theater. (Courtesy of Amir Pishdad)

result from interior room surfaces. (The reflections in indoor systems produce reverberation, which can be a negative or positive effect, depending on the nature of the program. Reflections also can result in greater uniformity of sound pressure level throughout the audience area to be covered.) Outdoor systems may have to cope with such ambient noise problems as aircraft flyovers, nearby rail lines or highways, and factory noise. This limits the dynamic range, or range of loudness, from outdoor systems such that the softest passages of a musical work must be brought up in loudness to override intruding noise, while the loudest passages cannot be so loud as to bother some listeners.

Care must be taken in outdoor facilities to prevent the sound system from becoming a nuisance to the community. For example, outdoor rock concerts should not take place in the midst of populated areas. The use of directional loudspeakers (such as constant-directivity high-frequency horns), the proper placement and orientation of these loudspeakers, and the use of architectural barriers (such as walls and berms) can minimize the annoyance created by the use of an outdoor sound system.

System development. Development of a successful sound-reinforcement system begins with analysis of the uses of the system: just who or what requires reinforcement, and what is the spatial relationship of the live sound source with respect to the audience. Next comes the acoustical, architectural, and electronic design, including choice of a central or distributed system, locations for loudspeakers, their orientation, choice of microphones, preparation of a signal-flow chart which is developed into a line diagram for the electronic portion of the system, and proper selection of a sound control position if one is necessary.

The design of a system may be greatly simplified through the use of acoustic simulation by computer software—based on the image source or ray tracing methods—to optimize the design with respect to loudspeaker directional characteristics, room geometry, and acoustic conditions. Using auralization (audible sound field simulation based on such simulation), it is possible to listen binaurally to a simulation of the sound-reinforcement system's performance in advance of its being installed. *See* BINAURAL SOUND SYSTEM; VIRTUAL ACOUSTICS.

Often the sound-reinforcement installation must be put to competitive bidding, and this process requires detailed specifications indicating the overall performance expected from the system and from critical individual components, the system's functional requirements, wiring practices, details of installation, documentation, and so forth. This leads to bid review by an impartial expert, who may have designed the system. Careful installation is necessary, followed by adjustment and equalization, in turn followed by acceptance tests by the owner's representatives. Instruction of operators takes place at this time and is critical to the performance of the system. Good service can keep a high-quality system in operation for periods of 20 years or more. Usually service is best provided by the installing contractor.

System control. The sound system operator should be located where it is possible to hear the same mix of live and amplified sound as heard by the audience. Backstage is a poor location for the controls of a theater sound-reinforcement system. Giving the operator a monitor loudspeaker in the control booth, or even a "dummy head" (mannequin with two "ear" microphones) feeding headphones, can never be as good as a large open window in a correctly located control booth (Fig. 8). The use of binaural technology may help in controlling the sound in a large venue by the use of many dummy heads. Many performing arts centers have cockpits in the middle of the audience area, placing the sound system operator directly in the audience. Compromise positions at the rear or at the side are more frequent and can also work. A frequent control position in churches is at a balcony rail.

Several types of automatic mixers have been developed. These devices recognize which microphone is being addressed, increase the amplification from that microphone while suppressing the signal from others, and reduce the gain appropriately when two microphones are being addressed to keep the system from feeding back. The best of these devices also provide overall gain control, reducing the amplification from the loudest voices while bringing up amplification from the softest. For a simple church speech-reinforcement system, one of these devices can replace a human operator with good results. As sophistication of the equipment and techniques of use have improved, automatic mixers in theaters assist, but do not replace, the human operator.

David Lloyd Klepper; Mendel Kleiner Bibliography. D. Davis and C. Davis, *Sound System Engineering*, 2d ed., 1997; G. Davis and R. Jones, *The Sound Reinforcement Handbook*, 2d ed., 1990; J. Eargle, *JBL Sound System Design Reference Manual*, 1999; D. L. Klepper, *Sound Reinforcement:* An Anthology, vol. 1, 1978, vol. 2, 1996; H. F. Olson, Acoustical Engineering, 1957; S. H. Stark, Live Sound Reinforcement: A Comprehensive Guide to P.A. and Music Reinforcement Systems and Technology, 2d ed., 1996.

Sound-reproducing systems

Systems that attempt to reconstruct some or all of the audible dimensions of an acoustic event that occurred elsewhere. A sound-reproducing system includes the functions of capturing sounds with microphones, manipulating those sounds using elaborate electronic mixing consoles and signal processors, and then storing the sounds for reproduction at later times and different places.

The design of audio products has been progressively more scientific, thanks to a better understanding of psychoacoustics, the relationship between what we measure and what we hear. Such investigations can be productive only if it is possible to elicit consistent opinions from individual listeners and agreement among numbers of different listeners. Although some with deteriorated hearing certainly do "hear differently," the majority of listeners, given some practice and the opportunity to form opinions about sound in circumstances free from biases (such as seeing the products, loudness differences, and variable program sources), are remarkably consistent judges of sound quality. *See* PSYCHOACOUSTICS.

When opinions obtained in a controlled manner are studied, it is found that they correlate strongly with certain measurable quantities. Of all audio components, electronic devices are the closest to technical perfection, exhibiting the smallest audible differences. Transducers present many more problems. Microphones, which convert sound into an electrical signal, exhibit fewer problems than loudspeakers, which do the reverse. This is because microphones are relatively simple devices, and they are small compared to most sound wavelengths. Loudspeakers must be physically larger in order to generate the required sound levels. Because the sound interacts with the acoustical space (a room or automobile interior, for example), as it propagates to the ears, reflected sounds modify and sometimes dominate the audible impressions. The acoustical space is the final component in a sound-reproduction system. See ARCHITECTURAL ACOUSTICS; LOUDSPEAKER; MICROPHONE; TRANSDUCER.

The original event for most recordings occurs in the control room of a recording studio or the dubbing stage of a film studio, where the artistic decisions are made about what is to be stored on disk or tape. These decisions are made by listening through loudspeakers in rooms. The loudspeakers and room used in capturing the sound influence what is recorded. If these are significantly different from those used during the reproduction, the sound will not be heard as it was created. As there are no universal standards governing the design and use of loudspeakers and rooms, there are substantial variations in the qual-



Fig. 1. Loudspeakers influence the sound of recordings while they are being made, and then those recordings are used to evaluate other audio products, including loudspeakers. Subjective evaluation, used alone, is a flawed process.

ity of recordings, as well as how they sound when reproduced through different loudspeakers in different rooms. This is a problem for a sound-reproducing system (**Fig. 1**). *See* SOUND-RECORDING STUDIO.

Technical variables. Sound-reproduction systems deal not only with the sounds of live performers but also with synthesized sounds. The same system can reasonably be expected to do justice to the nuances of chamber music, the dynamics and drama of a symphony, and World War II gunfire reenacted on film. The following technical variables are relevant to the reliable reconstruction of audio events.

Frequency range. Traditionally the audible frequency range has been considered to be 20 Hz to 20 kHz. However, lower frequencies can generate interesting tactile impressions, and it is now argued by some that higher frequencies (even 50 kHz and beyond) add other perceptual nuances. Whether these frequencies are significant, or even exist, is a matter of current debate. Most scientific data indicate that only exceptional persons, in the prime of youth, respond to sounds beyond about 20 kHz. *See* HEARING (HUMAN).

Dynamic range. In humans, this is the range of sound level from the smallest audible sound to the largest sound that can be tolerated. In devices, it is the range from background noise to unacceptable distortion. The contours of equal loudness show that from the threshold of audibility to the onset of discomfort is about 120 decibels at middle and high frequencies. In reality, much less than this is required to give a satisfactory impression of accurate reproduction. When recording a live performance, microphones are not a limiting factor since the best have dynamic ranges in excess of 130 dB. In the recording studio, dynamics are manipulated by electronic gain compensation, so the ultimate limitation is the storage medium. For example, a 16-bit compact disk



Fig. 2. Some important reference sound levels relating to what is recorded and what is heard.

(CD) can exceed 100 dB signal-to-noise ratio, while 24-bit digital media can exceed any reasonable need for dynamic range. Background acoustical noise in studios and concert halls sets the lower limit for recordings, just as background noise in homes and cars sets a lower limit for playback. Quiet concert halls and homes are around 25 dB sound pressure level (SPL) at middle and high frequencies. Crescendoes in music and movies approximate 105 dB, giving a working dynamic range of about 80 dB for fussy listeners and less for casual listeners. Today, this is no challenge for electronics, and it is not even difficult for loudspeakers in homes and cars, except perhaps at very low frequencies where a lot of air must be moved to fill a large room (Fig. 2). See ACOUSTIC NOISE; COMPACT DISK; DECIBEL; DIS-TORTION (ELECTRONIC CIRCUITS); GAIN; LOUDNESS; SIGNAL-TO-NOISE RATIO; SOUND; SOUND PRESSURE.



Fig. 3. Just-audible deviations in a measured frequency response when the program material is a complex orchestration in a reverberant space. From left to right are shown deviations caused by a high-Q, a medium-Q, and a low-Q resonance. The magnitudes of the deviations change little as a function of where they are located on the frequency scale.

Linear distortions—amplitude/frequency response. Variations in amplitude and phase as functions of frequency are known as linear distortions, since (ideally) they do not vary with signal level. In practice they often do, and in the process generate nonlinear distortions. In terms of their importance in preserving the timbre of voices and musical instruments, the amplitude versus frequency characteristic, known commonly as the frequency response, is the dominant factor. Humans are very sensitive to variations in frequency response, the amount depending on the bandwidth of the variation.

It is convenient to discuss this in terms of the quality factor (Q) of resonances, since most such variations are the result of resonances in loudspeaker transducers or enclosures. High-Q resonances (such as Q = 50) have a narrow bandwidth; they are very frequency specific. Such resonances are just audible when their deviation from an otherwise flat frequency response is about 10 dB for a sound that is not very revealing (nonreverberant, multimiked pop or jazz), to something less for complex orchestrations in reverberant spaces, to about 3 dB for pink noise. As the Q drops (wider bandwidth), so do the amplitudes at which the resonances are audible. A broadband Q of 1 is just audible at an amplitude deviation of about 0.3 dB with pink noise (Fig. 3). So, when we speak of frequency responses having deviations of +/- so many dB, over a frequency range, such a statement is meaningless unless there is a curve to show the bandwidth of the variations; a narrow spike can be much larger than a broad bump, for the same audible effect. Critical listeners report that they like the sound of devices with flat, smooth, broadband

frequency responses. *See* DISTORTION (ELECTRONIC CIRCUITS); Q (ELECTRICITY); RESONANCE (ACOUSTICS AND MECHANICS); RESPONSE; REVERBERATION.

Linear distortions-phase/frequency response. A pattern of phase shift accompanies the amplitude variation caused by a resonance. We can assume that this visual clue indicates something potentially audible. However, there is no clear evidence that phase shift, by itself, is an audible problem. Since both amplitude and phase responses must be linear to allow accurate waveform reproduction (Fourier transform), the conclusion is that the reproduction of accurate waveforms is not a requirement for good sound. The reality of everyday listening is that we hear a combination of direct sound and multiple reflections from room boundaries and furniture. Moving either the head or the sound source can cause huge changes in the waveform at the entrance to the ear, yet the perception of timbre is hardly, if at all, changed. See FOURIER SERIES AND TRANSFORMS; PHASE (PERIODIC PHENOMENA).

Nonlinear distortions. These distortions occur when a device behaves differently at different signal levels. The waveform coming out of a distorting device will be different from the one entering it. The difference in shape, translated into the frequency domain, is revealed as new spectral components. If the input waveform is a single frequency (pure tone), the additional spectral components will be seen to have a harmonic relationship to the input signal. Hence they are called harmonic distortion. If two or more tones are applied to the device, nonlinearities will create harmonics of all of the tones (harmonic distortion) and, in addition, more new spectral components that are sum-and-difference multiples of combinations of the tones. These additional components are called intermodulation distortion.

It should be remembered that it is the same nonlinearity that causes all of the distortion, and that the measurement methods described here are simply different ways to quantify the nonlinearity. In reality, we listen only rarely to anything resembling a pure tone, and only occasionally to multiples of pure tones. We listen to much more complicated sounds and, as a result, the distortions that result from their interaction with the nonlinearity are of many forms, continuously changing with time. It is therefore impossible to predict, from our simple distortion measures, just how offensive, or even audible, the unwanted spectral contamination will be.

Multiway loudspeaker systems complicate intermodulation distortion measurements because that form of distortion occurs only when the multiple test signals are reproduced by the same transducer. Thus it is possible for a nonlinear system to yield similar harmonic distortion data but very different intermodulation distortion data, depending on how many crossovers there are in the system and the frequencies at which they occur. For example, a woofer that becomes mildly nonlinear at low frequencies may be acceptable if it is restricted to frequencies below 80 Hz, but annoying if it operates to 1000 Hz where important voice and music frequencies can be modulated by simultaneous strong bass signals.

Perceptually, listeners are aided by a phenomenon called masking, in which loud sounds prevent some less loud sounds from being heard. This means that the music causing the distortion inhibits the ability to hear it. Rough guidelines suggest that, in music, much of the time we can be unaware of distortions measuring in whole percentages, but that occasionally small fractions of a percent can be heard. The only safe guideline seems to be to keep distortion, however it is measured, as low as possible. *See* MASK-ING OF SOUND.

Dispersion, directivity, sound power—of loudspeakers. The preceding technical issues apply to both electronic devices and loudspeakers. However, loudspeakers radiate sound in all directions, so that measurements made at a single point represent only a tiny fraction of the total sound output. In rooms, most of the sound we hear from loudspeakers reaches our ears after reflections from room boundaries and furnishings, meaning that our perceptions may be more influenced by measures of reflected and total sound than by a single measurement, say, on the principal axis.

At low frequencies, common box loudspeakers radiate omnidirectionally. At middle and high frequencies, loudspeakers can be designed to radiate in a predominantly forward direction, bidirectionally (forward and back) with the transducers in phase (the so-called bipole) or out-of-phase (the dipole), omnidirectionally, or any of several other multidirectional possibilities. Many of these designs evolved during the era of two-channel stereo, as means of enhancing the reflected sound field in rooms, compensating for the lack of true surround information. However, the most popular configuration has been loudspeakers with transducers arranged on the front surface, facing the listeners.

To be useful, technical measurements must allow us to anticipate how these loudspeakers might sound in rooms. Consequently, it is necessary to measure sounds radiated in many different directions at points distributed on the surface of an imaginary sphere surrounding the loudspeaker. From these data, it is possible to calculate the direct sound from the loudspeaker to the listener, estimates of strong early reflected sounds from room boundaries, and estimates of later reflected or reverberant sounds. It is also possible to calculate measures of total sound output, regardless to the direction of radiation (sound power) and of the uniformity of directivity as a function of frequency. All of these measured quantities are needed in order to fully evaluate the potential for good sound in rooms. See DIRECTIVITY.

Power compression of loudspeakers. As more power is delivered to a transducer, the voice coil heats up, the resistance increases, and the sound output is lower than it should be. This is called power compression. In systems with woofers, midranges, and tweeters, the different units may heat by different amounts, leading to audibly different spectral balance at different playback levels. Another power-related performance change has to do with the vent or port in bass-reflex woofer systems. At high sound levels, the high air velocity in the port can generate turbulence, causing noise and changing the carefully designed tuning of the system. Thus loudspeaker systems can have quite distinctive sounds at different playback levels. *See* AERODYNAMIC SOUND.

System configurations. In the earliest systems, there was monophonic sound (monaural means one ear, not one channel). In the 1930s, Bell Telephone Laboratories published research results indicating that, to successfully preserve the "auditory perspective" of a frontal sound stage, there were two practical techniques: binaural (meaning two ears) and multichannel.

Binaural. This refers to a sound-reproduction technique in which a dummy head, equipped with microphones in the ear locations, captures the original performance. Listeners then audition the reproduced performance through headphones, with the left and right ears hearing, ideally, what the dummy head "heard." The system is good, but flawed in that sounds that should be out in front (usually the most important sounds) tend to be perceived to be inside, or close to, the head. Some laboratory exercises claim to have solved the problem, but common experience is different. Addressing this limitation, systems have been developed that use two loudspeakers, combined with signal processing to cancel the acoustical crosstalk from the right loudspeaker to the left ear, and vice versa. The geometry of the loudspeakers and listener is fixed. In this mode of listening, sounds that should be behind are sometimes displaced forward, but the front hemisphere can be very satisfactorily reproduced. See BINAURAL SOUND SYSTEM; EARPHONES; VIRTUAL ACOUSTICS.

Multichannel audio. This began with stereophonic systems using two channels, not because two was an optimum number, but because in the 1950s the technology was limited to one modulation for each groove wall of a record. Bell Labs, many years before, had recommended many channels and, as a practical minimum, three (left, center, and right) or, for a single listener, two. Good stereo imaging is possible only for listeners on the axis of symmetry, equidistant from both loudspeakers. At its best, it has been very enjoyable. Stereo auditioned through headphones is much less realistic. *See* MODULATION; STEREOPHONIC SOUND.

Quadraphonic systems. These appeared in the 1970s, and added two more loudspeakers behind the listen-



Fig. 4. Evolution of multichannel audio. (*a*) Two-channel stereophonic reproduction. (*b*) Quadraphonics. (*c*) 5.1-channel "home theater." The larger loudspeaker symbol is a subwoofer. (*d*) 7.1-channel multichannel movie and music entertainment system.

ers, mirroring the ones in front. The most common systems exhibited generous interchannel crosstalk or leakage. To hear the proper spatial perspective, the preferred listening location was restricted to front/back as well as left/right symmetry. The lack of a center channel was a serious omission. The failure to agree on a single standard resulted in the system's demise.

Dolby Surround. For film sound applications, Dolby modified the matrix technology underlying one of the quadraphonic systems, rearranging the four matrixed channels into a left, center, right, and surround configuration. In cinemas, the single, limitedbandwidth, surround channel information is sent to several loudspeakers arranged along the side walls and across the back of the audience area. For home reproduction, the surround channel is split between two loudspeakers placed above and to the sides of the listeners. To enhance the spatial effect of the single surround channel, some systems introduced electronic decorrelation between the two loudspeakers, some used multidirectional loudspeakers in the surround channels, and some did both.

Digital discrete multichannel systems. Even with the best active (electronically steered) matrix systems, the channels are not truly discrete. Sounds leak into unintended channels where they dilute and distort directional and spatial impressions. Several competing digital recording systems now provide five discrete full-bandwidth channels, plus a low-frequency special effects channel for movies with truly big bass sounds. For homes, 5.1 channels refers to five satellite loudspeakers, operating with a common low-frequency subwoofer. The subwoofer channel is driven through a bass-management system that combines the low frequencies of all channels, usually crossing over at 80–100 Hz so that, with reasonable care in placement, it cannot be localized.

Movies provided the motivation for multichannel audio in homes, but music may be where it offers its greatest opportunities. Movies tend to limit the use of off-screen sounds mainly to vague spatial effects, ambient music, reverberation, and carefully rationed directional effects. Music, on the other hand, is not constrained. Traditional acoustical music takes advantage of the greater spatial realism that is possible with more channels. Modern music uses the additional channels to surround listeners with moving and static sounds, instruments, and voices. Listeners may find these new directional and spatial effects variously stimulating and annoying. At least now, with more channels and more loudspeakers, what we hear at home and in our cars is likely to more closely resemble what was created in the studio. See MUSICAL ACOUSTICS

The 5.1 channels do not satisfy all listeners. The configuration of three across the front and two at the side/rear locations originated with films. It assumed that all listeners would face a screen. Music listeners are less disciplined in where they sit and in which direction they face. More channels improve the situation and open more creative opportunities (**Fig. 4**).

System components-electronics. The number of signal sources is large and growing. The control device of any multichannel system must accept analog and digital inputs from several stereo and multichannel storage devices and from real-time delivery systems, including terrestrial broadcasting, cable, satellite, and the Internet. Mass storage options now include "juke boxes" holding hundreds of CD's and DVD's, as well as computerlike hard drive storage for compressed audio and video. To this complexity must be added the variety of encoding/decoding options, including different lossy and lossless digital compression algorithms for multichannel audio and digital video. Confusion is common. These are all digital devices and, with enough intelligent programming, a comprehensible user interface should make it all workable. Today's audio/video receivers are much more complicated than the stereo devices of just a few years ago, yet they are easy to use. These, and especially the stand-alone surround processors, are really very powerful computers. Sound quality in electronic devices is currently excellent and is getting even better. See VIDEO DISK.

System components-loudspeakers. The one audio component that remains stubbornly evolutionary is the loudspeaker. In basic form, cone and dome transducers of today look and work remarkably the same as those of decades ago. Other variations in transducers-electrostatic and electromagnetic planar panels, horns, compression drivers, arrays, an so on-have been developed, but none has prevailed. The excellent fidelity, versatility, and cost effectiveness of moving-coil cone and dome transducers justify their widespread acceptance. Better mechanical and acoustical measurements, new engineering design aids, and more accurate performance targets have led to considerable improvements in sound quality. New diaphragm materials allow resonances to be better controlled or even eliminated by forcing them above the highest frequency reproduced by the transducer. Acoustical waveguides tailor the directivity of both direct radiators and compression drivers so that loudspeaker systems radiate sounds more uniformly at all frequencies. The benefit is a greatly improved interaction with rooms, large and small, matching the perceived timbral signatures of the direct, early reflected, and reverberant sounds. See SOUND FIELD ENHANCEMENT; SOUND-REINFORCEMENT SYSTEM.

The marriage of electronics and transducers is a welcome trend, as the performance of the entire system can then be optimized. Designers have more scope to perfect the crossovers between transducers in multiway systems. Digital techniques can refine the performance of the transducers by reducing residual linear and nonlinear distortions. Built-in equalization can compensate for acoustical problems specific to the listening room. Supplying the signals through a digital connection makes it possible to link the loudspeakers into a whole-room or whole-house plug-and-play audio/video network. All of these functions have been demonstrated. It remains to be seen which will become commonplace.

Bibliography. K. B. Benson (ed.), Audio Engineering Handbook, 1988; J. Borwick (ed.), Loudspeaker and Headphone Handbook, 3d ed., 2001; S. E. Olive, B. Castro, and F. E. Toole, A New Laboratory for Evaluating Multichannel Systems and Audio Components, 105th AES Convention Audio Eng. Soc., Preprint 4842, September 1998; J. C. Steinberg and W. B. Snow, Auditory perspective: Physical factors, Elec. Eng., 53:12-17, January 1934; F. E. Toole, The future of stereo, Audio, 81(5):126-142, May 1997, and 8(6):34-39, June 1997; F. E. Toole, Listening tests: Turning opinion into fact, J. Audio Eng. Soc., 30:431-445, June 1982; F. E. Toole, Loudspeaker measurements and their relationship to listener preferences, J. Audio Eng. Soc., 34:227-235, April 1986, and 34:323-348, May 1986; F. E. Toole, Subjective measurements of loudspeaker sound quality and listener performance, J. Audio Eng. Soc., 33:2-32, January/February 1985; F. E. Toole and S. E. Olive, The modification of timbre by resonances: Perception and measurement, J. Audio Eng. Soc., 36:122-142, March 1988.

Sourwood

A deciduous tree, *Oxydendrum arboreum*, of the heath family, indigenous to the southeastern section of the United States, and found from Pennsylvania to Florida and west to Indiana and Louisiana. Sourwood is hardy and often cultivated in the north. It is usually a small or medium-sized tree, but it sometimes attains



Sourwood (*Oxydendrum arboreum*). (a) Branch has urn-shaped flowers. (b) Leaf, twig, and enlargement of twig showing axial bud and leaf scar are below. a height of 75 ft (23 m) and a diameter of nearly 2 ft (0.6 m). The leaves are simple, finely toothed, longpointed, and 4-8 in. (10-20 cm) long, and have an acid taste which explains the popular name. They turn scarlet in the fall. The flowers are white and urnshaped and grow in long clusters (see **illus.**). The dry fruit remains on the tree through the fall and winter. The wood is not used commercially. Sourwood is also known as sorrel tree, and it is widely planted as an ornamental. *See* ERICALES; FOREST AND FORESTRY; TREE. Arthur H. Graves; Kenneth P. Davis

South America

The southernmost of the New World or Western Hemisphere continents, with three-fourths of it lying within the tropics. South America is approximately 4500 mi (7200 km) long and at its greatest width 3000 mi (4800 km). Its area is estimated to be about 7,000,000 mi² (18,000,000 km²). South America has many unique physical features, such as the Earth's longest north-south mountain range (the Andes), highest waterfall (Angel Falls), highest navigable fresh-water lake (Lake Titicaca), and largest expanse of tropical rainforest (Amazonia). The western side of the continent has a deep subduction trench offshore, whereas the eastern continental shelf is more gently sloping and relatively shallow. *See* CON-TINENT.

Regional characteristics. South America has three distinct regions: the relatively young Andes Moun-

tains located parallel to the western coastline, the older Guiana and Brazilian Highlands located near the eastern margins of the continent, and an extensive lowland plains, which occupies the central portion of the continent. The regions have distinct physiographic and biotic features.

The Andes represent the highest and longest north-south mountain system on Earth. Altitudes often exceed 20,000 ft (6000 m) and perpetual snow tops many of the peaks, even along the Equator (Fig. 1). So high are the Andes in the northern half of the continent that few passes lie below 12,000 ft (3600 m). Over most of their length the Andes are not just a single range but two or three parallel ranges. Within the parallel peaks lie a vast series of intermontane basins and plateaus. Ecuador contains a string of 10 such basins. Bolivia has the Altiplano, an extensive basin about 400 mi (640 km) long and up to 200 mi (320 km) wide that is almost entirely surrounded by rugged and lofty peaks. The Altiplano is cold and high, averaging 12,000 ft (3600 m), and is mostly level. The Altiplano is distinctive in that it possesses the highest navigable fresh-water body, Lake Titicaca, measuring approximately 3488 mi2 (8965 km2) at 12,497 ft (3750 m) elevation and approximately 918 ft (275 m) in depth.

On the northeastern and eastern continental periphery lie the Guiana and the Brazilian Highlands. These vast areas of hilly uplands and low mountains are separated from each other by the Amazon River drainage (**Fig. 2**). Together these two plateaus form



Fig. 1. Andes Mountains, Chile. The Andes, which parallel the Pacific coast, constitute one of the Earth's most majestic features. Only the Himalayas exceed them in elevation and mass. (United Press International)



Fig. 2. Representative view of the hilly uplands and low mountain country of the Guiana Highlands. One of five waterfalls on the Canaima River, Gran Sabana in southern Venezuela. (Venezuelan Information Service)

the geologically oldest part of South America. Rocks of ancient igneous and metamorphic origin are partially covered by younger sedimentary beds and sandstones. The vegetation and animal life associated with the flat-topped mesas of the highlands, called *tepuis*, show a high proportion of endemicity (in-situ speciation) resulting from long periods of isolation and ecosystem stability.

More than 65% of South America's total area is characterized by lowland plains under 1000 ft (300 m) in elevation. Some 40% of the continent is less than 650 ft (200 m) above sea level. The plains lie from about 8°N to 40°S between the lofty Andean backbone on the west coast and the Guiana and the Brazilian Highlands on the east coast, and between the Río Orinoco in the north and the Río Colorado in the south. Also in this region are the Llanos of Venezuela and Colombia, the Amazon plain of Brazil, the Paraguayan Chaco, and the Pampas of Argentina. Some of these areas are quite flat while others are undulating. Some, such as the Llanos and Chaco, are alternately flooded and baked.

Physiographic features. The three Andes regions, the lowlands, and the eastern highlands are composed and bounded by many physiographic features, including mountains, intermontane basins, extensive lowland plains, coastlines, and rivers.

Mountains. Because of the vast extent of the Andes, a greater proportion of South America than of any other continent lies above 10,000 ft (3000 m). The young, rugged, folded Andean peaks stand in sharp contrast to the old, worn-down mountains of the eastern highlands. Although the Andes appear to be continuous, most geologists believe that they consist of several structural units, more or less joined. They are a single range in southern Chile, two ranges in Bolivia, and dominantly three ranges in Peru, Ecuador, and Colombia.

Except in Bolivia, where they attain their maximum width of 400 mi (640 km), the Andes are seldom more than 200 mi (320 km) wide. They do not equal the Himalayas in height, but have at least 30 peaks above 20,000 ft (6000 m). The average height of the Andes is estimated to be 13,000 ft (3900 m). However, it is only north of latitude 35°S that the mountains exceed elevations of 10,000 ft (3000 m) [see **table**].

Both active and quiescent volcanic formations are common in southern Colombia, Ecuador, central and southern Peru, and western Bolivia. These volcanic peaks are the surface expression of the subduction of the Nazca tectonic plate beneath the South American plate, forming the offshore Peru-Chile Trench. The Andes were raised up by tectonic forces from 6560 ft (1970 m) to an average 13,120 ft (3940 m) during the Pliocene uplift approximately 5 million years ago, and they continue to rise incrementally each year. As a result of these powerful tectonic forces, there is a spectacular series of volcanic peaks in western Bolivia and on each side of the structural depression in Ecuador. *See* PLATE TECTONICS.

Stratification of climate and vegetation with

altitude can be readily observed in the Andes. At their eastern base is a zone with hot, humid lowland and foothills up to 3300 ft (1000 m), known as *tierra caliente*. *Tierra templada*, the zone from 3300 to 6600 ft (1000 to 2000 m), has a relatively mild, frostfree climate and was preferred for European settlement and the production of plantation crops such as coffee and coca. From 6600 to 13,200 ft (2000 to 4000 m) is *tierra fria*, a montane zone with occasional frosts. *Tierra belada*, at 13,200 to 19,800 ft (4000 to 6000 m), occupies the zone between the daily frost line and the final zone of permanent snow called *tierra nevada*.

Because of the great north-south extent of the Andes, the processes of their erosion and denudation have varied. Southward from about 40°S, and especially in the far south, the Andes were heavily glaciated during the Ice Age, and an extensive area north of the Strait of Magellan still has a broad mantle of permanent ice. Glaciers descend to the heads of many fiords on the Pacific coast or into lakes on the eastern side of the mountains.

Coastal features. From the southern tip of Cape Horn north to 41°S latitude, the western coastal zone consists of a broad chain of islands where a mountainous strip subsided and the ocean invaded its valleys. This is one of the world's finest examples of a fiorded coast. Nowhere along the Pacific coast is there a true coastal plain. South of Arica, Chile, the bold, precipitous coast is broken by only a few deep streams, the majority of which carry no water for

Peak Aconcagua, Argentina Ampato, Peru	Height, ft (m) 22,835 (6960) 21,702 (6615)
Aconcagua, Argentina Ampato, Peru	22,835 (6960) 21,702 (6615)
Ampato, Peru	21,702 (6615)
Caca Aca, Bolivia	20,329 (6196)
Cachi, Argentina	21,326 (6500)
Chimborazo, Ecuador	20,577 (6272)
Cincel, Bolivia	20,102 (6127)
Condoriri, Bolivia	20,043 (6109)
Coropuna, Peru	22,802 (6950)
Cuzco (Ausangate), Peru	20,187 (6153)
Del Acay, Argentina	20,801 (6340)
Dos Conos, Argentina	22,507 (6860)
Falso Azufre, Argentina-Chile	22,277 (6790)
Huascaran, Peru	22,188 (6763)
Illampu, Bolivia	21,276 (6485)
Illimani, Bolivia	21,282 (6487)
Incahuasu, Argentina-Chile	21,720 (6620)
Llullaillaco, Argentina-Chile	22,015 (6710)
Mercedario, Argentina-Chile	21,884 (6670)
Ojos del Salado, Argentina-Chile	22,573 (6880)
Payachata, Bolivia	20,768 (6330)
Pissis, Argentina	22,245 (6780)
Porongos, Argentina-Chile	20,512 (6252)
Pular, Chile	20,342 (6200)
Sajama, Bolivia	21,390 (6520)
Sarmiento, Chile	20,670 (6300)
Socompa, Argentina-Chile	19,787 (6031)
Tocorpuri, Bolivia-Chile	22,163 (6755)
Tortolas, de las, Chile	20,018 (6101)
Tres Cruces, Chile	21,720 (6620)
Turpungato, Chile	21,490 (6550)
Valadero, Argentina	20,735 (6320)

*The elevations are approximate and some are controversial. Data based mostly on tables in *Goode's World Atlas*, 12th ed., Rand McNally, 1964.



Fig. 3. Uplifted beaches on desert coast, east side of San Juan Bay, Departmento de Ica, Peru. The sea cliff is about 100 ft (30 m) high. Each horizontal line represents a former beach. (*Photograph by F. Atchley*)

years at a time. Between Arica and Caldera, Chile, there are no natural harbors and almost no protected anchorages. In fact, South America's coastline is the least indented of all the continents except Africa's. The only exceptions are the fiorded coast and the offshore archipelagoes of southern Chile, the Gulf of Venezuela including Lake Maracaibo in the north, the Gulf of Guayaquil in the west, and the estuaries of the Amazon and the Río de la Plata in the east. *See* FIORD.

Many raised beaches are seen along the coast (Fig. 3), supporting the premise that the Andes are of recent origin and that the forces that raised them are still at work. Nearly every port in coastal Peru is an open roadstead (less enclosed than a harbor). Coastal Ecuador, separated by mountain spurs and a narrow coastal plain east of the Gulf of Guayaquil, varies in width from 50 to 150 mi (80 to 240 km). Two important river systems cross it, the Esmeraldas in the north and the Guayas in the south. But the most important segment of the entire area is the Guayas Lowland, east and north of the Gulf of Guayaquil. Here are found one-quarter of Ecuador's total population, the most important commercial agricultural lands, and the principal port, Guayaquil. The climatic transition that occurs in coastal Ecuador is one of the most pronounced in the world. Within a distance of only 425 mi (680 km), one passes from areas having a rainy tropical climate, to wet and dry tropical, to semiarid tropical, and finally to arid tropicalin short, from the world's rainiest to its driest climates.

The Caribbean coast of Colombia is a lowland formed largely of alluvium, deposited by the Magdalena and Cauca rivers, and bounded by mountains on three sides. In Venezuela, the Central Highlands rise abruptly from the Caribbean, with lowlands around Lake Maracaibo, west of Puerto Cabello, and around the mouth of the Río Tuy of the Port of Guanta. The coastal region of Guyana, Suriname, and French Guiana is a low, swampy alluvial plain 10–30 mi (16–48 km) wide, and as much as 60 mi (96 km) wide along the larger rivers. This coastal plain is being built up by sediments carried by the Amazon to the Atlantic and then deflected westward by the equatorial current and cast upon the shore by the trade winds. There is no broad coastal plain south of the Amazon and east of the Brazilian Highlands to afford easy access to the interior as is characteristic of North America's coast. The rise from the coastal strip to the interior is quite gradual in northeastern Brazil; but southward, between Bahia and Río Grande do Sul, the steep Serra do Mar is a formidable obstacle to transportation. Nonetheless, the first railroad was constructed here connecting São Paulo in the coffee region with the port of Santos. Though only 40 mi (64 km) long, it was a remarkable feat since the abrupt escarpment of the Serra do Mar behind Río de Janeiro posed difficult engineering problems and required the use of cog tracks and cables.

Along coastal Uruguay there is a transition between the hilly uplands and plateaus of Brazil and the flat Pampas of Argentina, whereas coastal Argentina as far south as the Río Colorado, in Patagonia, is an almost featureless plain. In Patagonia, steep cliffs rise from the water's edge. Behind these cliffs lies a succession of dry, flat-topped plateaus, surmounted occasionally by hilly land composed of resistant crystalline rocks. Separating southern Patagonia from Tierra del Fuego is the Strait of Magellan, which is 350 mi (560 km) long and 2–20 mi (3-32 km) wide. Threading through numerous islands, the strait is lined on each side with fiords and mountains.

Majorrivers and river systems. There are three great river systems in South America and a number of important rivers that are not a part of these systems. The largest river system is the Amazon which, with its many tributaries, drains a basin covering 2.7 million square miles (7 million square kilometers), or about 40% of the continent. The next largest is the system composed of the Paraguay, Paraná, and La Plata rivers, the last being a huge estuary. The third largest river system, located in southern Venezuela, is the Orinoco, which drains 365,000 mi² (945,000 km²) of land, emptying into the Atlantic Ocean along the northeast edge of the continent.

Amazon River system. The Amazon River begins in the eastern Andean highlands and stretches across the north-central part of the continent, draining the interior lowlands. Not only is it the longest, but it carries the largest volume of water and silt of any river in the world. It is also navigable for oceangoing vessels for a greater distance than any other river. The Amazon discharges 3.4 billion gallons/min 13 billion liters/min and has an average flow of 7.5 million cubic feet/s (210,000 cubic meters/s). Its discharge is about 18% of that of all rivers of the world. In places, for example, between Obidos, Brazil (500 mi or 800 km upstream) and the mouth, its depth exceeds 300 ft (90 m). It drains an area of 2.3 million square meters (6 million square kilometers), and its gradient averages only 0.2 in./mi (3 mm/km) or about 35 ft (11 m) for the last 2000 mi (3200 km) to the sea. Considering the enormous amount of silt, sand, and alluvial debris the Amazon carries to the Atlantic Ocean, it might be expected that a huge delta would have formed. But the Amazon has no delta because of the steady submergence of the land where the river reaches the sea. The Amazon is joined by some 500 tributaries descending the Andes and the Brazilian and Guiana Highlands.

Paraguay-Paraná-La Plata system. From its headwaters in southwestern Brazil, the Paraguay River courses southward 1300 mi (1780 km), discharging into the Paraná at the southern edge of Paraguay. The western bank of the Paraguay is low, and during the rainy season the western side of the entire basin becomes inundated, flooding thousands of square miles of low-lying country.

The Paraná ranks among the world's major rivers. It is longer and carries more water than the Mississippi River in the United States. Its source in the Brazilian Highlands is 2450 mi (3920 km) from its outlet in the Río de la Plata. The Paraná has cut a deep canyon in the flat-topped plateau of its upper course. As the river drops over the edge of the lava formation, famous waterfalls are found-La Guayra and Iguazú, the latter on the Iguazú, a tributary of the Paraná. However, from Corrientes, where the water of the Paraguay is added, the gradient is gentle. Since the Paraná rises in areas with characteristically heavy summer rainfall, its water volume fluctuates widely, the variation from high to low water reaching 15 ft (4.5 m) in the lower course. At the mouth of the Paraná is an enormous delta, shared with the emptying Uruguay River. The delta consists of numerous low, flat islands that are submerged for weeks at a time.

The largest indentation on the east coast of South America occurs from the head of the estuary of La Plata to the open Atlantic and is almost 200 mi (320 km) long. At Buenos Aires the estuary is about 25 mi (40 km) wide, at Montevideo 60 mi (96 km), and at its mouth 138 mi (221 km). The Paraná and the Uruguay transport huge quantities of silt into La Plata, some of which settles in navigation channels, necessitating costly dredging.

Navigation for vessels drafting 6-7 ft (1.8-2.1 m) is possible as far as Corumba, Brazil, some 1800 mi (88 km) upstream from Buenos Aires via the Paraná and Paraguay rivers. The remainder of the Paraná is rarely used for navigation, and its hydroelectric potential is as yet undeveloped because markets are too distant. The Paraguay, however, serves the interior, and is considered the lifeline to the outside world for Paraguay.

Orinoco River system. The headwaters of the Orinoco River lie in southern Venezuela near the Brazilian border. Near Esmeraldas, in southern Venezuela, is a stretch of about 220 mi (350 km) known as the Casiquiare canal, joining the Orinoco with the Negro of the Amazon system.

The Orinoco, approximately 1600 mi (2560 km) long, is noted for its variability in volume from wet to dry season. It is a broad stream, and is a mile wide at Ciudad Bolivar, Venezuela, the "narrows." Rainfall from June to October is so heavy that the Orinoco and its many tributaries are unable to handle all the water, and enormous areas become inundated. The river rises 39 ft (12 m) at Puerto Ordaz, Venezuela. From January to March, the dry season, the waters recede and only the larger rivers flow freely. The smaller rivers are gradually converted into chains of pools and swamps lying along the valley bottoms. Where the Orinoco flows into the Atlantic, it forms a low, wide delta consisting of small islands and swamps and an intricate maze of distributaries.

Biotic features. The plants and animals of the South American tropics are classified as Neotropical, defined by the separation of the South American and African continents during the Middle Cretaceous (95 million years ago). The Paraná basalt flow, which caps the Brazilian shield in southern Brazil and adjacent parts of Uruguay and Argentina, as well as western Africa, indicates the previous linkage between the South American and African continents. During the Middle Cretaceous Period, angiosperms (flowering plants) expanded in distribution across the Earth's landmasses, largely replacing the gymnosperms, which previously dominated the Gondwana landmass. While there are regions of important endemicity for plants and animals within South America, many evolved early enough to share significant similarities with North American, Eurasian, and African fauna and flora. South America has many biotic environments, including the constantly moist tropical rainforest, seasonally dry deciduous forests and savannas, and high-altitude tundra and glaciated environments.

Tropical rainforest. Amazonia contains the largest extent of tropical rainforest on Earth. It is estimated to encompass up to 20% of the Earth's higher plant species and is a critically important source of fresh water and oxygen. Structurally complex, the rainforest is composed of up to four distinct vertical layers of plants and their associated fauna. The layers often cluster at 10, 20, 98, and 164 ft (3, 6, 30, and 50 m) in height. The highest levels of canopy experience intense competition for light and water. Epiphytes and lianas are often found in the taller trees of this layer. The lower canopy and forest floor are usually open spaces because of the low intensity of light (around 1%) that reaches the forest floor. Over 75% of Amazonian soils are classified as infertile, acidic, or poorly drained, making them undesirable for agriculture because of nutrient deficiencies. Most of the nutrients in the tropical rainforest are quickly absorbed and stored in plant biomass because the high annual rainfall and associated leaching make it impossible to maintain nutrients in the soils. In addition to the high structural complexity of the tropical rainforest, there is considerable horizontal diversity or patchiness. As many as 300 separate species of trees can be found in a square mile (2.6 km²) sample tract of rainforest in Brazil. The high complexity and species diversity of the rainforest are the result of long periods of relative stability in these regions. Many areas of Amazonia show unique concentrations of species that suggest these territories may have acted as Pleistocene refugia where plants and animals were able to survive the glacial advances of higher latitudes. See PLEISTOCENE.

Deciduous and conifer forest. Deciduous forests are found in areas where there is seasonal drought and the trees lose their leaves in order to slow transpiration. The lower slopes of the Andes, central Venezuela, and central Brazil are areas where these formations are found. Conifer forests occur in the higher elevations of the Andes and the higher latitudes of Chile and Argentina. Many of the species of these formations derive from the Late Cretaceous Southern Gondwana flora, including the *Araucaria* (star pine), *Podocarpus*, and *Dacrydium*. *See* DECID-UOUS PLANTS.

Savanna. Tropical savannas occupy an extensive range in northern South America through southeastern Venezuela and eastern Colombia. Temperate savannas are found in Paraguay, Uruguay, the Pampas of Argentina, and to the south, Patagonia. Savannas are composed of a combination of grass and tree species. The climate in these areas is often quite hot with high rates of evapotranspiration and a pronounced dry season. Most of the plants and animals of these zones are drought-adapted and fire-adapted. Tall grasses up to 12 ft (3.5 m) are common as are thorny trees of the Acacia (Fabaceae) family. Many birds and mammals are found in these zones, including anteater, armadillo, capybara (the largest rodent on Earth), deer, jaguar, and numerous species of venomous snake, including rattlesnake and bushmaster (matranare).

Desert. South America is unique in having a westcoast desert that extends almost to the Equator, probably receiving less rain than any on Earth (the Atacama), and an east coast desert located poleward from latitude 40° S (the Patagonian).

The Atacama-Peruvian Desert dominates the Pacific coast for nearly 2000 mi (3200 km), lying between the ocean and the higher slopes of the Andean ranges. In many places, particularly in Chile, the land rises abruptly from the sea to an elevation of 2000-3000 ft (600-900 m). The driest stretch, one of the driest places in the world, is between Arica and Caldera, Chile (Fig. 4). Above the seaward scarp is a low range of hills and a desert trough, 30-50 mi (48-80 km) wide, from which nitrate deposits are extracted. The persistence of this desert through latitudes at which deserts are not found elsewhere is the result of the north-moving Humboldt Current (Peru Current), accompanied along the immediate shore by the upwelling of cold water from considerable depths, creating stable, high-pressure atmospheric



Fig. 4. Port of Matarani on the desert coast of southern Peru. This oblique aerial view looks coastwise southward over one of the world's driest and most barren coastal zones. (*Servicio Aerofotográfico Nacional*)

conditions. Prevailing winds, blowing from the west, carry cold air onto the warmer land with a drying effect.

South America's east-coast desert, the Patagonian, is found between the Río Colorado and the Strait of Magellan. This desert is partly located in the rain shadow of the Andes, and the meager annual rainfall is ordinarily 5–6 in. (13–15 cm). Also contributing to the region's aridity and to the presence of the desert near to the shores of the Atlantic are the cold waters of the Falkland Current, moving northward off the eastern coast, and the ceaseless winds, which hasten evaporation. Wind velocities can exceed 70 mi/h (31 m/s). What precipitation there is occurs mostly in the winter. The region's very cold winter temperatures are the result of high latitude and high elevation.

Tundra. In Bolivia and Peru the zone from 10,000 to 13,000 ft (3000 to 3900 m), though occasionally to 15,000–16,000 ft (4500 to 4800 m), is known as the *puna.* Here the hot days contrast sharply with the cold nights. Above the *puna*, from timberline to snowline, is the *paramo*, a region of broadleaf herbs and grasses found in the highest elevations of Venezuela, Colombia, and Ecuador. Many of the plant species in these environments are similar to those found at lower elevations; however, they grow closer to the ground in order to conserve heat and moisture.

Deforestation and environmental contamination. The 1992 Quincentennial and the 1992 Earth Summit (United Nations Conference on Environment and Development) have drawn widespead attention to environmental policy and economic development in South American countries that have encouraged the destruction of natural vegetation. Intense political discourse has focused on the removal of tropical rainforest and the contamination of both lowland and highland environments across the continent. An estimated 20% of the Amazonian forest has already been destroyed, and more than 7700 mi² (200,000 km²) are estimated to be replaced by agriculture, cattle ranching, and urbanization each year. The loss of tropical rainforest in particular has been widely publicized, although many other biotic environments, including deciduous forest, coastal mangrove, and tundra environments, are also being degraded by human use.

Deborah A. Salazar; C. Langdon White Bibliography. B. Box, *South American Handbook*, 82d ed., 2006; C. Caviedes and G. Knapp, *South America*, 1995; B. Cox and B. Moore, *Biogeography*, 7th ed., 2005; A. Morris, *South America*, 3d ed., 1991; S. E. Place, *Tropical Rainforests*, rev. ed., 2001; A. H. Strahler and A. Strahler, *Introducing Physical Geography*, 4th ed., 2005.

South Pole

That end of the Earth's rotational axis opposite the North Pole. It is the southernmost point on the Earth and one of the two points through which all meridians pass (the North Pole being the other point). This is the geographic pole and is not the same as the south magnetic pole, which is near the coast of Antarctica, south of Tasmania. The South Pole lies inland from the Ross Sea, within the landmass of Antarctica, at an elevation of about 9200 ft (2800 m).

Since the Earth's axis is inclined $23^{1/2}^{\circ}$ from the vertical to the plane of the ecliptic, the South Pole experiences six months when the Sun is continuously above the horizon, reaching a maximum altitude of $23^{1/2}^{\circ}$. It also has 6 months when the Sun remains continuously below the horizon, from March 22 to September 22. During the latter period, however, there is an interval of 7 weeks at the beginning and at the end when twilight conditions preclude total darkness. Twilight prevails when the Sun is located between 0° and 18° below the horizon.

There is no natural way to determine local Sun time because there is no noon position of the Sun, and shadows point north at all times, there being no other direction from the South Pole. *See* MATHEMAT-ICAL GEOGRAPHY; NORTH POLE. Van H. English

Bibliography. A. H. Strahler and A. Strahler, *Introducing Physical Geography*, 4th ed., 2005.

Southeast Asian waters

All the seas between Asia and Australia and the Pacific and the Indian oceans. They form a geographical and oceanographical unit because of their special structure and position, and make up an area of $3,450,000 \text{ mi}^2$ ($8,940,000 \text{ km}^2$), or about 2.5% of the surface of all oceans.

Ocean floor. The ocean floor consists of two large shelves and a number of deep-sea depressions. The Sunda Shelf is covered with littoral sediments, and the submerged valleys of two large river systems are found on it. The Arafura Shelf connects New Guinea with Australia. The China Basin, whose maximal depth is 16,450 ft (5016 m), is connected with the Pacific Ocean over a sill about 6000 ft (2000 m) in depth. The Sulu Basin (maximal depth 18,300 ft or 5580 m) has the highest sill (1400 ft or 420 m). The Celebes Basin (maximal depth 20,400 ft or 6220 m) is connected with the Pacific Ocean over three sills of about 4200 ft (1400 m) each. The Banda Basin is subdivided into several smaller basins, including the Weber Deep of 24,400 ft (7440 m); its sill depth is 6170 ft (1880 m). In the Celebes, Banda, and Flores basins, volcanic ashes are found. See EAST INDIES.

Surface circulation. The surface circulation is completely reversed twice a year (**Figs. 1** and **2**) by the changing monsoon winds.

Monsoon current. During the north monsoon, which is fully developed in February, the monsoon current is formed in the northern China Sea, flows along the coast of Vietnam into the Java Sea, and into the Flores Sea. Parts of its water masses return into the Pacific Ocean; parts sink in the Banda Sea or flow into the Indian Ocean.



Fig. 1. Water circulation in South China and Indonesian seas, February. Transports in $10^6 \ m^3/s.$

During the south monsoon in August, the monsoon current is formed by water flowing from the Java Sea into the South China Sea and farther into the Pacific Ocean.

The transports of the monsoon current are $110,000,000 \text{ ft}^3/\text{s}$ (3,000,000 m^3/s) in August and $180,000,000 \text{ ft}^3/\text{s}$ (5,000,000 m^3/s) in February, but are small compared with those in the adjoining parts of the Pacific and Indian oceans (Figs. 1 and 2).

Indonesian throughflow. Indonesian throughflow is a complex system of currents within the depth range of the thermocline to 1000 m (3300 ft) that allows water to flow from the Pacific to the Indian Ocean. It is the only place in the global oceans where warm



Fig. 2. Water circulation in South China and Indonesian seas, August. Transports in $10^6 \ m^3/s.$

equatorial waters pass from one ocean basin into another. Throughflow water must be replaced in the Pacific by inflowing, colder water at a high latitude in the South Pacific. Consequently, Indonesian throughflow exports a large amount of heat from the Pacific. The shift in heat between ocean basins, and the subsequent release of the heat to the atmosphere over the Indian Ocean, is recognized as a critical component of the global climate system. Consequently, Indonesian throughflow has been the subject of intensive oceanographic research. Attempts to understand why it exists led to fundamental theoretical advances in geophysical fluid dynamics. The studies showed that the mean throughflow is a response to the wind field over the whole Pacific Ocean.

The main path of Indonesian throughflow is from the Celebes Sea near Mindanao, through Makassar Strait, eastward along the Indonesian archipelago through the Flores and Banda seas, and finally into the Indian Ocean passing south of Timor. Some flow passes through the lesser Sunda Islands. The strength of throughflow is probably not larger that 360,000, 000 ft³/s (10,000,000 m³/s), while the temperature difference from the South Pacific is about 10° C (18° F). The large temperature difference is the factor that makes Indonesian throughflow critical in the global climate system. The throughflow is strongest during the south monsoon (Fig. 2), when upwelling water from the Banda Sea contributes to the flow. Its strength varies considerably from one year to the next in response to changing winds over the equatorial Pacific and Indian oceans. The variability is related to the El Niño Southern Oscillation phenomenon. See EL NIÑO.

Salinity and temperature. The monsoons cause a pronounced rainy and dry season over most parts of the region and consequently strong annual variations of the surface salinity. During the north monsoon, rainfall is strongest to the south of the Equator and surface salinity decreases, while in the China Sea the excess of evaporation over rainfall increases salinity. During the south monsoon, conditions are reversed; strong rainfall over the China Sea decreases its salinity, while evaporation and upwelling increase salinity in the Java, Flores, and Banda seas. Regions of permanent low salinities are in the Malacca Straits, the Gulf of Thailand, and the waters between Borneo and Sumatra, because of the discharge of the large rivers. The surface temperature is always high, between 79 and 84°F (26 and 29°C), and drops only in the northern parts of the South China Sea and in the Arafura Sea to about 75° F (24°C) during the dry season. See MONSOON METEOROLOGY.

Deep circulation. The deep basins are supplied with Pacific Ocean deep water that enters over the various sills (**Fig. 3**). Only the Timor and Aroe basins are filled with Indian Ocean deep water. The renewal of water in the deep-sea basins is sufficiently rapid, on the order of 100 years, to maintain an adequate level of oxygen for the deep-sea fauna.

Tides. The tides are mostly of the mixed type. Diurnal tides are found in the Java Sea, Gulf of Tonkin, and Gulf of Thailand. Semidiurnal tides with high amplitudes occur in the Malacca Straits. Strong tidal



Fig. 3. Flow of bottom waters in the eastern part of the Indonesian Archipelago.

currents occur in many of the small shallow passages. *See* INDIAN OCEAN; PACIFIC OCEAN; TIDE.

Klaus Wyrtki; Gary Meyers Geology and geophysics. The Southeast Asian seas are characterized by the presence of numerous major plate boundaries. Plate boundaries represent sites where the Earth's crust is created or destroyed, or where rigid blocks of crust slide past one another. In Southeast Asia the plate boundaries are identified, respectively, by young, small ocean basins with their spreading systems and associated high heat flow; deep-sea trenches and their associated earthquake zones and volcanic chains; and major strike-slip faults such as the Philippine Fault (similar to the San Andreas Fault in California). The Southeast Asian seas are thus composed of a mosaic of about 10 small ocean basins whose boundaries are defined mainly by trenches and volcanic arcs. The dimensions of these basins are much smaller than the basins of the major oceans. The major topographic features of the region are believed to represent the surface expression of plate interactions, the scars left behind on the sea floor. See BASIN; FAULT AND FAULT STRUCTURES.

The proximity of important metalliferous ore deposits in selected islands of Southeast Asia, coupled with the fact that other nearby areas show significant potential for oil and gas, has led many scientists to investigate models that could relate plate tectonic processes to the formation of ore deposits and oil.

Ocean basins. The small ocean basins which make up the Southeast Asian waters are relatively young when compared with the ages of the major ocean basins. The basins of Southeast Asia that have been dated with some confidence are all younger than approximately 50,000,000 years. As part of the Deep Sea Drilling Project, the vessel *Glomar Challenger* drilled several holes along an east-west line (near 18°N) from the central part of the Philippine Sea into the Pacific Basin. One purpose of the drilling program was to determine the age and chemical properties of the underlying sea floor for comparison with the sea floor of the major ocean basins. The ultimate goal of the drilling is to determine the evolutionary processes which have given rise to the extremely complex tectonic features which are observed in the Southeast Asian seas.

Sediments. The sediments that blanket the sea floor in the Southeast Asian seas primarily reflect the proximity of land (that is, sediment sources) and the age and depth of the underlying sea floor. The thickly sedimented areas are found close to the continental margins and island arcs. The sediments here are composed mostly of materials transported from the adjacent continents and islands. It is possible that an additional component of the sediments found near the margins was first deposited on the deep-sea floor but was subsequently uplifted onto the margins when certain plates collided. *See* CONTINENTAL MARGIN; MARINE SEDIMENTS.

Tectonic processes. The study and understanding of the tectonic processes active in this region are of fundamental importance to many unresolved problems of the Earth's history. For example, it is a widely considered hypothesis that the continental regions may have evolved through the subduction (underthrusting) and the obduction (overthrusting) of ancient island arc systems with one another and with preexisting continental fragments. Another important hypothesis regarding the origin of metalliferous ore deposits is known as the geostill hypothesis. This postulates that rich metalliferous deposits were formed on or in the ocean crust near sites of plate accretion, that they were transported laterally to the arc-trench areas by sea-floor spreading processes and there were thrust back into the Earth's interior; eventually they were remelted and migrated back to the Earth's surface. When these metals (for example, copper) are found in concentrations such as in the Philippines, they are of great economic importance.

Finally, the release of energy by earthquakes and volcanic eruptions at the collisional plate boundaries underlines the need for more studies so that the destructive effects of these events can be minimized through accurate monitoring and prediction programs. *See* MARINE GEOLOGY; PLATE TECTONICS. Dennis E. Hayes

Bibliography. G. Dietrich, General Oceanography: An Introduction, 2d ed., 1980; A. B. Duxbury and A. C. Duxbury, Fundamentals of Oceanography, 2d ed., 1995; J. J. Gotthold and D. W. Gotthold, Indian Ocean, 1988; D. E. Hayes (ed.), The Tectonic and Geologic Evolution of Southeast Asian Seas and Islands, pt. 1, 1980, pt. 2, 1983; D. E. Hayes et al., A Geophysical Atlas of East and Southeast Seas, Geological Society of America, 1978; R. Lukas, T. Yamagata, and J. P. McCreary (eds.), Pacific Low-Latitude Western Boundary Currents and the Indonesian Throughflow, American Geophysical Union reprint from J. Geophys. Res., 101(C5):12209-12488, May 15, 1996; H. U. Sverdrup, M. W. Johnson, and R. H. Fleming, The Oceans: Their Physics, Chemistry, and General Biology, 1946; K. Wyrtki, Oceanographic Atlas of the International Indian Ocean Expedition, 1987.

Soybean

Glycine max, a legume native to China that is a major source of vegetable protein and oil for human and animal consumption and for industrial usage. Introduced into North America in 1765, soybeans were first grown as a curiosity and then as a forage and soil-building crop. The valued portion of the plant is the seed, which contains about 40% protein and 20% oil on a dry-weight basis.

Economic importance. The United States is the world's largest producer of soybeans. Illinois, Iowa, Minnesota, Indiana, Ohio, Missouri, Nebraska, Arkansas, South Dakota, Kansas, Michigan, and Mississippi are the major soybean producers there. Ranking after the United States are Brazil, China, and Argentina. About 64% of the soybeans produced in the United States are crushed or processed domestically, 31% are exported as whole beans, 3% are used for seed, and 2% have miscellaneous uses. Countries that import large quantities of whole beans for processing include Japan, the Netherlands, Taiwan, Spain, Mexico, South Korea, Germany, and Belgium. Russia, Canada, Algeria, and Venezuela import large quantities of soybean meal; and Pakistan, Morocco, Tunisia, Ecuador, and the Dominican Republic are the major importers of soybean oil.

Livestock consume about 97% of the soybean protein that is utilized in the United States. There is a limited use of the protein for human consumption as soy flours, grits, and texturized proteins in combination with other foods. Soybean oil is widely used throughout the world for human consumption as margarine, salad and cooking oils, and shortening. A small amount of protein and oil is used for industrial purposes.

In China, Taiwan, Japan, Korea, and Indonesia, soybeans are a regular part of the human diet as a source of protein and oil. Soybean milk, made by soaking beans in water, grinding the soaked beans, and filtering the insoluble pieces out, is directly consumed primarily by the Chinese; it is the starting material for preparation of soybean curd, probably the most important and popular soybean food in Asia. The soft cheeselike curd, known as tou fu, tofu, tubu, and other local names, has a bland taste and can be flavored with seasonings or blended with other foods. Soybean curd is made into a variety of products by frying, dehydration, and freezing, and is consumed daily in the same manner as highprotein foods in Western cultures. Soybeans also are prepared by fermentation with microorganisms to form highly flavored foods and seasonings such as soy sauce and fermented soybean paste. Soybean sprouts, immature green seed, and deep-fried mature seed provide additional sources of food for human consumption.

Plant characteristics. The soybean is an annual plant that reaches a mature height of 25–50 in. (64–127 cm). The cotyledons are the parts of the seed that emerge from the soil as the seedling develops. Two unifoliolate leaves (single leaflets) develop opposite each other above the cotyledonary nodes.

Other leaves produced on the plant generally are trifoliolate (three leaflets). Branches develop from the main stem in response to the amount of space between adjacent plants. A plant may produce few branches and only a few seeds in close spacings or many branches and over a thousand seeds in wide spacings.

Flowers of soybeans are self-fertilized and white or varying shades of purple. Pubescence (hair) on the stems, leaves, and pods is brown or gray. Pod color of commercial varieties in the United States is brown or tan, although genetic types are available with black pods. The pods generally contain one to four seeds ranging in size from about 800 to 10,000 seeds per pound (1800 to 22,000 seeds per kilogram); however, most commercial varieties have two or three seeds per pod with a size of 1600 to 3600 seeds per pound (3600 to 8000 seeds per kilogram). The seedcoat of commercial varieties is yellow, and the common colors of the hilum (seed scar showing point of attachment to pod) are yellow, gray, buff, brown, imperfect black, or black. Varieties of soybean native to Asia have yellow, green, black, brown, and varicolored seedcoats.

As a legume, soybeans can obtain nitrogen needed for growth from nitrogen-fixing bacteria that live in their roots. The bacteria can be applied at the time of planting by mixing them with the seed or by direct application to the soil. They can survive in most soils for several years and do not have to be reapplied when soybeans are grown in a field every 2 or 3 years. *See* LEGUME; NITROGEN FIXATION.

The growing season for soybean varieties is controlled by their response to photoperiod and temperature; consequently, the most productive varieties for the northern United States are different from those grown in the South. Number of days from planting to maturity can range from about 90 to 180, depending on the variety and environmental conditions.

More than 15,000 varieties and selections of soybeans have been introduced into the United States from other areas of the world, and in the past some of them were grown commercially. Current production, however, is entirely devoted to varieties produced by the breeding programs conducted by the U.S. Department of Agriculture, state agriculture experiment stations, and private companies.

Most soybeans in the United States are planted in a well-prepared seedbed, although in some locations they are planted in unplowed fields to reduce soil erosion. In areas where the growing season is long enough, soybeans are planted as a second crop immediately after wheat or other crops are harvested. Planting is done in rows ranging in width from 7 to 40 in. (18 to 102 cm) with equipment also used to plant corn, cotton, wheat, and similar crops. Weed control includes the use of chemical herbicides, mechanical cultivation, and hand labor. Harvest of soybeans with combines begins when the leaves have dropped and moisture in the seed is at or below the 13% level required for safe storage.

Diseases

Most diseases of soybean that occur throughout the world can be found in the United States. Diseases tend to be chronic in nature, occurring with some regularity from year to year and taking a small (and often unnoticed) toll. Annual losses from all diseases have been estimated at 12%, and this figure can fluctuate considerably, depending upon environmental conditions in any one particular season. Many soybean diseases are usually found in any one field, and a disease-free field is rare indeed.

Although the term disease is often difficult to define, it is usually considered that soybean diseases can be caused by both living and nonliving entities. In the first category are the parasitic diseases caused by fungi, nematodes, bacteria, and viruses. The nonliving causes of disease include conditions of weather and soil and the deficiency or excess of various minerals. Iron, potassium, and manganese deficiencies are common, and varying degrees of chlorosis accompany such deficits. Herbicide and lightning damage are also often reported by growers.

Fungus diseases. Fungi cause more diseases than other parasitic organisms. Leaves can be attacked by Peronospora manshurica and become severely spotted with yellowish-green areas on upper leaf surfaces. If the case is severe enough, the leaves can fall prematurely; the leaves more often remain intact, and the spots take on a dark-brown appearance with fungus growth apparent on the lower side of leaves. This symptom gives rise to the common term mildew. The fungus spreads rapidly from plant to plant and may overwinter in fallen leaves or invade pods and survive on seeds as a white crust. Many pathogenic variants of this fungus are found, and resistant varieties of soybean are being developed. Up to 8% reduction in yield has been reported from this disease alone in Iowa.

In the Midwest, brown spot, caused by *Septoria glycines*, is another disease of importance on leaves. It appears early in the season as angular lesions on primary leaves and may continue to move up the plant and cause defoliation, attack pods, and ultimately invade seed. Target spot, a disease caused by *Corynespora cassiicola*, is found farther south in the soybean-growing region and has caused losses in the range of 18–32% in Mississippi. Its name is derived from concentric dark rings formed as large brown and often irregular spots develop on leaves. Fungus leaf spots of minor importance include those caused by *Cercospora sojina*, *Phyllosticta sojicola*, and *Alternaria atrans*.

A series of fungus root and stem rots may plague soybeans. Root rot, caused by *Phytophthora megasperma*, is probably the most spectacular and serious disease in areas where it occurs. Plants die in all stages of growth, usually in low places in fields and along end turn rows. Several states and Canada have reported damage from this disease, and control is effected by the use of resistant varieties. Inheritance of resistance is controlled by a single factor, thus making it relatively simple to incorporate this resistance into promising (or preexisting) varieties of soybeans. The pathogen is not seed-borne.

By 1980, 19 pathogenic races had been described, and some researchers believe that single-gene resistance or tolerance no longer is a viable approach to control. Either a generalized resistance or tolerance has been considered. Most commercial cultivars possess considerable field resistance to *Pbytophtbora*.

Brown stem rot, caused by Cephalosporium gregatum, is a disease of major concern in the Midwest, and the fungus is widespread in Minnesota, Illinois, Iowa, and Indiana. Crop rotation (on a 3-4-year basis) is a well-known control. Resistance as a viable method of control was investigated in the 1970s, and by the end of that decade at least two major producing states were considering this method of control. Seed rot and root rot by Pythium and Rhizoctonia are common and chronic on seedlings over vast areas of the soybean region. Plants able to emerge usually outgrow this problem; however, stand reduction may be severe enough to warrant replanting. Stem canker and pod and stem blight are two stages of an important disease caused by Diaporthe phaseolorum which are favored by high humidity and can be controlled to some extent by chemical treatment. A widespread and serious seed quality problem in the warmer and more humid regions of the South and Midwest results from pod infection by Diaporthe. Systemic fungicides have provided some improvement in seed quality and yields.

Soybean rust, a fungus-induced disease, is assuming more importance as acreages expand in traditional soybean growing areas and as production moves into new areas. The disease is serious in Asia, especially in tropical and subtropical areas. In Taiwan, Thailand, the Philippines, and Java it is the most damaging fungal disease affecting soybean crops. The pathogen, Phakopsora pachyrbizi, can infect and sporulate on species in 17 known legume genera but is economically important only on soybean. The pathogen has been reported in Africa on cowpea but not on soybean. In 1976 soybean rust was found on soybean cultivars Biloxi, Hardee, Santa Rosa, and Williams in an experimental planting in the highlands of western Puerto Rico. This was the first record of rust on soybean in the Western Hemisphere, and is cause for concern in the Americas. It is likely that the pathogen will be found on soybean plantings elsewhere in the Caribbean area.

Bacterial diseases. Bacterial diseases usually are a chronic problem of this crop. Bacterial blight (**Fig. 1**) caused by *Pseudomonas glycinea* is found primarily in the upper Midwest and bacterial pustule caused by *Xanthomonas phaseoli* predominates in the southern United States. These pathogens enter plants through pores or injuries in leaves and are spread rapidly by windblown rain. Seed-borne infection is common and constitutes the major means of overwintering. Infected seedlings can thus emerge, and the disease cycle begins again. Antibiotic sprays are impractical, and resistant strains of the blight pathogen develop rapidly during their use. Some varieties of soybean are more resistant to blight than



Fig. 1. Bacterial blight symptoms on soybean leaves under field conditions. The condition results principally in chlorate spots and malformations.

others (**Fig. 2**), and inheritance of resistance in the soybean host depends upon the particular strain of the pathogen being tested.

A chlorosis (yellowing of leaves) may develop from strains of *Rbizobia*, which are abnormal in that they cannot fix nitrogen but are aggressive in entering soybean roots. These bacteria thus occupy space that efficient nitrogen fixers would otherwise occupy. A bacterial wilt caused by *Corynebacterium flaccumfaciens* has been reported in certain areas of the Midwest but is not of great economic importance.

Virus diseases. The viruses may enter soybean plants by any of several methods, although each virus usually is very specific in its mode of entrance. They



Fig. 2. Response of soybean varieties to bacterial blight. (a) Complete resistance. (b) Complete susceptibility.

may be injected into the plant by insects or nematodes or through wounds or may go from one generation to the next via the seed. Only a few of the viruses of this crop have been studied extensively.

Soybean mosaic is common and is easily transmitted mechanically or by insects or seeds. It usually appears on leaves as crinkling and distortion of young tissues, and damage is similar to that produced by sprays used in weed control. It affects the seed by producing a mottled appearance; the seed becomes discolored when pigments from the seed scar appear to streak across seed coat surfaces during their development. Control is difficult, but yield losses have not been excessive.

Bud blight, on the other hand, is a very destructive virus disease, and no resistance is known. It, too, is seed-borne; it is apparently transmitted by insects. Usually, isolated plants in a field are found scattered at random (originating from infected seed), or plants along the edge of fields bordering alfalfa or fence rows may be heavily infected. This phase is thought to be due to an insect vector moving the virus from other plants to soybeans. Loss in yield can be 100% of the crop. This virus causes the plant to produce a rank green stem which stays green after frost, and seed pods are either absent or small; if infected while young, the growing points of plants die and curve downward, giving rise to the common name "bud blight." The same virus causes ring spot in tobacco. See PLANT VIRUSES AND VIROIDS.

Nematodes. These pathogens are microscopic, eellike worms in the soil. Members of the genus Meloidogyne cause root knots or galls which form as large swellings and cause severe yellowing and stunting of aboveground portions of the plant. Crop rotation and the use of resistant varieties constitute the usual control methods. The soybean cyst nematode Heterodera glycines represents a real threat to soybean production and is now so widespread that quarantine no longer is effective and has been abandoned. It was first found in 1952 in North Carolina and since then has been found in isolated places in Virginia, Mississippi, Arkansas, Illinois, Kentucky, Missouri, and Tennessee. It is now feared from southern to northern climates. It moved northward in Illinois for many years, and was found for the first time as far north as Minnesota in 1978 and in Iowa in 1979; it appears to be destructive in northern as well as southern climates. Symptoms vary but are not very dissimilar to root knot symptoms on aboveground parts of the plants. Below ground, small white cysts are noted; once present, these cysts can exist many years in the soil, and complete control becomes almost impossible. Yield loss may be severe. Resistant varieties and rotation with nonsusceptible crops are recommended. See NEMATA (NEMATODA); PLANT PATHOLOGY. Bill W. Kennedy

Processing

Processing of soybeans into oil and defatted cake or meal began in the Orient. The oil was used primarily in foods, whereas the meal was used as a fertilizer and in animal feed. In the United States, soybeans have



Fig. 3. Steps involved in the solvent extraction method for soybeans.

been processed primarily to obtain oil. The oil was processed for use in shortenings, margarines, and salad dressings. The defatted cake or soybean meal, once a by-product of soybean processing, has become the principal product from soybeans because of its protein value for livestock feeding. Soybean protein products are prepared primarily for direct human consumption. Direct food usage has been limited, but has been expected to increase dramatically because of population pressure on the food supply and limited land and energy resources available for expansion of the livestock industry.

The soybean processing industry converts the oil and meal from the soybean seed into numerous food and industrial products. The products can be generally classified into three categories: soybean oil, soybean meal, and soybean protein products. The processing steps are oil extraction, oil refining, and protein processing.

Oil extraction. The basic processing of soybeans consists of separation of the oil and meal. The equipment used varies considerably from plant to plant although the general process, solvent extraction of the oil, remains the same. Older, obsolete methods used either continuous screw presses or hydraulic presses to squeeze out the oil. Solvent extraction has almost universally been adopted because of its efficiency and low operating cost.

The steps involved in the solvent extraction of soybeans are outlined in **Fig. 3**. Storage of soybeans in elevators or silos is included as part of the processing since it ensures a continuous, year-round supply of beans for processing without the day-to-day price fluctuations of the commodity market. The soybeans, prior to storage, are cleaned and dried to 12–14% moisture for improved stability until processed.

Preparation. The processing operation consists of preparation, extraction, and meal-finishing phases. Preparation for extraction consists of steam-conditioning the beans, cracking, dehulling, and flaking. Cracking the beans into small particles, usually six to eight pieces, loosens the hulls from the soybean meats and assists in the flaking operation. Corrugated rollers are used for the cracking operation. Dehulling or separation of the hull material from the cracked meats is necessary to increase extraction

efficiencies. Dehulling is performed by a combination of air aspiration, screening, centrifugal separation, and density separation methods. Flaking the cracked, dehulled soybean meats facilitates the extraction process by disruption of the internal cellular structure of the bean and by allowing better contact of the extracting solvent with the internal parts of the seed. Flaking is done by heating and steaming the cracked meats and passing them through differential roller mills. The resulting flakes are paper-thin (0.009-0.02 in. or 0.25-0.40 mm).

Extraction. This operation consists of countercurrent washing of the flakes with a petroleum solvent, usually hexane. The solvent is kept near 120° F (50° C) during extraction to speed the solubilization of oil from the flakes. The oil-laden solvent, called micella, is then separated from the flakes and distilled for recovery of the crude soybean oil. The solvent is recovered for further use. The crude soybean oil is then pumped to storage tanks for further processing or to be sold as crude soybean oil. *See* SOLVENT EXTRACTION.

Finishing. The solvent-wet, defatted flakes are desolventized by steam and heat. The severity of the desolventizing depends on the uses for the meal. If the extracted flakes are to be used in animal feeds, moisture is adjusted and the flakes are heat-treated (toasted) to increase their feed value as a protein supplement. Flakes to be used for edible soy flour or protein products are subjected to less severe heating to prevent heat damage to the proteins present. After desolventizing, the flakes are screened and ground. The protein content may also be adjusted. Anticaking agents may be added to ensure ease in handling.

Yields. The processing yields based on a 60-lb (27.2-kg) bushel are 10.2-11.6 lb (4.6-5.3 kg) of crude oil and 48.4-49.8 lb (22-22.6 kg) of meal and feed products. The maximum yield of crude oil is always attempted. The meal from the extraction contains 49-50% protein. The hulls, accounting for about 5% of the beans, are ground and may be added back to the extracted meal for production of a 44% protein meal. The quantities of each meal prepared depend upon the market value at the time of production. The ground hulls may be stored for later use, sold as mill feed to be used as a source of fiber in mixed livestock feed products, or used in a patented process as a carrier for molasses in feed concentrates.

Soybean oil processing. Soybean oil is used in salad dressings, shortenings, and margarines and in various nonfood applications such as paints and varnishes. Most (93%) is used in foods. To be suitable for edible products, crude oil from the extraction requires further processing.

The processing of soybean oil, generally referred to as refining, is outlined in **Fig. 4**. Individual processes are degumming, alkali refining, bleaching (decolorizing), hydrogenation, deodorization, and



Fig. 4. Steps involved in the soybean oil refining process.

winterization. Semirefined oils are also marketed, depending upon the customer's demands. Hydrogenation is varied, depending upon the final use of the oil. All edible oils are deodorized.

Degumming. Phospholipids, or lecithin, present in the crude soybean oil will precipitate during storage and contribute to deterioration during use of the oil. During degumming, the soybean phospholipids are removed from the crude oil. Degumming itself refers to the gummy consistency of the wet phospholipids. The process involves the addition of a small amount of water or steam to the crude oil, resulting in the hydration of the phospholipids and their subsequent precipitation. Acidic catalysts are sometimes added with the water to improve the hydration. The precipitated phospholipids are then separated by centrifugation and are vacuum-dried. The dried phospholipids are sold as soybean lecithin.

Lecithin. Commercial soybean lecithin is widely used in foods, particularly chocolate, margarine, shortening, and greaseless frying compounds. Soybean lecithin from the degumming step contains 60-65% phospholipids and 30-35% entrained soybean oil. The principal phospholipids present are phosphatidylcholine, phosphatidylethanolamine, and phosphatidylinositol. The crude lecithin is dark-colored and may be bleached with hydrogen peroxide to produce a light, straw-colored product. Fatty acids may be added to make a fluid product, or the lecithin may be treated with ammonia to form a firm, plastic consistency. More refined grades (pharmaceutical lecithin) are produced by acetone extraction of the entrained soybean oil. The phospholipids are insoluble and precipitate. Further processing consists of desolventizing and granulating.

Alkali refining. Free fatty acids decrease the heat stability of the oil. The degummed soybean oil is alkali-refined to remove the free fatty acids, residual phospholipids, and color. Alkali refining consists of addition of caustic soda, soda ash, or a combination of the two to neutralize the free fatty acids. Careful addition of the alkali is required to prevent further degradation of the oil. The neutralized free fatty acids or soaps are removed from the oil by centrifugation or settling. The soaps, referred to as soapstock, are sold to soap manufacturers, fatty acid industries, or livestock feed formulators in the alkaline form, or in an acidulated form after treatment with sulfuric acid. The alkali-refined oil is further washed with water to remove all traces of alkali and soaps present and then is dried. The neutralized oil is often sold to industrial users for nonfood applications.

Bleaching. Bleaching (decolorizing) removes the color bodies or pigments present in the oil by adsorption on activated earth or carbon. The bleaching agent is simply dispersed in the oil and removed by filtration. The bleached oil is sent to the hydrogenation unit or to the deodorizers, depending upon the final use. Foods requiring improved oxidative stability or plastic fats such as margarines or bakery shortenings require hydrogenation.

Hydrogenation. Development of the hydrogenation process for vegetable oils has permitted the substi-

tution of vegetable oils for animal-derived fats and the substitution of the various types of vegetable oils with soybean oil. Variations in the hydrogenation conditions can be used to produce high-stability soybean oil, hardened soybean oils that melt rapidly at body temperature, and shortening-like fats.

Hydrogenation is performed by purging the oil with hydrogen gas at an elevated temperature and pressure in the presence of a catalyst. Nickel catalysts are generally used. The addition of hydrogen to the fat saturates the molecular structure. The degree of saturation depends upon the time, temperature, and pressure of hydrogenation. *See* HYDROGENATION.

Deodorizing. Deodorization of the oil removes the undesirable flavor and odor constituents. Oils must be odorless and bland in flavor to be suitable for food use. Deodorization is carried out under high temperature and vacuum with steam injected to assist volatization of the undesirable components.

Winterization. Winterization is performed on lightly hydrogenated soybean oil that is to be used as a salad oil. The oil, in addition to conforming to oxidative stability requirements, must remain clear at refrigerator temperatures. Winterization consists of cooling of the oil and filtering off the cloudy haze that forms.

Reversion. Refined soybean oil is a highly unsaturated oil and will develop, after storage, characteristic flavors and odors described as painty or fishy. The development of these flavors is called reversion, and is possibly a result of the oxidation of the linolenic acid component of the oil. Care in the exposure of the oil to air during processing and controlled hydrogenation to partially saturate the linolenic acid minimizes the tendency of soybean oil to reversion. Metal sequestrants, such as citric acid, are sometimes added during deodorization to inactivate the prooxidant metals such as iron and copper. Antioxidants may also be added. *See* ANTIOXIDANT; FAT AND OIL (FOOD).

Soybean protein products. The soybean protein products are prepared from the extracted meal. Three distinct product types are prepared: soy flour and grits, soy protein concentrates, and soy protein isolates. These are used both in foods and in nonfood applications.

Flour and grits. Soy flour and soy grits contain approximately 50% protein based on the nonlipid fraction. Flours and grits are used in processed meats, bakery mixes, cereals, breads, and baby foods. A large amount has been utilized in nutrients for export to undeveloped, underfed countries.

The term soy flour refers to an extremely fine powdered product, whereas the term grits refers to a granular product. Various products are prepared varying in oil content, oil type, particle size, and water solubility of the protein. In general, the oil content varies from less than 1% (defatted) to 18–20% (full fat); the oil source is soybean oil or lecithin. Particle size varies from powders that pass a 200-mesh screen (flour) to those that pass a 14-mesh (grits) one. Water solubility of the protein ranges from 10% (toasted) to 80% (untoasted).

Soy flour or grits is produced similarly to soybean



soy protein isolate

Fig. 5. Production of soy protein concentrates and isolates.

meal, except with thorough removal of the hulls before extraction. Desolventizing and toasting are varied to alter the water solubility of the protein, with minimum heating being required for high protein solubility. The defatted, desolventized flakes are ground and screened to the desired particle size.

The oil-containing products are generally prepared by blending soybean oil or lecithin back to the extracted meals. Full-fat soy flour may also be prepared by steaming the unextracted, dehulled flakes to remove the bitter flavors, and then drying, grinding, and screening them to the desired particle size.

Concentrates. Soybean protein concentrates and soybean protein isolates are generally prepared from soy flour or grits. The processing methods vary considerably between processors. A generalized process is shown in **Fig. 5**.

Soybean protein concentrates contain 70% protein. Concentrates are prepared by washing the water-soluble carbohydrates from the meal under conditions designed to ensure minimum protein solubility. Dilute acids or aqueous alcohol solutions have been used. The washing is performed in agitated tanks. The remaining insoluble proteins and carbohydrates are removed from the extraction by centrifugation or filtration, washed, dried, and ground. Products are marketed with ranges of particle sizes and protein solubilities similar to soy flour and grits. Soy protein concentrates are usually bland in flavor, possess a lighter color, and are more functional than soy flour (having emulsification, waterbinding, gelling, and other properties). *Isolates.* Soy protein isolates are the purest form of soy proteins marketed (90% protein minimum). The isolates are prepared by solubilizing the protein present in the grits by using mildly alkaline solutions, separating the protein solution from the fibrous residue by centrifugation or filtration, and acidifying the protein solution to precipitate the protein as a curd. The curd is recovered, washed, resolubilized, and spraydried into a fine, flourlike powder.

Soy protein isolates are manufactured to possess specific functional properties such as gelling, whipping, foaming, emulsifying, thickening, and binding. Isolates are used in meats, simulated dairy products, and bakery products.

Textured proteins. Each of the soybean protein products can be used to form textured soy proteins. Textured soy proteins have a meatlike texture and appearance when hydrated and can be used for extension of ground and comminuted meats (such as ground meat and bologna) and for the preparation of simulated meats. When they are used as extenders, the textured proteins increase cook yield, aid in binding of meat constituents, and improve emulsion stability.

Two general processes (with variations) are used for texturizing soy proteins: extrusion cooking and spinning. Extrusion cooking consists of moistening the protein, working the doughlike mass under high temperature and pressure, and forcing the cooked mass through an orifice, which permits the heated compressed mass to expand. The product is then dried and sized. Colored and flavored extruded soy protein products are prepared by addition of these additives either before or after extrusion. Imitation bacon bits that are based on extruded soy flour are widely marketed.

Soy protein isolates are required for spinning. Spun proteins are somewhat more adaptable to simulated meat production than extruded proteins. Spun proteins are produced by solubilizing the isolates in an alkaline solution and forcing this solution through a multiholed die immersed in an acid bath. The solubilized protein immediately coagulates in the acid bath, forming a continuous filament. The filaments are stretched to orient the protein molecules within the filament. Toughness of the filaments is dependent upon the degree of protein orientation. Filament diameters are approximately 0.003 in. (0.076 mm). Groups of filaments when combined with egg white, flavors, and colors, followed by compression and heat setting, form imitation meats such as ham, chicken, or beef. See FOOD MANUFACTURING. F. T. Orthoefer

Bibliography. D. Erickson et al., *Handbook of Soy Oil Processing and Utilization*, 1980; W. O. Scott and S. R. Aldrich, *Modern Soybean Production*, 2d ed., 1983; J. B. Sinclair, *Compendium of Soybean Diseases*, 4th ed., 1999; S. R. Singh et al. (eds.), Soybeans for the Tropics: Research, Production, and Utilization, 1988; H. E. Snyder and T. W. Kwon, Soybean Utilization, 1987; D. Swern (ed.), Bailey's Industrial Oil and Fat Products, 2 vols., 4th ed., 1986.

Space

Physically, space is that property of the universe associated with extension in three mutually perpendicular directions. Space, from a newtonian point of view, may contain matter, but space exists apart from matter. Through usage, the term space has come to mean generally outer space or the region beyond the Earth. Geophysically, space is that portion of the universe beyond the immediate influence of Earth and its atmosphere. From the point of view of flight, space is that region in which a vehicle cannot obtain oxygen for its engines or rely upon an atmospheric gas for support (either by buoyancy or by aerodynamic effects). Astronomically, space is a part of the space-time continuum by which all events are uniquely located. Relativistically, the space and time variables of uniformly moving (inertial) reference systems are connected by the Lorentz transformations. Gravitationally, one characteristic of space is that all bodies undergo the same acceleration in a gravitational field and therefore that inertial forces are equivalent to gravitational forces. Perceptually, space is sensed indirectly by the objects and events within it. Thus, a survey of space is more a survey of its contents. See EUCLIDEAN GEOMETRY; LORENTZ TRANSFORMATIONS; RELATIVITY; SPACE-TIME.

The space around the Earth and extending out perhaps 10 earth radii (40,000 mi or 64,000 km) has properties which differ from those of interplanetary space. The chief reasons for this difference are the existence of the gravitational and magnetic fields of Earth. These fields affect the densities and motions of neutral and charged particles in the vicinity of Earth. The corona of the Sun, which was once believed to be a static and limited atmosphere, fills the entire solar system, although in a highly tenuous form. This corona, in the form of a plasma of hydrogen particles, is pervaded by a magnetic field, and it flows past the Earth at speeds as high as 250 mi/s (400 km/s). *See* VAN ALLEN RADIATION.

The geomagnetic field of the Earth, unlike an undistorted dipole field, is confined to a limited region of space, as data supplied by satellite measurements suggest. The actual field is distorted by radiation belts and primarily by its interaction with the solar wind. This volume of space is called the magnetosphere, and is bounded by a thin layer called the magnetopause that fluctuates in response to the solar wind. The magnetosphere appears to have the shape of an elongated comet whose tail points away from the Sun.

Between 10 to 20 earth radii, a shock wave is produced by the flow of the solar wind around the magnetosphere.

The gravitational field allows Earth to retain its atmosphere but is not strong enough to prevent escape of the atmosphere completely. Above a level of about 310 mi (500 km), where the atmosphere is so rarefied that collisions between atoms can be neglected, the atoms describe ballistic orbits and form the exosphere. Light atoms, especially hydrogen, but also helium, occasionally obtain enough energy so that they escape completely from Earth's gravitational field. *See* ATMOSPHERE; MESOSPHERE.

Earth's magnetic field deflects moving charged particles. Particles of the highest energy, namely cosmic rays, are deflected by the magnetic field so that only those having an energy of more than 15 GeV can enter near the Equator. At the poles, however, cosmic rays of all energies may enter because there the lines of force of the magnetic field are vertical. As a result, cosmic ray intensity at the pole is 10–50 times higher than at the Equator. *See* GEOMAGNETISM.

Solar corpuscular radiation, containing particles of much lower energies, is affected more strongly by Earth's magnetic field and is deviated into zones at high latitudes. Many of these solar particles are trapped in Earth's magnetic field and remain in it for long periods of time, moving back and forth along a line of force, but being reflected from each end. In this process particles may be accelerated to higher energies and then be energetic enough to produce luminous displays in the upper atmosphere. In their trapped condition these particles contribute to Earth's radiation belt, first observed in early Earth satellite experiments. Another source for the radiation belt comes from cosmic rays which plunge into Earth's atmosphere, disintegrate atmospheric nuclei, and throw back into space debris which can be trapped. See ASTRONOMY; AURORA; COSMIC RAYS.

S. F. Singer

Space biology

Collectively, the various biological sciences concerned with the study of organisms and their components when exposed to the environment of space (or of a spacecraft beyond the immediate influence of Earth and its atmosphere). Space biology studies include all the biological science disciplines, such as molecular, cellular, developmental, environmental, behavioral, and radiation biology; physiology; medicine; biophysics; and biochemistry. Space biology encompasses both basic and applied sciences, and focuses on using space research to understand and resolve biological problems on Earth, as well as alleviating or preventing the deleterious physiological changes associated with space flight or habitation in non-Earth environments. *See* SPACE.

As crewed space flight developed, medical research focused on the health and well-being of astronauts was paramount. This medical research has historically been included in the area of aerospace medicine. From these research efforts, directed at survival of living systems in the space environment, has evolved the broader discipline of space biology, which uses the space environment to gain an understanding of basic biological processes. *See* AEROSPACE MEDICINE.

Space biology considers all of the inherent physical factors of the space environment and their effects on living organisms: (1) factors due to the environment of space, such as radiation and reduced pressure; (2) factors due to the flight dynamics of spacecraft, including acceleration, weightlessness, and vibration; (3) factors due to the environment of the spacecraft, such as air composition, atmospheric pressure, toxic materials, noise, and confinement; and (4) factors due to the environment on nonterrestrial bodies, such as the level of gravity, composition of soil, and temperature extremes. Of these factors, only weightlessness and space radiation are unique to the space environment, but the effects of the other factors must be considered and mimicked in ground-based or flight controls when conducting space biological experiments. In addition, space experiments must take into consideration the physical effects of weightlessness on the environment surrounding and supporting the organisms. For example, neither gases nor liquids behave in a space environment as they do under the conditions found on Earth, as sedimentation, buoyancy, and convection do not occur in a weightless environment. See AC-CELERATION; WEIGHTLESSNESS.

Space Gravitational Biology

Among the effects of space flight, weightlessness (free fall) associated with orbital flight presents the greatest scientific opportunity and poses the greatest challenge. All life has evolved under the most constant and ubiquitous of the Earth's environmental factors, gravity. The near-weightless (microgravity) environment of space flight provides the means to manipulate gravity below 1 *g* in order to understand its effects on biological systems. Studies in microgravity provide information as to the influence of gravity on living systems, and insights into possible life in environments beyond Earth.

Periods of near-weightlessness, ranging from a few seconds to about 15 minutes, can be achieved by using drop towers, parabolic airplane flights, and sounding rockets. To extend the exposure time pe-



Fig. 1. *Cosmos* 782, one of a series of Soviet and Russian spacecraft that carried biological experiments. Launched on November 25, 1975, on a 20-day mission, it carried more than 20 different species. Scientists from France, Czechoslovakia, Hungary, Poland, Romania, the United States, and the Soviet Union participated in these investigations.

riod to days, weeks, and months requires orbital space flight. Biological specimens have been sent into space in a variety of satellites and crewed space ships (**Figs. 1** and **2**). The collection of biological data in space has been complicated by a number of limiting and confounding factors: few flight opportunities, short-duration space flights, adverse spacecraft environmental conditions, complicating spacecraft flight dynamics, accommodations for only a few species, limited means to manipulate specimens in space, and the general lack of a 1-g reference control centrifuge on board the spacecraft. Yet many significant observations have been made that establish a foundation for future space research.

Ground-based research. A variety of ground-based experimental methods have been used successfully to complement space experiments. Each method depends upon changing the force imposed on part of an organism or the entire organism-for example, by using centrifugation to increase the level of gravity or changing the gravitational vector by clinostating (randomly rotating). These methods are restricted by the untoward additional biological effects they create. Organs and tissues such as bone and muscle that have evolved to support animals against the Earth's gravitational force have been the most responsive to ground-based modeling. Rats exposed to space flight and those subjected to 1-g and 2-g centrifugation showed that total body calcium was decreased by reduced gravitational loads and increased by elevated gravitational loads. Computerized analysis of the morphology of the rat inner-ear vestibular system, crucial for orientation and balance, has shown that the number of nerve cells innervating these organs is also gravity-dependent. There are more nerve cells in space-flight-exposed tissue, for which the gravitational stimulus has been reduced, while the number of cells is decreased when the gravitational load is increased to 2 g. Further, bone cells unloaded by rotation demonstrate differential responsiveness to growth mediators compared to static cultures. Imposition of a load to the cells results in cell-type differentiation. During centrifugation, the increase in the level of gravity produces distinct differences in the behavior of rodents. Activity is reduced, and there are shifts in circadian rhythms opposite to those observed during space flight. Ground-based research using these types of experimental models provides direction for and insights into space flight experiments.

Effects on unicellular organisms. Space experiments have been conducted with a variety of unicellular organisms, including bacteria, fungi, protozoa, and algae. Developmental, physiological, and behavioral effects have been noted. In prokaryotic cells (cells without nuclei, such as bacteria), changes in metabolism have been repeatedly documented. Bacteria grow two to three times more rapidly in microgravity, are more resistant to antibiotics, and have more complete exchange of nuclear material during conjugation than on the ground. Decreased sporulation is observed, suggesting a change in bacterial differentiation in microgravity. While no differences are

found in the genetic recombination mechanisms of transformation or transduction of *Escherichia coli*, conjugation is increased three- to fourfold in space flight. *See* BACTERIAL GENETICS; BACTERIAL PHYSIOL-OGY AND METABOLISM; TRANSDUCTION (BACTERIA); TRANSFORMATION (BACTERIA).

Under microgravity conditions the eukaryotic single-celled organism Paramecium tetraurelia displays an increased growth rate accompanied by increased cell volume and decreased dry weight, protein content, and calcium. There are possible changes in cell membrane assembly and extrusion of trichocysts as well. The alga Chlorella vulgaris shows changes in the ultrastructure of the cells, and rearrangement of organelles. The degree and spectrum of change are correlated with the duration of the space flight. Mitochondrial density and volume and the number of cristae in plastids are the first to undergo change. Chlamydomonas cultures show increased cell size and a prolongation of the active growth phase. The slime mold Physarum polycephalum demonstrates that both light and gravity regulate the contraction frequency of plasmodial strands. The results implicate the mitochondria, since they play a major role in changing the contraction rhythm, and suggest that the mitochondria not only supply most of the energy necessary for the contraction but also may be involved in its regulation and possibly even in the perception of the gravity stimulus. See CELL ORGANIZATION; CELL PLASTIDS; MITOCHONDRIA.

Effects on animals. Physiological and developmental effects of microgravity have been observed in many types of animals.

Physiological effects. Bone in growing rats is subtly altered in space flights lasting up to 21 days. The mineral concentration, size, and bone resorption are minimally changed, while bone formation, mineralization, and maturation are suppressed. While the bone increases in size in growing rats during flight, it has reduced strength or resistance to breaking. Bone vascularity, nerves, ligaments, tendons, and vertebral discs of young rats all respond in some way to space flight.

Skeletal muscles that normally function to overcome gravity atrophy during space flight. Although muscle mass changes occur quickly, the underlying physiological changes may continue even after the mass achieves maximum loss. There is a decrease in the cross-sectional area of muscle fibers (primarily of the slow-twitch oxidative fibers), associated with a substantial loss of muscle protein, especially myofibril proteins; a decrease in alpha-actin mRNA; increased glucose uptake in the presence of insulin; and increased activities of glycolytic enzymes; while oxidation of fatty acids appears to be reduced. In cardiac muscle, mitochondrial numbers decrease, and increased lipid and glycogen content indicate a reduced work capacity. Cardiac muscle cells appear deficient in their ability to utilize energy, based on adenosine triphosphate (ATP) content and fatty acid metabolism. Regulatory proteins in the ventricle show reduced responsiveness to hormone stim-



Fig. 2. Payload specialist Mamoru Mohri using a microscope to produce photomicrographs of mammalian cells during the Spacelab SL-J mission, a joint effort of Japan and the United States aboard flight *STS 47* of the space shuttle *Endeavour* in September 1992. The biopack contains a health-monitoring experiment. (*NASA*)

ulation, and heart cells show mild degradation. *See* CARDIOVASCULAR SYSTEM; SKELETAL SYSTEM.

In rats flown in space, there is a reduction in the volume of red blood cells. Hemolysis of red blood cells is about three times greater in rats exposed to space flight than in control rats, and the mean life-span of these cells is decreased by about 4 days during flight. This reduction is in part due to a decrease in plasma levels of erythopoietin, reducing production of red blood cells. There is a reduced immune response to mitogens. Livers from rats exposed to weightlessness contain up to 100% more glycogen than those from control rats, and bone and other organs also show increases in lipid content. Analysis of the body constituents of space-flight-exposed young rats shows an increase in potassium, creatine, nitrogen, and magnesium and a decrease in calcium, phosphorus, and sodium when compared to 1-g controls on Earth. Conversely, rats subjected to an increase in gravity to 2 g by centrifugation exhibit a decrease in potassium, creatine, and nitrogen and an increase in calcium, phosphorus, magnesium, and sodium. Pituitary oxytocin and vasopressin are decreased, suggesting in-flight adjustments in the regulation of body water content. Rats, like humans, appear dehydrated after flight, with changes in water-salt metabolism. The extent of this water loss appears to be related to the time after landing at which the measurements are made. The longer the time after landing, the greater the loss of body water. This is corrected within 3-5 days after return to Earth. Upon return to Earth, animals have a pronounced increase in urine output, further supporting a reduction in regulation of water and electrolyte homeostasis.

Decreased levels of growth hormone are related to increased concentration of growth hormone in pituitary cells; these cells release less growth hormone when cultured postflight, suggesting a secretory dysfunction. These changes in growth hormone also appear to be associated with a reduction in translation, as adequate amounts of mRNA are present. Most changes induced by space flight are corrected within 1 month of return to Earth, and an artificial gravity level of 1 g during flight appears to normalize many flight-induced effects. *See* ENDOCRINE MECHA-NISMS.

In rhesus monkeys, cardiovascular measurements taken during space flights indicate a decrease in heart rate, increase in blood pressure, and reciprocal changes in blood flow velocity associated with adaptation to space flight. Opposite responses are noted with exposure to increased gravity produced by centrifugation. After space flight, studies of muscle function and morphology show a shift in muscle typing from slow twitch to fast twitch. This is predominantly noted in those muscles used to support and move the body against gravity. In addition, the function of the muscle during movement is altered. Circadian rhythms of activity and body temperature persist during space flight, but both regulation and circadian timing of these systems are significantly influenced by microgravity. At the same time, heart rate maintains a normal 24-hour rhythm, indicating it is unaffected by microgravity. Studies of vestibular function suggest that acclimation to microgravity causes changes in the way that body orientation information is processed in the nervous system. These changes in vestibular function, body orientation, and muscle morphology and function contribute to altered posture and movement upon return to Earth. There are significant alterations in walking and running gait. Fine motor control also appears to be affected. These changes are similar to those reported in astronauts.

Reproduction and development. The experimental difficulty in initiating fertilization in space and the short duration of most space flights have been limiting factors in research on animal reproduction and development in space. The scientific questions fall in two main categories: (1) the effects on fertilization and early development, and (2) the effects on the different organ systems, since each has a critical development period that is especially sensitive to environmental perturbation. The majority of the research has focused on fertilization and early development, and in general has found that microgravity has minimal effect on the animals tested.

Normal fertilization and development of fruit flies can occur in microgravity. However, larval hatching and the number of adults emerging after recovery are significantly reduced, oogenesis is affected, and there is a slight increase in abnormal head morphology. Stick insects also show a decrease in hatching with apparent normal development. The developmental process in polyps and ephyrae of jellyfish (*Aurelia aurita*) appears to be normal. However, their pulsing movement is abnormal and may reflect abnormal development of the graviceptors, the nerve net, or the neuromuscular system. Successful fertilization of the sea urchins *Paracentrotus lividus* and *Lytechinus pictus* has been achieved; however, significant abnormalities were noted in the centrosomecentriole complex of dividing cells. These abnormalities suggest alterations in developmental processes in microgravity. The nematode *Caenorhabditis elegans* exhibits normal development and chromosome mechanics.

The fish Fundlulsu reproduces and develops normally, and gravid guppies give birth to normal young when returned to Earth. Female frogs of the species Xenopus laevis have been stimulated to ovulate in space, and their eggs have been fertilized in microgravity. Exposure to microgravity from oocyte maturation through hatching does not deleteriously affect amphibian morphogenesis or permanently alter behavior. The duration of the flights with developing frogs has been short, and some developmental differences have been noted that may affect morphogenesis. For example, the lungs of tadpoles reared in space are not inflated, either because they are not needed for buoyancy or because the young cannot break through the air-water interface to inflate them. Upon return to Earth, tadpoles fail to swim appropriately, possibly due to the difference in buoyancy.

Apparently normal chicks have hatched from fertilized Japanese quail eggs (Coturnix coturnix japonica) incubated in space. They react to visual and auditory stimuli and display normal motor activity but have orientation problems. Early chicken embryos have died during space flight following early preincubation on the ground. The deaths have been suggested to result from the absence of buoyancy during space flight. The yolk normally floats at the top of the egg, allowing the embryo close contact with the shell for easy diffusion of oxygen. During space flight the yolk does not float to the top of the egg, and thus the embryo may be starved for oxygen and die. Later in development, surviving chicks from embryos exposed to space flight show decreased vestibular sensitivity to gravitational stimuli.

In rats, early development during gestation does not appear to be affected by microgravity. Pregnant rats launched into space early in gestation deliver physiologically and behaviorally normal rat pups upon return to Earth. However, maternal differences occur that could affect the young. The metabolism of the mammary glands of rats flown in space is reduced and could result in a decrease in milk production. Young rats flown on the space shuttle prior to weaning do show deleterious effect. If they are too young, mortality is very high, as the microgravity environment affects behavioral interactions between the mother and neonate. Animals that survive appear to have neurological deficiencies that persist into later life when returned to Earth.

Effects on cultured animal cells. Most mammalian cells grown in microgravity (Fig. 2) show a reduced growth rate compared to ground controls. Reduced proliferation is observed with cultured human lung cells, monkey kidney cells, mouse hybridoma cells, and human lymphocytes. The rate of glucose utilization may be reduced, and there may be changes in membrane structure or receptors. Culture experiments with cells that are dependent upon attachment to a substrate are compromised in microgravity, where all objects tend to float. Culturing of lymphocytes in microgravity is not similarly impacted because the cells are capable of autonomous motility. Immune-cell cultures in space produce interferon more rapidly and with four to eight times higher titers. Human mononuclear cells grown in space lose 50-90% of their ability to respond to stimulation with the mitogen concanavalin A, compared to control cells grown on Earth or in 1-g centrifuge in space. Immune cells also display major depression of blast cell transformation in response to mitogens, loss of cytokine production or function, and changes in natural killer cell activity and response to colonystimulating factor. Changes in cellular function are associated with alteration of gene levels. Human kidney cells grown in space show alterations in signaling pathways, and structures within the cell possibly contribute to functional changes. See CELLULAR IMMUNOLOGY.

Effects on plants. Plants were the first organisms to demonstrate a biological role for gravity. Experimental interest centered on gravitropism, the differential growth induced by gravity that causes shoots to grow up and roots to grow down. This process allows plants to correct for changes in direction imposed by the environment, such as roots growing around rocks. *See* PLANT MOVEMENTS.

Although seeds and plants have been flown in space for over 40 years, experiments have been infrequent and experimental conditions limiting. Major environmental influences due to the physical properties of the space flight environment have been identified. Successes of early plant experiments were influenced by the physical properties of water in the microgravity environment. For example, freefloating water forms into a ball, while contact with a surface causes the water to form a film. These properties led to inappropriate nutrient delivery. Airflow was also limited in early flights due to the lack of convection, which allowed the creation of boundaries around leaves limiting nutrient delivery and increased concentrations of waste products. These problems were overcome by engineering interventions in the design of subsequent habitats to grow plants. Results of experiments on plants in space suggest an extensive direct influence of gravity on all aspects of plant growth, development, and metabolism.

Physiology, orientation, and tropisms. Plants are exquisitely sensitive to gravity. Graviperception (the magnitude of acceleration required to induce a directed growth response) has been calculated at approximately 10^{-3} g in shoots and 10^{-4} g in roots in flight experiments.

Orientation of roots and shoots (**Fig. 3**) is directed by external stimuli. In microgravity, light directs the orientation of wheat and pea shoots while roots remain disoriented. These studies suggest that, at least in these species, gravity directs the orientation of root growth, while gravity and light guide shoot orientation. Both circumnutation (a gravity-induced touch response) and epinastic movements (bending of leaves downward due to growth of their upper



Fig. 3. Shoots in space.

surface) occur in microgravity. See PLANT GROWTH.

Flight experiments indicate that microgravity affects a plant's mineral balance. It has been postulated that cytoplasmic calcium concentration increases, calcium ion adenosine triphosphatase activity is altered, and ion intake by roots is disrupted in microgravity. In fact, calcium has been linked to both the mechanisms of gravitropism and the mitotic defects seen in seedlings grown in microgravity.

Reproduction and development. Space flight does not affect germination, and its effects on growth are unclear. Generally, by gross observation plants appear to grow normally for short periods of time in microgravity. Over extended periods of microgravity, plants slow or stop growing and then die, most often at the flowering stage. Success was finally achieved with small, fast-growing Arabidopsis seeds which were sown in space, completed a full life cycle, and produced fertile seeds in 1982. Growth was not vigorous, few seeds were produced, and the seeds had a significant level of abnormalities. Many of these observations were demonstrated to be the product of the physical properties of the seeds' space environment. Stagnant air layers that form in the absence of convection during microgravity stress the plants by producing localized high humidity, diminished availability of metabolic gases, buildup of harmful gases, and high temperatures at the top of the plant near light sources. Implementation of protocols to compensate for these properties has greatly improved the reproductive development and germination of Arabidopsis such that, in every aspect examined, values were similar to those of ground control plants.

Cell division is generally reduced or inhibited in microgravity. In fact, in one space shuttle experiment, oat seedlings had only one-tenth as many



Fig. 4. Groups of reconstructed gravity-sensing cells from the root tips of 3-day clover seedlings. (a) Cells grown on Earth. Gravity-sensitive particles (amyloplasts) are sedimented to the bottom of the cell. (b) Cells grown in space (microgravity; no gravity direction), with larger particles that display grouping. (c) Cells grown in a clinostat where they are subjected to slow tumbling (gravity in all directions), with particles that are fewer in number and dispersed, showing that slow tumbling does not mimic the effect of space flight on plants. Particle size and number respond to gravity sensitivity, while particle distribution responds to biochemical signals.

dividing cells as ground controls. Carrot and day lily plants grown in microgravity from somatic cells develop a substantial number of binucleate cells, while the ground control cells are uniformly uninucleate. This may be due to a defect in the ability to put down wall material. Chromosomal perturbations, including breakage, translocations, and deletions, are seen in a variety of microgravity-grown plants derived from somatic embryos, tissue-culture plantlets, and seeds. Dosimetry data collected in conjunction with some of these experiments have ruled out radiation as a cause of the damage, and in-flight fixation of the plants eliminates reentry and 1-g ground recovery perturbations as possible causes of the observed chromosomal damage. Microgravity caused disturbances in the mitotic process. Abnormalities are believed to be associated with spindle disturbances, although this has yet to be demonstrated. See CELL DIVISION; CHROMOSOME ABERRATION; MITOSIS.

The pattern and rate of cellular differentiation may be altered by microgravity. Undifferentiated clusters of anise cells differentiate more rapidly in space. Carrot and lettuce embryo roots and grape hyacinth gametophytes show accelerated development as well; and roots of lettuce, *Crepis*, and oat seedlings differentiate prematurely. This could indicate that microgravity accelerates organ formation and development. *See* CELL DIFFERENTIATION.

Cellular effects. The size of plant cells may be increased, decreased, or normal in microgravity. Rounded cells are seen in space-grown shoots of pine and maize and the roots of maize and *Crepis*, possibly due to the increased effect of surface tension and reduced cell-to-cell mechanical pressure in microgravity. Root caps do not regenerate on some plants in microgravity; however, this failure could be associated with the environment of their growth chamber.

Growth on Earth is responsible for the arrangement of at least some cell organelles, for in microgravity the distribution of not only amyloplasts but also endoplasmic reticulum is altered. Amyloplasts in root caps of plants grown in microgravity (**Fig. 4**) are always distributed randomly. Moreover, their structure and starch content are changed. Chloroplasts of space-grown pea plants show morphological changes. *See* ENDOPLASMIC RETICULUM.

Nuclei of space-grown pea and *Arabidopsis* root cells show abnormal distribution and increased amounts of condensed chromatin, which correlates well with the general observation of decreased mitosis and increased chromosomal aberrations in cells grown in microgravity. This suggests reduced functional activity that could indicate an effect on protein synthesis. *See* CELL NUCLEUS.

Mitochondria of microgravity-grown plants are characteristically swollen. This would suggest an effect of microgravity on cellular metabolism. Microgravity-grown seedling root cells contain 50-90% less dictyosomal volume than cells of ground control plants. This correlates with the thinning of cell walls in space.

It has been proposed that cellulose and lignin should be reduced in space because they function in a structural capacity that is unneeded there. Flight experiments have supported this assumption. *See* CELL (BIOLOGY); SPACE FLIGHT.

Ionizing Radiation

The physical environment of space includes a broad spectrum of ionizing radiation, ranging from the ultraviolet to galactic cosmic rays, that is able to penetrate the life-supporting enclosures in which organisms live (habitats, spacecraft, and space suits). In addition, biological specimens are exposed to products derived from interactions of radiation with other biological systems and materials in the spacecraft. These by-products, such as free radicals and superoxides, can impact biological processes.

The components of space radiation consisting of energetic charged particles are the ones of greatest concern in terms of possible biological effects. They can be categorized as galactic cosmic radiation, geomagnetically trapped radiation (the Van Allen belts), and other particles of solar origin. The number of particles arriving from the Sun varies as a function of solar activity, which has a cycle of approximately 11 years. During solar particle events, large numbers of energetic charged particles, mainly protons, are ejected from the Sun and increase the interplanetary radiation levels by several orders of magnitude. *See* COSMIC RAYS; MAGNETOSPHERE; SUN; VAN ALLEN RA-DIATION.

On Earth the atmosphere shields the planet from short-wavelength ultraviolet radiation and lowenergy charged particles. The magnetic field of the Earth shields the planet from low- and mediumenergy charged-particle radiation by deflecting the particles along magnetic field lines, where they remain captured for long periods of time. The effectiveness of this magnetic shielding varies with latitude, being highest at the Equator and disappearing at the magnetic poles. The displacement of the magnetic poles with respect to the geographic poles results in regions of higher radiation (for example, the South Atlantic Anomaly) within some low-altitude spacecraft orbits.

Measurements of galactic cosmic radiation from space probes show that these particles consist of 87% protons (hydrogen nuclei), 12% alpha particles (helium nuclei), and 1% energetic heavy nuclei, also known as HZE (high charge and energy) particles, ranging from lithium to tin. The energy of individual particles is very high, reaching 10²⁰ eV in some instances. Consequently, it is virtually impossible to passively shield against these particles.

Effects of HZE particles. The energy transfer (the number of ionizations) that occurs along the track of a charged particle is identified in terms of linear energy transfer. High linear energy transfer characterizes highly ionizing radiation, while low linear energy transfer characterizes sparsely ionizing radiation. When HZE particles enter cells, they produce intense ionization along their path and generate a cascade of secondary ionization events. The number of ions produced per unit length of track varies with the charge and velocity of the particles. This variation in ion distribution within a cell is an important factor in the induction of biological damage. *See* LINEAR ENERGY TRANSFER (BIOLOGY).

Radiation hazard and exposure risks are expressed in several interrelated units used to define radiation levels and biological effects due to exposure. For example, a rad of radiation corresponds to the absorption of 100 ergs of energy per gram of any material $(10^{-2}$ joule per kilogram, or 10^{-2} gray). However, a unified theoretical understanding of the radiation quality of cosmic radiation has not been achieved. *See* RADIATION BIOLOGY; UNITS OF MEASUREMENT.

The major radiation hazards for living systems occurring during exploration-class missions outside the Earth's magnetosphere are due to solar-particle-event protons and to the HZE particles in galactic cosmic rays. These hazards are complicated by the fact that there may be a much lower probability of observing the biological effects of HZE particles (and, to some extent, of protons and helium nuclei) than of lightly ionizing radiation (x-rays, gamma rays).

The biological effects of radiation depend on complex energy transduction factors (including energy deposition or dose, dose rate, ionization density, radiochemical intermediaries, molecular biology lesions, and cellular response mechanisms and their regulation and expression in the intact organism). The difference in biological effects due to x-rays and gamma rays on the one hand and heavy charged particles on the other hand are largely derived from the much denser local pattern of ionization around charged-particle trajectories. For lightly ionizing radiation (x-rays and gamma rays), effects are mainly dependent on dose and dose rate. For HZE particles, the concept of dose is inadequate; in particular, the density of energy deposited around a single track is high enough that there is no such thing as a low dose, even for a single particle. See RADIATION INJURY (BIOLOGY).

Measurement. Measurements of radiation on board spacecraft are performed with a large variety of detectors. Dosimeters measuring the average energy deposited by radiation are used both inside and outside spacecraft in order to understand the way in which spacecraft materials modify the radiation field. Materials suffering permanent (or long-term) changes in their physical properties due to radiation are used as so-called passive dosimeters. Examples are film emulsions, certain plastics, and thermoluminescent materials. Active dosimeters are radiation detectors whose response to radiation can be monitored electronically as it occurs. These detectors are used to provide ongoing warnings of radiation hazards during space flight activities. See DOSIMETER.

Since the biological effects of HZE particles are strongly dependent on the particle type and energy (which result in variations in the linear energy transfer), instruments to detect these particles need to identify their charge and energy. Such instruments are called particle identification spectrometers. They generally consist of stacked semiconductor detectors in conjunction with some velocity-sensitive detector. The characteristic signature identifying the charge of the particle is given by its pattern of energy loss as it traverses the detector stack; for particles stopping in the stack, the total energy deposition is also a measure of the particle energy. *See* JUNCTION DETECTOR.

A simplified approximation of particles and their spectra is obtained by use of a low-pressure, gas-filled ionization chamber of precisely defined geometry designed to have the same electron density as the tissue of humans. The ionization produced in the chamber is proportional to the energy loss of the incident radiation. Theoretical models are then used to interpret these measurements in terms of the biological impact of the radiation. *See* IONIZATION CHAMBER; PARTICLE DETECTOR.

Ground-based experiments. Ground-based experiments provide the means to study the effects of

individual components of space radiation in a 1g environment uncomplicated by all of the other physical factors of space flight, including microgravity. The availability of a ground-based accelerator facility to simulate galactic cosmic-ray irradiation by means of accelerated beams of HZE nuclei is essential for the theoretical understanding of the phenomena involved. Ground-based and space-flight studies of HZE effects have revealed that radiobiological mechanisms of densely ionizing radiation are qualitatively different from those of sparsely ionizing radiation. Ground-based experiments have shown that x-ray irradiation followed by heavy ion exposure is far more damaging to an organism than irradiation in the reverse order. *See* PARTICLE ACCELERATOR.

Experiments in space. The HZE-particle component of galactic cosmic radiation was discovered in 1948, and in 1952 it was predicted that humans exposed to these particles could experience a visual light flash sensation. The crews of Apollo 1 and subsequent missions reported experiencing a visual light flash. Microscopic examination of the eyes of rats flown on space flights indicates that light flashes and eye damage result from cosmic rays passing through the eye retina. The particles stimulate the photoreceptors of the eye as they pass through the tissue and are thus perceived as light. Even more important to the understanding of HZE particle effects, however, is the finding that HZE particles produce channellike or tunnel lesions. Ground-based accelerator experiments show that each tunnel lesion is produced by a single HZE particle.

Packaged series of monolayers of bacterial spores, plant seeds, animal eggs, and lower animals interlayered with physical track detectors permit evaluation of individual tracks, and allow identification of each penetrating particle and determination of its possible effects on biological matter in its path. It is found that the very high concentration of absorbed energy produced by a single HZE particle of cosmic radiation can have serious biological effects. Mutations are produced in corn seeds, but in general bacterial spores and seeds are highly resistant. The eggs of the brine shrimp Artemia salina are extremely sensitive. An egg that survives a hit by an HZE particle demonstrates abnormal growth and behavior when it continues development. Two species of bacteria show reduced viability and antibiotic sensitivity that is believed to be due to defective DNA repair mechanisms. Space radiation also results in an increase in sex-chromosome-linked mutations in the fruit fly, Drosophila melanogaster. Radiobiological models derived from ground-based models fall significantly short of reproducing the observed space effects. See MUTATION.

Both synergistic and antagonistic relationships are observed between space radiation and the microgravity environment. However, embryonic development is more sensitive to HZE particle damage under microgravity conditions than inside a centrifuge simulating the acceleration of gravity at the Earth's surface (1 g). When the eggs of the stick insect, *Carausius morosus*, are exposed in space both at 1 g and in a weightless environment, statistically significant differences in development are observed between them. These results indicate that microgravity affects biological processes directly at the cell level, and that these changes in turn modify or interfere with the expression of the effects of the various components of space radiation. It has been proposed that microgravity may affect cell repair and recovery processes, as well as cytoskeleton and membrane functions, thereby making cells more sensitive to radiation damage. Charles Edwin Wade

Bibliography. H. J. Appell, H. Hoppeler, and H. Oser, Muscle research in space, Int. J. Sports Med., Suppl. 4, 18:S255-S334, 1997; A. Berthoz and A. Güell (eds.), Space neuroscience research, Brain Res. Rev., 28:1-233, 1998; Cell and molecular biology research in space, FASEB J., 13:S1-S178, 1999; International Workshop on Bone Research in Space, Bone, 22:71S-161S, 1998; R. Letaw, R. Silberberg, and C. H. Tsao, Radiation hazards on space missions, Nature, 303:709-710, 1988; S. A. Moody and C. Golden, Developmental biology in space: Issues and directions in the era of the International Space Station, Dev. Biol., 228:1-5, 2000; A. Mortimer, International Workshop on Radiation Biology, Mutation Res., 430:165-7, 1999; A. Sievers, B, Buchen, and T. K. Scott (eds.), Plant Biology in Space, Planta, Supplement to vol. 203, 1997; K. Souza, G. Etheridge, and P. X. Callahan, Life into Space: Ames Research Center and Kennedy Space Center 1991-1998, NASA/SP-2000-534, 2000; K. Souza, R. Hogan, and R. Ballard, Life into Space: Ames Research Center 1965-1990, NASA/RP-1372, 1992.

Space charge

The net electric charge within a given volume. If both positive and negative charges are present, the space charge represents the excess of the total positive charge diffused through the volume in question over the total negative charge. Since electric field lines end on electric charge, the space-charge density ρ may also be defined in terms of the divergence of the electric field **E** or the laplacian of the electric potential *V* by Eq. (1) (Poisson's equation). Here ϵ is

$$-\frac{\rho}{\epsilon} = -\operatorname{div} \mathbf{E} = \nabla^2 V$$
$$= \frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2}$$
(1)

the dielectric constant of the medium and x, y, and z are rectangular coordinates defining the position of a point in space. If, under the influence of an applied field, the charge carriers acquire a drift velocity v, the space charge becomes j/v, where j is the current density. For current carried by both positive and negative carriers, such as positive ions and electrons, the space charge density is given by Eq. (2). Here the

$$\rho = \frac{j_+}{v_+} - \frac{j_-}{v_-} \tag{2}$$

subscripts + and - indicate the current density and

drift velocity for the positive and negative carriers, respectively. Thus a relatively small current of slowmoving positive ions can neutralize the space charge of a much larger current of high-velocity electrons. *See* CHILD-LANGMUIR LAW. Edward G. Ramberg

Space communications

Communications between a vehicle in outer space and Earth, using high-frequency electromagnetic radiation (radio waves). Provision for such communication is an essential requirement of any space mission. The total communication system ordinarily includes (1) command, the transmission of instructions to the spacecraft; (2) telemetry, the transmission of scientific and applications data from the spacecraft to Earth; and (3) tracking, the determination of the distance (range) from Earth to the spacecraft and its radial velocity (range-rate) toward or away from Earth by the measurement of the round-trip radio transmission time and Doppler frequency shift (magnitude and direction). A specialized but commercially important application, which is excluded from consideration here, is the communications satellite system in which the spacecraft serves solely as a relay station between remote points on Earth. See COM-MUNICATIONS SATELLITE; DOPPLER EFFECT; MILITARY SATELLITES; SATELLITE (SPACECRAFT); SCIENTIFIC AND APPLICATIONS SATELLITES; SPACE FLIGHT; SPACE NAV-IGATION AND GUIDANCE; SPACE PROBE; SPACECRAFT GROUND INSTRUMENTATION; TELEMETERING.

General features. Certain characteristic constraints distinguish space communication systems from their terrestrial counterparts. Although only line-of-sight propagation is required, both the transmitter and the receiver are usually in motion. The movement of satellites relative to the rotating Earth, for example, requires geographically dispersed Earth stations to achieve adequate communication with the spacecraft on each orbit. The signal power received is expressed by

$$P_r = KP_t G_t G_r \left(\frac{\lambda}{4\pi R}\right)^2$$

where P_r = received signal power; P_t = transmitter power output; G_t = transmitter antenna gain; G_r = receiver antenna gain; K = a factor to allow for system losses, such as cable and connector losses, or losses due to antenna pointing error; R = range (the distance between transmitter and receiver); and λ = wavelength of the transmitted signal in the same units as the range.

Because enormous distances are involved (over a billion miles to the planets beyond Jupiter), the signal received on Earth from deep-space probes is so small that local interference, both artificial and natural, has to be drastically reduced. For this purpose, the transmitted frequency has to be sufficiently high, in the gigahertz range, to reduce noise originating in the Milky Way Galaxy (galactic noise background). The receiver site must be remote from technologically advanced population centers to reduce artificial noise,



Fig. 1. Goldstone 210-ft (64-m) antenna. (Jet Propulsion Laboratory)

and at a dry location such as Goldstone, California (Fig. 1), or in the Australian desert, to avoid precipitation attenuation of the radio signal as well as the higher antenna thermal noise associated with higher atmospheric absolute humidity and relatively warm cloud droplets. The receiving antennas must be steerable and large, typically 85 ft (26 m) or at times 210 ft (64 m) in diameter, to enhance the received signal strength relative to the galactic noise background. Special low-noise preamplifiers such as cooled masers are mounted on the Earth receiver antenna feed to reduce the receiver input thermal noise background. Sophisticated digital data processing is required, and the ground-receiver complex includes large high-speed computers and associated processing equipment. See MASER; PREAMPLIFIER; RADIO RE-CEIVER; RADIO TELESCOPE.

The spacecraft communications equipment is constrained by severe power, weight, and space limitations. Typical communications equipment mass ranges from 25 to 220 lb (12 to 100 kg). Another major challenge is reliability, since the equipment must operate for years, sometimes for decades, unattended, in the difficult radiation, vacuum, and thermal environment of space. Highly reliable components and equipment have been developed, and redundancy is employed to eliminate almost all single-point failures. For example, it is not unusual to have as many as three redundant command receivers operating continuously, because without at least one such receiver in operation no command can get through, including a command to switch from a failed command receiver to a backup radio. Power can be saved by putting some or all of the redundant radios on timers, and to switch to a backup receiver if no commands have been received through the primary receiver within a predetermined interval; but the saved power may come at the cost of a possible delay in emergency response initiation.

Spacecraft power is always at a premium, and other techniques must also be used to minimize its consumption by the communication system. The transmitter is a major power consumer, so its efficiency must be maximized. All aspects of data transmission must contribute to error-free (very low bit error rate) reproduction of the telemetry data using no more power or bandwidth than is absolutely essential. Pulse-code modulation is a common technique which helps meet this goal. In general terms, space communication systems are far less forgiving than terrestrial systems and must be designed, constructed, and tested to much higher standards. *See* PULSE MODULATION; SPACE POWER SYSTEMS; SPACE TECHNOLOGY.

Before describing several typical space communication systems, a brief examination of some important subsystems is in order.

Data sources. The kinds of data that must be transmitted by the space communication system vary, but certain basic types (internal spacecraft environment, scientific, and applications) may be distinguished.

Internal spacecraft environment (housekeeping) data. Onboard sensors provide information from selected locations about a number of functions of internal spacecraft behavior, such as switch position, temperature, pressure, acceleration, voltage, and current. Earth-based monitoring of such data is essential to detect and correct or bypass malfunctions as well as to evaluate trends such as battery, solar cell, and thermal coating degradation.

Scientific data. These data contain the information needed to accomplish the objectives of the mission being flown. For the most part, they originate in sensors of electromagnetic radiation at wavelengths as long as radio waves, ranging to photons as energetic as primary cosmic rays. The sources of this radiation vary from violent astrophysical events which occurred in deep space as long as 10 billion years ago, to solar flare outbursts whose timing and duration over the 23-year solar magnetic-field (sunspot) cycle are still very difficult to understand and to predict. Many of these data are collected and transmitted to Earth in a raster-scan image similar to that used on computer and television displays. Like terrestrial television, the desired data bandwidths are often in the multimegahertz range. Deep-space probes, for which such unchallenged bandwidth needs would produce prohibitively massive and costly spacecraft, trade more transmission time for less bandwidth. Should the mission be a planetary flyby, the data can be acquired in a few hours, stored in solidstate broadband recorders in the probe, and transmitted relatively slowly over the following days or weeks. The ratio of the data collection time interval to the transmission time interval equals the ratio of the required transmission bandwidth to the original data collection bandwidth. Additional bit-errorrate reduction may be accomplished by the statistical treatment of successive transmissions of the same image from the space probe.

Another class of scientific investigations has to do with the immediate external environment of the spacecraft, particularly involving the measurement of the complex terrestrial and solar vector (magnitude and direction) magnetic fields and associated energetic plasma (ionized particles). In lower-altitude denser terrestrial and other planetary atmospheres, measurements of the neutral and ionized atmospheric composition, pressure, temperature, and density are also valuable. Such investigations require telemetry bandwidths in the range of tens of kilohertz, substantially less than that which is required for images.

Applications data. A third category includes data which are supplied regularly as a service to users of images of Earth's land and water surfaces (Earth resources), and of the atmosphere (weather and climatology). The former use is exemplified by Landsat 7 images in the 0.45-12.50-micrometer bands, and the latter by the National Oceanic and Atmospheric Administration (NOAA) polar and Geostationary Operational Environmental Satellite (GOES) images in both visible and infrared spectral bands. Such images again use telemetry bandwidths in the multimegahertz range. Another applications data type is information transmitted from data collection platforms on Earth and relayed to ground processing stations via telemetry. Here the data bandwidth is in the kilohertz range. In general, applications data are part of a guaranteed service; that is, if an operational spacecraft fails, another already has been or shortly will be placed in orbit to continue the service.

Antennas. Antennas for space communication may be classified as ground or spacecraft antennas according to their function.

Ground antennas. The use of paraboloidal (dish) reflectors for large-aperture antennas is practically universal. As the size increases, so do cost and structural problems, in step with the volume which is proportional to the aperture diameter cubed. The signal collection area increases only as the diameter squared. The magnitude of reflector surface irregularities sets an upper limit to the usable frequency. Standard reflector diameters in the National Aeronatics and Space Agency (NASA) deep-space network are 85 ft (26 m) and 210 ft (64 m). The feed system has to be carefully designed for optimum performance. For large antennas a common and efficient feed system uses the Cassegrain configuration in which a hyperboloidal reflector is interposed between the focal point of the main antenna and its reflecting surface (Fig. 1). Its main advantage is that the feed horn and associated microwave receiving equipment are near the surface of the paraboloid rather than at its focus, which results in critical operational advantages and higher efficiencies.

Microwave antennas for near-Earth missions usually have smaller diameters, typically 40 or 30 ft (12 or 9 m). Smaller antennas are used for this purpose because they cost considerably less, and larger systems are not needed because the range is so much smaller. For the reception of lower-frequency VHF (very high frequency) signals from near-Earth spacecraft, an array of crossed-dipole Yagi antennas is often used. *See* ANTENNA (ELECTROMAGNETISM); YAGI-UDA ANTENNA.

Spacecraft antennas. An important requirement for a spacecraft antenna is high reliability in the space environment, since it is usually a component in line to a single-point failure of the communication system. For deep-space missions, spacecraft microwave antennas have diameters up to 16 ft (5 m), depending upon transmitter power, data rate, and range involved. Near-Earth spacecraft use both high-gain and omnidirectional antennas to transmit telemetry data, the particular design depending mainly upon data rates, type of orbit, and the receiving system being employed. Nearly all spacecraft use omnidirectional antennas for receiving commands and transmitting low-data-rate housekeeping information. With this mode of operation, if a partial system failure occurs, the problem can be analyzed and corrective actions taken via the command system to keep the spacecraft in operation.

Modulation techniques. Pulse-code modulation (PCM), in which the data are time-multiplexed, sampled, digitized, and transmitted as a binary stream of pulses grouped into code words, is employed almost universally for transmitting scientific data and command signals. This information is usually phase-modulated on a very stable transmitter carrier to optimize system performance. Techniques such as block coding or convolutional coding are often used to improve the system sensitivity by up to 6 dB (a factor of 4 in power). These techniques require more bandwidth for a given data rate and additional complexity of signal-processing equipment.

Spacecraft transponders. The major function of a spacecraft transponder is to provide a means of measuring range and range rate of a spacecraft by a ground station. This information is then used to compute spacecraft orbits and trajectories. The basic principle of operation is to use a phase-locked loop (PLL) to lock an oscillator coherently to the carrier of the received signal by an exact ratio (typically 240:221 for NASA S-band systems). Comparison of the transmitted and received signals at the ground station provides a coherent two-way Doppler-shift measurement which translates directly into range rate. A ranging signal is modulated on the ground transmitter carrier and returned to the ground station via the transponder carrier; the time delay between transmission and reception on the ground provides a means of computing the range of the spacecraft. Ranging modulation is usually in the form of pseudonoise (PN) PCM codes or coherently related subcarriers. By suitable multiplexing, the transponder also performs the functions of a command receiver and a telemetry transmitter. See DOPPLER EFFECT; PHASE-LOCKED LOOPS.

Tracking and Data Relay Satellite System. The Tracking and Data Relay Satellite System (TDRSS) consists of a series of geostationary spacecraft and an Earth terminal located at White Sands, New Mexico. The purpose of TDRSS is to provide telecommunication services between low-Earth-orbiting (LEO) user spacecraft and user control centers (**Fig. 2**). A principal advantage of the system is the elimination of the need for many of the worldwide ground stations for tracking such spacecraft. The *Tracking and Data Relay Satellite (TDRS)* provides no processing of data; rather, it translates received signals in frequency and retransmits them. User orbits are calculated from range and range-rate data obtained through the *TDRS* by using transponders on the user spacecraft.

The first *TDRS* was launched in 1983, and the firstgeneration spacecraft will near the end of their operational life early in the first decade of the new millennium. The upgraded second-generation replacement series was initiated in 2000 with the launch of *TDRS* H (the eighth of the series).

Forward and return link data are transmitted over three basic services in the second-generation TDRSS.

Multiple access. The multiple-access (MA) service provides simultaneous return-link service from as many as 20 LEO user spacecraft, with data rates up to 3 megabits per second (Mbps) for each user. Forward-link service to the user spacecraft is timeshared with a maximum data rate of 300 kilobits per second (kbps), one user at a time. The multipleaccess service is provided at S-band frequencies of 2106.4 MHz for the forward link and 2287.5 MHz for the return link. Multiple-access users operating on the same frequency are discriminated by unique pseudonoise codes and antenna beam pointing. Multiple-access antennas on each TDRS are adaptive arrays in which signals received by each antenna element are transmitted to White Sands on separate carriers. Ground-based processing provides the ability to form multiple antenna beams simultaneously.

S-band single access. The S-band single-access (SSA) service provides return-link data rates up to 6 Mbps for each user spacecraft. The forward S-band access link provides data at 300 kbps. Carrier frequencies are assigned in the bands from 2025 to 2120 MHz for the forward link and 2200 to 2300 MHz for the return link.

Ka- and Ku-band single access. The K-band single-access (KSA) service provides wide-band forward and return access links. Both forward KSA links provide data at 25 Mbps. The return-link data rates are up to 300 and 800 Mbps for the Ku and Ka bands, respectively. The Ku forward link carrier operates at 13.775 GHz, the return link at 15.0034 GHz. Corresponding Ka assignments are 22–23 and 25–27.5 GHz.

Complete systems. Two examples—the automated *Landsat* 7 and the crewed space shuttle—illustrate TDRSS service to modern space systems.

Landsat 7. This applications spacecraft is designed to monitor the resources and environment of Earth and supply these data to users worldwide. The U.S. Geological Service (USGS) and NASA share


Fig. 2. Tracking and Data Relay Satellite System. (Goddard Space Flight Center)

the responsibility for *Landsat* 7 operations, data processing, archiving, and distribution. The Earthobserving sensor system on board *Landsat* 7 is the Enhanced Thematic Mapper Plus (ETM+), optimized for global change studies, land-cover monitoring and assessment, and large-area monitoring. Seven Earth-observing spectral bands are provided at wave-lengths between 0.45 and 12.50 μ m, as well as a panchromatic band. Applications of *Landsat* 7 data include crop and forestry monitoring and resource mapping, land-use mapping, geological feature mapping, surface water coverage, coastal resource mapping, and environmental change monitoring. *See* AGRICULTURE; REMOTE SENSING.

Landsat 7 is controlled by the Mission Operations Center (MOC) at the NASA Goddard Space Flight Center. Communication with this satellite may be divided into two parts: the wideband user data generated by the ETM+ sensor, and the narrow-band housekeeping telemetry and command traffic.

Landsat 7 has the capability to transmit wideband ETM+ data to as many as three ground stations simultaneously. A ground network receives these data via an X-band direct downlink at a data rate of 150 Mbps. The primary receiving station is at the USGS Data Center in Sioux Falls, South Dakota. Processing of the image data is done after receipt at the data center by the *Landsat* 7 processing system (LPS). The image data may be transmitted as it is collected, in real time, or it may be stored on an onboard solid-state recorder (SSR) for delayed playback to the ground. The X-band wideband data system transfer is done by the use of gimbaled X-band antennas (GXA) [**Fig. 3**]. Three frequencies are employed: 8082.5, 8212.5, and 8342.5 MHz. Quadrature phase-shift keying (QPSK) is used, permitting division of each 150-Mbps link into two 75-Mbps data streams, one of which is modulated on the in-phase (I) channel and the other on the quadrature (Q) channel. Each QPSK modulator drives a solid-state power amplifier (SSPA) which transmits the amplified modulated carrier at the appropriate frequency to the ground station through the designated GXA using right-hand circular polarization (RHCP).

Telemetry and commanding of Landsat 7 is done through two S-band transponders which operate continuously in their receive mode. Communications with Landsat 7 can be done either directly from ground stations of the NASA Spaceflight Tracking and Data Network (STDN) or through the TDRSS. Each transponder is equipped with a transmit/receive S-band omnidirectional antenna which is left-hand circular-polarized (LHCP). Each receiver in these S-band transponders simultaneously searches for TDRSS pseudonoise code correlation and STDN continuous-wave (CW) signals greater than the detection threshold. Upon acquisition of either command signal type, the other mode is inhibited until loss of the initial signal and a subsequent return to the search mode occurs.

The S-band telemetry downlink has three data rates: 1.216, 4.864, and 256.0 kbps. This downlink



Fig. 3. Landsat 7 satellite. (Goddard Space Flight Center)

is also used for two-way Doppler tracking. The nominal output frequency is 2287.5 MHz. The telemetry data are phase shift-key (PSK) modulated onto a 1.024-MHz subcarrier. When the recorded data playback is commanded, the 256-kbps SSR data are phase-modulated onto the baseband channel. The PSK subcarrier is summed with the baseband channel data, and this composite signal phase-modulates the transponder transmitter. Command uplinking to the user spacecraft is done through the S-band system at a data rate of 2.0 kbps.

Space shuttle. The space shuttle is a crewed spacecraft system which is used to provide a heavy-lift reusable space transportation capability to low Earth orbit. Since 2000 its primary use has been to support the construction, continued occupation, and long-term microgravity research aboard the International Space Station (ISS). The launch configuration consists of the orbiter, external tank, and solid rocket boosters (SRBs). Shortly after launch, the SRBs are separated from the space shuttle and are recovered by oceangoing tugs for reuse on later flights. The external tank is jettisoned when the main orbiter engines are shut down, and the orbital maneuvering system (OMS) engines provide the principal orbiter propulsion for the rest of the mission. The space shuttle carries up to seven crew and mission specialists, and each orbiter is designed to be reused for up to 100 missions. It carries a mass of up to 65,000 lb (29,000 kg) of payload in the cargo bay (Fig. 4) on ascent and up to 30,000 lb (14,000 kg) on landing. Major cargo includes logistics support items to and from the ISS; massive modules to the ISS during its construction, such as its robotic arm; and automated spacecraft for launch or recovery from orbit for refurbishment on Earth and possible reuse. *See* SPACE STATION.

The Orbiter Communications and Tracking System operates with the NASA ground-based STDN, the TDRSS, the U.S. Air Force Satellite Control Facility (SCF), the *ISS*, and crew members performing extravehicular activities. The shuttle uses UHF-, S-, and Ku-band frequencies.

The Ku-band system is designed to operate as a radar during a space rendezvous and, when not so employed, as a two-way link with the TDRSS. The Kuband antenna is a 3-ft (0.9-m) monopulse paraboloid tracking antenna with a gain of 38 dB and a beam width of 0.36°. This antenna is deployed after the orbiter is in space and the cargo bay doors are opened. The transmitter uses a 50-W traveling-wave tube amplifier. The Ku-band data from the orbiter occupy a 192-kbps channel for voice and operations data, and a channel of up to 50 Mbps for science data. On command, the 50-Mbps channel can be replaced by a 4-MHz analog channel for television transmission. Television is used for monitoring onboard activities, for monitoring the cargo bay area, and by astronauts in extravehicular activities. See TRAVELING-WAVE TUBE.

The S-band communication system is composed of two independent subsystems: the network subsystem and the payload subsystem. The network subsystem provides tracking and two-way communication direct to ground stations or through TDRSS. The payload subsystem provides two-way communication with automated orbiting spacecraft in the vicinity of the orbiter. All S-band antennas are low-gain and flush-mounted. Locations are chosen to favor the desired direction of coverage. The flush antennas are



Fig. 4. Cutaway view of the space shuttle orbiter. (NASA)

overlaid with thermal protective material to enable them to survive the heat of entry.

The network subsystem is designed to be compatible with both the NASA and the Department of Defense (DOD) networks. The NASA-compatible transponders use a transmit-to-receive frequency ratio of 240:221 and are phase-modulated. When transmitting to NASA ground stations, a power output of 2 W is used. Transmissions to the TDRS use 100 W, and the data are convolutionally encoded to increase the link margins. The DOD-compatible transponders use a frequency ratio of 256:205 and transmit to the SCF with a power of 2 W. *See* PHASE MODULATION.

These transponders provide the capability of two digitally encoded voice channels plus a command channel to the orbiter, and a low-data-rate telemetry and two digital voice return channels. A specialpurpose frequency-modulated (FM) transmitter operates on a frequency of 2250 MHz and has a minimum power output of 10 W. This transmitter is used primarily for television video data transmission to ground stations. *See* FREQUENCY MODULATION.

The S-band payload subsystem operates with a large variety of spacecraft communication systems. It is used during the release and recovery of spacecraft by the orbiter and is effective at ranges from a few feet to over a mile. The maximum power output is 2 W, with the ability to transmit at reduced power levels when the spacecraft is nearby.

The UHF communication system is used during extravehicular activities by astronauts and operates at frequencies of 260 and 297 MHz. These transmitters are amplitude-modulated (AM), with direct modulation of voice plus a 5.4-kHz subcarrier of biomedical data. Two orbiter ultrahigh-frequency (UHF) antennas are used, on the inside of the air lock and on the lower portion of the orbiter. *See* AMPLITUDE MODULATION; SPACE SHUTTLE.

John F. Clark; Donald Platt

Bibliography. B. H. Batson (ed.), *IEEE Trans. Commun.*, COM-26:11, November 1978; W. Froehlich, *New Space Network: The Tracking and Data Relay Satellite System*, NASA EP-251, 1986; *Interface Control Document Between Landsat 7 and the Landsat 7 Ground Network (LGN)*, NASA 430-14-01-001-0, 1997; H. B. Poza, TDRSS telecommunications payload: An overview, *IEEE Trans. Aerosp. Electr. Sys.*, AES-15:414-429, May 1979; *Tracking and Data Relay Satellite System (TDRSS) Users' Guide*, Rev. 4, STDN 101.2, Goddard Space Flight Center, January 1980; J. H. Yuen (ed.), *Deep Space Telecommunications Systems Engineering*, 1983.

Space flight

The penetration by humans into the reaches of the universe above the terrestrial atmosphere, and investigation of these regions by automated, remotecontrolled and crewed vehicles.

The purpose of space flight is to provide significant contributions to the physical and mental needs of humanity on a national and global basis. Such contributions fall specifically in the areas of (1) Earth resources of food, forestry, atmospheric environment, energy, minerals, water, and marine life; (2) Earth and space sciences for research; (3) commercial materials processing, manufacturing in space, and public services. More general goals of space flight include expansion of knowledge; exploration of the unknown, providing a driving force for technology advancement and, hence, improved Earth-based productivity; development and occupation of new frontiers with access to extraterrestrial resources and unlimited energy; strengthening of national prestige, self-esteem, and security; and providing opportunity for international cooperation and understanding.

This article focuses on crewed space flight. For discussion of other missions *see* APPLICATION SATEL-LITES; COMMUNICATIONS SATELLITE; METEOROLOG-ICAL SATELLITES; MILITARY SATELLITES; SATELLITE (SPACECRAFT); SATELLITE NAVIGATION SYSTEMS; SCI-ENTIFIC SATELLITES; SPACE PROBE.

To conduct crewed space flight, the two leading space-faring nations, the United States and Russia, formerly the Soviet Union, have developed spacecraft systems and the necessary ground facilities, research and development base, operational knowhow, planning experience, and management skills. In the United States, crewed space programs are conducted by the National Aeronautics and Space Administration (NASA), a federal agency established in 1958 for the peaceful exploration of space. In Russia, crewed space flights have been under the auspices of the U.S.S.R. Academy of Sciences; they are now the responsibility of the Russian Space Agency (Rosaviakosmos, or RKA). The first spacecraft with a human on board, the Vostok 1, piloted by Yuri A. Gagarin, was launched on April 12, 1961, from the Baikonur Cosmodrome in Kazakhstan and returned after completing one revolution of the Earth. The first American space flight of a human took place 3 weeks later, when NASA launched Alan B. Shepard on May 5 on the Mercury-Redstone 3 for a 15-min suborbital test flight.

The early spacecraft of both nations were built for only one space flight. The first multiply reusable space vehicle, the space shuttle, was launched by the United States on April 12, 1981. By December 31, 2006, 906 men and 109 women had flown a total of 254 missions. Of these, the United States alone, in a total of 148 missions, had put 668 men and 101 women into space (some of them more than once). Up until permanent occupancy of space began on November 2, 2000, with the boarding of the International Space Station (ISS) by its first resident crew, humans had spent a total of 11,008 days 3 hours in space (or about 519,525 crew-hours, accounting for crew sizes). By the beginning of permanent residency, the United States had accumulated a total of 1195 days in space (or about 149,174 crew-hours; Table 1).

In November 1987, the 13 member nations of the European Space Agency (ESA) agreed to become the third major power engaging in human space flight. European efforts today focus on a two-pronged program: participation as a partner in the *International Space Station*, primarily with the Columbus Orbital Facility (COF), and extension of the Ariane family of expendable launch vehicles with the heavy-lift carrier Ariane 5. By joining the original partnership behind the space station and assigning astronauts to United States space shuttle missions (as did the European Space Agency and Russia), Canada and Japan have also entered the ranks of space-faring

nations, while other countries, such as Brazil, are preparing to join.

Crewed Spacecraft

A crewed spacecraft is a vehicle capable of sustaining humans above the terrestrial atmosphere. In a more limited sense, the term "crewed spacecraft" is usually understood to apply to vehicles for transporting and sustaining human crews in space for time periods limited by prestored on-board supplies, as distinct from orbital space stations which support theoretically unlimited habitation of humans in space by autonomous systems, crewed maintenance, and periodic resupply.

Reentry. The basic requirements of crewed spacecraft are quite different from those of crewless space probes and satellites. The presence of humans on board necessitates a life-sustaining environment and the means to return safely to Earth. The major common feature of all crewed spacecraft, therefore, is the atmospheric return element or reentry module. It consists basically of a pressure-tight cabin for the protection, comfort, and assistance of the crew, similar to the cockpits of fighter aircraft, but shaped externally in accordance with the desired airflow and its forces, and surrounded by heat-resistant material, the thermal protection system (TPS), to cope with the high frictional and radiative heating accompanying the energy dissipation phase of the atmospheric return flight to Earth.

To better deal with this heating, the design of the forward-facing portion of the reentry module is usually relatively blunt and symmetrical. Temperatures on and around the stagnation point of the blunt face can reach 3000° F (1700° C) and more. When they are too high for metallic materials, a class of materials called ablators is used. Ablators can be composed of phenolic, silicone rubber, or epoxy compounds. At high temperatures, a great deal of heat is absorbed and carried away from the body by outgassing, charring, and ablating. The remaining heat influx can be dealt with by the insulating properties of the thermal protection system to keep the backface temperatures within the limits of conventional structural materials. *See* NOSE CONE.

Ballistic reentry. A ballistic reentry, flown with a blunt, symmetrical heat shield, is the simplest, most direct mode of return from space, but it also incurs the highest heating on the thermal protection system and the heaviest deceleration loads on the crew since the orbital energy has to be dissipated in the short time of the plunge. This mode also allows no further control of the landing point once the atmosphere has been entered. The first generation of spacecraft of both the United States and the Soviet Union was designed for ballistic reentry and landing by parachute. While the United States since 1981 has gone on to lifting reentry for human space flight, Russia still depends on ballistic reentry with the Soyuz spacecraft for crewed missions and the automated Progress drones for logistics flights.

Lifting reentry. To provide more flexibility to the spacecraft for selecting its touchdown point and to

June 14-19, 1963

June 16-19, 1963

1 lb = 0.45 kg.

Vostok 5

Vostok 6

1 mi = 1.6 km.

Valery F. Bykovsky

Valentina V. Tereshkova

1 ft = 0.3 m.

10,408 lb

10,392 lb

45 rev 143 mi

1 yd = 0.9 m.

76 rev 146 mi 119 h 6 min

70 h 50 min

1 lb/in.² = 6.9 kPa.

1 fathom = 1.8 m.

New time and distance space-flight records set, last 45 revolutions in tandem with *Vostok 6*. Radio contact with *Vostok 6*. Direct TV broadcast to Earth in real time. Extensive medical and biological

Orbited first woman and first person not trained as a pilot. Passed within proximity of *Vostok* 5 on first orbit, in different orbital plane. Radio contact with *Vostok* 5. Direct TV broadcast to Earth in real

time. Parachuted separately from spacecraft on landing.

experiments conducted.

TABLE 1. Crewed s	space flights 1961-2006			
Dates launched and recovered	Designation and crew	A. Weight B. Revolutions C. Max. distance from Earth	Duration	Remarks
Apr. 12, 1961	<i>Vostok 1</i> Yuri A. Gagarin	10,419 lb 1 rev 203 mi	1 h 48 min	First person in space, conducted radio and TV communication with Earth. Spacecraft contained life-support systems, telemetry ejection seat, and recovery system. Cosmonaut and spacecraft landed at preselected area. Automatic control with cosmonaut available as backup. Two-gas air supply at sea-level pressure.
May 5, 1961	Freedom 7 Alan B. Shepard, Jr.	2855 lb — 116 mi	15 min	Mercury spacecraft launched in ballistic suborbital trajectory by Redstone booster (Mercury-Redstone 3). Downrange distance 302 mi. First United States person in space; weightless for 5 min with no ill effects. Astronaut exercised manual control of spacecraft
July 2–21, 1961	Liberty Bell 7 Virgil I. Grissom	2836 lb — 118 mi	15 min	Mercury spacecraft launched into ballistic trajectory by Redstone booster (Mercury-Redstone 4). Downrange distance 303 mi Weightless for 5 min with no ill effects. Hatch opened prematurel during recovery, spacecraft filled with water and sank in 2500 fathoms in Atlantic Ocean. Astronaut was recovered.
Aug. 6–7, 1961	<i>Vostok 2</i> Gherman S. Titov	10,432 lb 16 rev 152 mi	25 h 11 min	First test of prolonged weightlessness; cosmonaut ate, worked, and slept in space. Monitored by TV and radio. Vestibular disturbances produced motion sickness but apparently no significant aftereffects
Feb. 20, 1962	Friendship 7 John H. Glenn, Jr.	2987 lb 3 rev 162 mi	4 h 55 min	Mercury spacecraft launched into orbit by Atlas booster (Mercury Atlas 6). First United States crewed orbital flight, achieving origina Project Mercury objectives of placing a human in orbit, observing reactions to space environment, and making a safe recovery. No adverse physiological effects from the 4.5 h of weightlessness.
May 24, 1962	Aurora 7 M. Scott Carpenter	2975 lb 3 rev 167 mi	4 h 56 min	<i>Mercury</i> spacecraft launched into orbit by Atlas booster (<i>Mercury Atlas 7</i>). Flight plan similar to <i>Friendship 7</i> . Manual attitude contro exercised during retrofire. Landed 250 mi beyond predicted impac point.
Aug. 11–15, 1962	<i>Vostok 3</i> Andrian G. Nikolayev	10,412 lb 60 rev 146 mi	94 h 22 min	Launched as first half of tandem mission with <i>Vostok 4</i> . Radiation measurements taken and effects of biological specimens studied Extensive physiological measurements recorded during prolonged exposure to space environment with no adverse effects noted. The monitoring, radio and TV contact with <i>Vostok 4</i> . Cosmonaut floated free of restraint for a total of 3.5 h, was ejected to land separately from spacecraft.
Aug. 12–15, 1962	<i>Vostok 4</i> Pavel R. Popovich	10,425 lb 45 rev 147 mi	70 h 57 min	Launched from same facility as <i>Vostok 3</i> into tandem flight. Passed to within 4 mi of <i>Vostok 3</i> at closest point. Extensive radiation and physiological experiments conducted. TV monitoring, radio and TV contact with <i>Vostok 3</i> . Cosmonaut floated free of restraint for a tota of about 3 h, was ejected to land separately from spacecraft.
Oct. 3, 1962	Sigma 7 Walter M. Schirra, Jr.	3029 lb 5 rev 176 mi	9 h 13 min	Mercury spacecraft launched into orbit by Atlas booster (Mercury Atlas 8). Automatic reentry mode utilized, landed 9000 yd fron prime recovery ship for first recovery in Pacific Ocean. Experiments included camera, radiation sensors, and ablative samples on the neck of spacecraft.
May 15–16, 1963	Faith 7 L. Gordon Cooper, Jr.	3033 lb 21 rev 166 mi	34 h 20 min	Mercury spacecraft launched into orbit by Atlas booster (Mercury Atlas 9). Long-endurance mission with no adverse physiological ef fects as astronaut ate, slept, and worked in the space environment Reentry sequence was with astronaut orienting spacecraft and firing retrorockets manually. A very accurate landing was made in the Pacific Ocean 7000 yd from prime recovery ship. Experiments in cluded infrared and standard photography, radiation measurement ejection of flashing light (sighted in fifth and sixth orbits), and obser vation of lights on the ground.

TABLE 1. Crewed space flights 1961–2006 (cont.)				
Dates launched and recovered	Designation and crew	A. Weight B. Revolutions C. Max. distance from Earth	Duration	Remarks
Oct. 12–13, 1964	<i>Voskhod 1</i> Vladimir M. Komarov Konstantin P. Feoktistov Boris B. Yegorov	11,731 lb 15 rev 254 mi	24 h 17 min	First multicrewed space flight; first "shirt-sleeve" environment in space flight, no pressure suits; record altitude for crewed space flight. Feoktistov, scientist, and Yegorov, physician, as experimenters and observers in addition to Komarov as pilot. Both Feoktistov and Yegorov experienced motion sickness and disorientation illusions. Crew worked, ate, drank, and slept unrestrained. Crew remained in spacecraft through landing.
Mar. 18–19, 1965	Voskhod 2 Pavel I. Belyayev Aleksei A. Leonov	12,529 lb 16 rev 308 mi	26 h 2 min	New record altitude for crewed space flight. During second orbit Leonov took first space walk with autonomous life-support system, moved 17 ft from spacecraft on tether, 23 min in space environment, including 11 min in air lock through which he exited from spacecraft. While outside, Leonov was observed through the spacecraft TV camera. First Soviet landing sequence using manual system in lieu of automatic. Crew remained in spacecraft through landing. Planned landing point was overshot: landed in deep snow.
Mar. 23, 1965	<i>Gemini 3</i> Virgil I. Grissom John W. Young	7111 lb 3 rev 140 mi	4 h 53 min	<i>Gemini</i> spacecraft launched into orbit by modified Titan 2 booster (<i>Gemini-Titan 3</i>). First United States two-person space flight, Grissom first person in space for second time. One orbital maneuver was conducted in each of the three orbits. Manual control was exercised throughout the reentry phase, using the limited lifting characteristics of the spacecraft to steer toward touchdown.
June 3–7, 1965	<i>Gemini 4</i> James A. McDivitt Edward H. White II	7879 lb 62 rev 184 mi	97 h 56 min	<i>Gemini</i> spacecraft launched into orbit by modified Titan 2 booster (<i>Gemini-Titan 4</i>). Endurance mission of 4 days to test crew and spacecraft in buildup to longer missions. Eleven scientific and engineering experiments, in addition to basic medical checks on 4 days of exposure to the space environment. White spent 22 min outside the spacecraft in EVA, propelling himself with an oxygen-jet gun. During space walk he was attached to the spacecraft with a 25-ft tether-and-oxygen line; he carried emergency oxygen and camera. Spacecraft computer malfunction necessitated a ballistic reentry sequence.
Aug. 21–29, 1965	Gemini 5 L. Gordon Cooper, Jr. Charles Conrad, Jr.	7947 lb 120 rev 217 mi	190 h 55 min	Gemini spacecraft launched into orbit by modified Titan 2 booster (Gemini-Titan 5). Endurance mission of 8 days confirmed the physiological feasibility of Apollo lunar landing mission. First flight of fuel-cell electrical power system. On Aug. 23 a simulated rendezvous was conducted with a series of four maneuvers through two orbits which raised the perigee to 124 mi and apogee to 192.6 mi. Five medical experiments measured physiological effects; extensive weather and terrain photography was conducted. Voice communication was conducted with Sea Lab 2 under the Pacific Ocean.
Dec. 4–18, 1965	<i>Gemini 7</i> Frank Borman James A. Lovell, Jr.	8076 lb 206 rev 204 mi	330 h 35 min	Gemini spacecraft boosted into orbit by modified Titan 2 booster (Gemini-Titan 7). Endurance mission of 14 days, longest scheduled flight in Gemini program. Station-keeping conducted with booster second stage after separation; spacecraft maneuvered to serve as rendezvous target for Gemini 6. Accurate controlled reentry was made, landing 7.6 mi from target. One astronaut at a time doffed his pressure suit to enhance comfort. No adverse physiological effects from 2 weeks in the space environment. Twenty experiments were incorporated into the mission, ranging from detailed medical measurements and sleep analysis to Earth photography and radiometry.
Dec. 15–16, 1965	Gemini 6 Walter M. Schirra, Jr. Thomas P. Stafford	7817 lb 16 rev 193 mi	25 h 51 min	Gemini spacecraft launched into orbit by modified Titan 2 booster (Gemini-Titan 6). Rendezvous with Gemini 7 as target. A series of posigrade maneuvers began the rendezvous chase and shortened the initial 1200 mi separation distance to 730 mi. During second revolution the distance was reduced to 300 mi with another thruster firing, followed by a plane change to match Gemini 7's orbit. Orbit was adjusted to near circular at beginning of third revolution; as fourth revolution began, the terminal phase of rendezvous was initiated, resulting in approach to within 1 ft of Gemini 7. Station-keeping and fly-around maneuvers were conducted for 5 h 19 min. Controlled reentry to landing within 13 mi of prime recovery ship.

		A. Weight		
Dates launched and recovered	Designation and crew	B. Revolutions C. Max. distance from Earth	Duration	Remarks
Mar. 16, 1966	<i>Gemini 8</i> Neil A. Armstrong David R. Scott	8351 lb 6½ rev 186 mi	10 h 41 min	Gemini spacecraft launched into orbit by modified Titan 2 booster, 1 h 41 min after Gemini-Agena target vehicle (GATV)- Atlas booster combination. Rendezvous with the GATV was accomplished in fourth orbit, 6 h after launch, and first docking of two vehicles in space was achieved 33 min later. After 28 min of flight in the docked configuration, a thruster in the orbital-attitude maneuvering system stuck open and the spacecraft began to spin. <i>Gemini</i> spacecraft was undocked from GATV and, after 25 min, was stabilized by use of reentry control system. Mission terminated early with landing in secondary recovery area in western Pacific.
June 3–6, 1966	<i>Gemini 9a</i> Thomas P. Stafford Eugene A. Cernan	8268 lb 45 rev 194 mi	75 h 21 min	Mission utilizing augmented target docking adapter (ATDA) as the target was placed into orbit on June 1 with an Atlas booster. The <i>Gemini</i> was launched into orbit on June 3 by its modified Titan 2 booster. Rendezvous with ATDA achieved on third orbit. Crew confirmed that shroud covering the docking collar had failed to separate, and docking was canceled. Two additional rendezvous were performed as planned, one using visual techniques, the last from above the ATDA. After 1 h of EVA, Cernan's visor accumulated fog, and communications between Stafford and Cernan were poor. Planned use of the self-contained life-support and propulsion systems to maneuver was canceled. Total EVA was 2 h 7 min. Controlled reentry to 0.42 mi of target.
July 18–21, 1966	<i>Gemini 10</i> John W. Young Michael Collins	8248 lb 43 rev 476 mi	70 h 47 min	GATV-Atlas booster combination launched 100 min before <i>Gemini</i> spacecraft launched into orbit by modified Titan 2 booster. Rendezvous and docking with GATV 10 accomplished in fourth orbit. GATV engine used to propel the docked combination to record attitude, then proper orbit to rendezvous with GATV 8. After separation from GATV 10, <i>Gemini 10</i> effected rendezvous with the passive target. First period of standup EVA was terminated at 50 min when both crew members suffered eye irritation. Second EVA period, after rendezvous with passive GATV, consisted of Collins on umbilical moving to GATV and recovering micrometeorite experiment package.
Sept. 12–15, 1966	Gemini 11 Charles Conrad, Jr. Richard F. Gordon, Jr.	8509 lb 44 rev 853 mi	71 h 17 min	GATV-Atlas booster combination launched 1 h 37 min before Gemini spacecraft was launched into orbit by modified Titan 2 booster within 2-s launch window. Rendezvous accomplished on first orbit using on-board information exclusively. Docking at 1 h 34 min after launch. Docking accomplished twice by each crew member. During umbilical EVA, Gordon removed the nuclear- emulsion experiment package and attached tether between Gemini spacecraft and GATV. Excessive fatigue from these activities caused early termination of EVA at 33 min. Using GATV primary propulsion system, docked configuration was propelled to an altitude of 853 mi. Hatch opened third time for 2 h 10 min standup EVA, during which scheduled photography was accomplished. The two spacecraft undocked, and station-keeping by use of a 100-ft nylon tether was exercised with a slow spin rate imparted to the tethered combination. After separation of the tether, a second rendezvous with GATV was conducted from 25 mi separation. Reentry sequence was fully automatic, with impact 2.9 mi from the aiming point.
Nov. 11–15, 1966	Gemini 12 James A. Lovell, Jr. Edwin E. Aldrin, Jr.	8297 lb 59 rev 187 mi	94 h 35 min	GATV-Atlas booster combination launched 1 h 38 min before Gemini spacecraft was launched into orbit by modified Titan 2 booster. Rendezvous was accomplished in third orbit with docking 4 h 14 min after Gemini launch. Two standup EVA periods, on first and third days, totaled 3 h 34 min. On second day Aldrin conducted 2 h 6 min umbilical EVA, testing restraint devices to overcome body-positioning problems experienced on prior flights. Resting frequently, Aldrin used portable hand rails, foot restraints, and various tethers. He completed 19 tasks, demonstrating that useful EVA work was feasible with proper planning and restraining devices. Gemini and GATV spacecraft undocked, remaining joined by 100-ft tether for station-keeping and gravity-gradient stabilization experiment. Fourteen experiments were conducted. Automatic reentry sequence was used for second time.

		A. Weight		
Dates launched and recovered	Designation and crew	B. Revolutions C. Max. distance from Earth	Duration	Remarks
Apr. 23–24, 1967	Soyuz 1 Vladimir M. Komarov	14,000 lb 18 rev 139 mi	24 h 45 min	After a 2-year hiatus, Soviet Union crewed space flight resumed wit Soyuz 1. During reentry on its eighteenth orbit, after 25.2 h of fligh Soyuz 1's parachutes fouled and the 14,000-lb spacecraft crashed killing Komarov.
Oct. 11–22, 1968	<i>Apollo 7</i> Walter M. Schirra, Jr. Donn Eisele Walter Cunningham	87,382 lb 163 rev 244 mi	260 h 9 min 3 s	First crewed flight of <i>Apollo</i> command and service modules Launched by a Saturn 1B, this was an engineering test flight and wa a complete success. Rendezvous and simulated docking exercise with the Saturn upper stage were practiced. A TV camera wa carried, and a series of in-flight programs were staged by th astronauts, showing their activities within the command module Transmissions were relayed live to home receivers all over the Unite States.
Oct. 26–30, 1968	<i>Soyuz 3</i> Georgi T. Beregovoi	14,000 lb 64 rev 151 mi	94 h 51 min	First successful crewed flight of the <i>Soyuz</i> spacecraft. Accomplished orbital rendezvous with the uncrewed <i>Soyuz</i> 2, which was orbited several days previously. Docking was not attempted. Automatic control was used to bring the two craft within 650 ft of eac other. At the point the pilot assumed manual control, and he mad at least two simulated dockings. Television pictures of the inside of the spacecraft were relayed to home receivers in the Soviet Unio and Europe.
Dec. 21–27, 1968	Apollo 8 Frank Borman James A. Lovell, Jr. William Anders	87,000 lb 10 rev 231,000 mi (lunar module)	146 h 59 min 49 s	First crewed circumlunar orbital mission. Also first use of Saturn launch vehicle for a crewed flight. Remained in an orbit about 69 n above the lunar surface for approximately 20 h (10 revolutions). Th ability to use lunar landmarks for navigation to the lunar landin sites was successfully verified. Communications with Earth wer excellent, and TV pictures of good quality, of both the lunar surfac and the Earth from lunar and translunar distances, were transmitte to the Earth from circumlunar reentry speed. Landing was withi 3 mi of the recovery ship.
Jan. 14–17, 1969	<i>Soyuz 4</i> Vladimir A. Shatalov	14,000 lb 48 rev 158 mi	71 h 14 min	Rendezvoused and docked with <i>Soyuz</i> 5 to form the "firsexperimental space station." Cosmonaut Shatalov in <i>Soyuz</i> performed the actual docking. The two spacecraft remained docke for 4 h 35 min, during which time cosmonauts Khrunov and Yeliseye transferred from <i>Soyuz</i> 4 to <i>Soyuz</i> 5, spending about 1 h outside the spacecraft in the process.
Jan. 15–18, 1969	Soyuz 5 Boris V. Volynov Yevgeny V. Khrunov Aleksei S. Yeliseyev	14,000 lb 50 rev 158 mi	72 h 46 min	After docking with <i>Soyuz 4</i> , cosmonauts Khrunov and Yeliseye transferred to <i>Soyuz 5</i> and cosmonaut Volynov remained in th spacecraft. Details of the transfer operation were transmitte over TV.
Feb. 28–Mar. 13, 1969	<i>Apollo</i> 9 James A. McDivitt David R. Scott Russell L. Schweikart	95,231 lb 151 rev 240 mi	241 h 1 min	Earth orbital flight designed to give the lunar module its first crewe flight test. This vehicle was crewed for 28 h by astronauts Schweika and McDivitt. During this period the hatches of both comman module and the lunar module were opened. Schweikart got out ont the "front porch" of the lunar module and thus was able to exercis the suit and backpack designed for operation on the lunar surface Some curtailment of this activity was deemed advisable becaus of Schweikart's previous nausea. Subsequently, the lunar modul was separated from the command module, and after 6 h 24 mi was rendezvoused and docked with the command module. Th astronauts in the lunar module. Numerous navigation and contro exercises and Earth photography experiments were accomplishee Because of weather conditions in the prime recovery areas, th landing area was moved 500 mi south. Splashdown occurred withi 3 mi of the recovery ship.
May 18–26, 1969	<i>Apollo 10</i> Thomas P. Stafford John W. Young Eugene A. Cernan	96,000 lb 31 rev 251,000 mi (lunar module)	192 h 3 min 23 s	Lunar module tested in lunar orbit. Astronauts Stafford and Cerna entered the lunar module, which separated from the comman module, descended to within 50,000 ft of the lunar surface, an then rendezvoused and docked with the command module, which had remained in an orbit about 69 mi above the lunar surface. Trans- mitted to the Earth were color TV pictures of the lunar surface, of the Earth from lunar and translunar distances, and of the activities of the astronauts inside the command module. This flight successful verified that the <i>Apollo</i> spacecraft was ready to attempt a lunar landing on the next flight.

Dates launched and recovered	Designation and crew	A. WeightB. RevolutionsC. Max. distance from Earth	Duration	Remarks
July 16–24, 1969	Apollo 11 Neil A. Armstrong Michael Collins Edwin E. Aldrin, Jr.	96,000 lb 31 rev 250,000 mi (lunar CM)	195 h 17 min 12 s	The objective of the Apollo program (a crewed lunar landing) was achieved in this flight. Astronaut Armstrong emerged from the lunar module on July 20, 1969, followed shortly afterward by Aldrin. This event was recorded by a TV camera and relayed to home receivers in many parts of the world by satellite. The astronauts gathered samples of the lunar surface material for return to Earth. After deploying a United States flag, they installed a passive seismic instrument and a laser reflector on the lunar surface so that data of scientific interest could be obtained after their return to Earth. The astronauts found that mobility on the lunar surface in the one-sixth gravity environment presented no unanticipated problems, and walking was relatively easy. During the whole flight all the systems functioned correctly and the other parts of the flight followed <i>Apollo 10</i> pattern closely.
Oct. 11–16, 1969	<i>Soyuz 6</i> Georgi S. Shonin Valeri N. Kubasov	14,000 lb 79 rev 144 mi	118 h 42 min	Tested effectiveness of various welding techniques in space. Along with <i>Soyuz 7</i> and 8, tested multiple launch capacity of Soviet Union and performed various scientific investigations.
Oct. 12–17, 1969	Soyuz 7 Anatoly V. Filipchenko Vladislav N. Volkov Viktor V. Gorbatko	14,000 lb — 141 mi	118 h 41 min	Along with Soyuz 6 and 8, tested capacity of Soviet Union for multiple launch and rendezvous; also performed navigation experiments and various scientific investigations.
Oct. 13–18, 1969	<i>Soyuz 8</i> Vladimir A. Shatalov Aleksei S. Yeliseyev	14,000 lb 140 mi	119 h	Along with <i>Soyuz</i> 6 and 7, tested capacity of Soviet Union for multiple launch and rendezvous; also tracked other ships' maneuvers and performed various scientific investigations.
Nov. 14–22, 1969	Apollo 12 Charles Conrad, Jr. Richard F. Gordon, Jr. Alan L. Bean	96,000 lb 45 rev 250,000 mi (lunar CM)	142 h 52 min	More lengthy exploration of the Moon's surface.
Apr. 11–17, 1970	<i>Apollo 13</i> James A. Lovell, Jr. Fred W. Haise, Jr. John L. Swigert, Jr.	97,000 lb — 250,000 mi	142 h 52 min	Rupture of a fuel-cell oxygen tank in service module forced cancellation of the third crewed trip to the Moon. The rupture was caused by high pressure which developed when the failure of two thermal switches responsible for regulating the tank heaters led to abnormally high temperatures within the tank, produced by the combustion of Teflon in the supercritical (cryogenic liquid) oxygen. Lunar module life-support system enabled crew to return to Earth.
June 1–19, 1970	<i>Soyuz 9</i> Andrian G. Nikolayev Vitali I. Sevastyanov	14,000 lb 300 rev 167 mi	424 h 59 min	Predominantly biomedical flight, testing human ability to live and work for prolonged periods of time in spacecraft environment. Earth resources applications experiments were also performed.
Jan. 31-Feb. 9, 1971	Apollo 14 Alan B. Shepard Jr. Stuart A. Roosa Edgar D. Mitchell	98,000 lb 34 rev 250,000 mi (lunar CM)	216 h 2 min	Third lunar landing. Precision touchdown accomplished by Shepard and Mitchell in "Antares." On two EVAs, crew deployed Apollo lunar surface scientific experiment package (ALSEP), completed other scientific tasks. First extraterrestrial use of a two-wheeled cart, modular equipment transporter (MET). New record surface time, 9 h 24 min. Surface samples returned to Earth, 94 lb.
Apr. 19, 1971	Salyut 1 Crewless	40,000 lb — 176 mi	175 days	First Soviet space station. Prototype. Launched crewless. Reentered Oct. 11, 1971.
Apr. 23–24, 1971	Soyuz 10 Vladimir A. Shatalov Aleksey S. Yeliseyev Nikolai Rukavishnikov	14,500 lb 32 rev 165 mi	47 h 46 min	Three-person spacecraft, docked with <i>Salyut 1</i> space station. Remained docked for 5.5 h, but no crew transfer took place, probably because of mechanical difficulties.
June 6–30, 1971	Soyuz 11 Georgi T. Dobrovolskiy Vladislav N. Volkov Viktor I. Patsayev	14,300 lb 385 rev 140 mi	570 h 22 min	Second docking with Salyut 1. Crew boarded Salyut and remained for 23 days 17 h 40 min, breaking previous Soyuz 9 record. After extensive on-board activities, crew transferred to Soyuz for Earth return and was killed during reentry by pulmonary embolisms, caused by rapid decompression of cabin through leaking hatch valve.
July 26–Aug. 7, 1971	Apollo 15 David R. Scott Alfred M. Worden James B. Irwin	107,000 lb 74 rev 250,000 mi (lunar CM)	295 h	Perfect landing by Scott and Irwin in "Falcon" at Hadley Rill. First use of lunar roving vehicle (LRV), driven a total of 17.5 mi during three surface EVAs. Brought back more than 170 lb of samples. First subsatellite, carrying particles and fields (P&F) experiment, launched in lunar orbit by crewed spacecraft. First inflight EVA performed by Worden in deep space on return leg.

Dates launched and recovered	Designation and crew	A. WeightB. RevolutionsC. Max. distance from Earth	Duration	Remarks
Apr. 16–27, 1972	<i>Apollo 16</i> John W. Young Thomas K. Mattingly II Charles M. Duke	103,000 lb 64 rev 250,000 mi (lunar CM)	265 h 51 min 5 s	Landing by Young and Duke in "Orion" at Descartes. Second use of LRV, for a total of 16.6 driven miles. Three EVAs totaling 20 h 14 mi yielded 209 lb of samples. Fourth ALSEP station was deployed o surface and second P&F subsatellite launched in lunar orbit. On retur leg, Mattingly undertook second deep-space EVA to retrieve film from external bay.
Dec. 7-19, 1972	Apollo 17 Eugene A. Cernan Ronald E. Evans Harrison H. Schmitt	103,000 lb 75 rev 250,000 mi (lunar CM)	301 h 51 min	Last <i>Apollo</i> lunar landing, by Cernan and Schmitt, a geologist, i "Challenger" in mountainous region at Taurus-Littrow. Third LRV wa driven 21 mi; fifth ALSEP deployed. Longest lunar stay time (74 59 min 38 s), with three EVAs setting new record of 22 h 5 min 6 s Most samples returned to Earth, 250 lb. Third deep-space film retrieva EVA, by Evans.
Apr. 3, 1973	Salyut 2 Crewless	40,000 lb —	_	Second Soviet space station. Failed in orbit.
May 14, 1973– July 11, 1979	Skylab 1 Crewless	170,000 lb (without payload shroud) — 276 mi	-	First United States (experimental) space station, launched crewless on two-stage Saturn 5. At 60 s after launch, the worksho meteoroid/thermal shield tore loose. Primary solar array wing was loss the other jammed closed by debris. As a result, inside temperature became too high and power output too low for long-term huma habitation. Station was saved through determined effort by NAS. engineers on ground and astronauts in space, to become successfu beyond expectations. Station reentered July 11, 1979.
May 25– June 22,1973	Skylab 2 Charles Conrad, Jr. Joseph P. Kerwin Paul J. Weitz	31,000 lb 404 rev 276 mi	28 days 49 min	Crew launch on Saturn 1B postponed from May 15 to analyz problems. After entering <i>Skylab</i> , crew deployed parasol-like therma blanket on workshop outside through small air lock. On June 7, Conra and Kerwin performed hazardous EVA to repair jammed solar wing Three EVAs totaled 5 h 51 min.
July 28– Sept. 25, 1973	<i>Skylab 3</i> Alan L. Bean Jack R. Lousma Owen Garriott	31,000 lb 858 rev 270 mi	59 days 1 h 9 min	Launch vehicle, Saturn 1B. Crew erected new twin-pole sun shield o EVA, installed new rate gyros. Extended solar observations with Apoll telescope mount (ATM), and 39 Earth observation passes. Three EVA totaled 13 h 44 min. Space manufacturing and maneuvering unit test added to mission schedule.
Sept. 27–29, 1973	Soyuz 12 Vasily G. Lazarev Oleg G. Makarov	14,460 lb 32 rev 216 mi	Approx. 48 h	First flight after <i>Soyuz 11</i> tragedy. Short-duration test of a modifie <i>Soyuz</i> with new hatch/valve design, improved maneuvering an manual control capabilities. Tested new space suit design and cabi suit loop.
Nov. 16, 1973– Feb. 8, 1974	<i>Skylab 4</i> Gerald P. Carr William R. Pogue Edward Gibson	31,000 lb 1214 rev 270 mi	84 days 1 h 16 min	Launch vehicle, Saturn 1B. Last <i>Skylab</i> mission with 4 EVAs for film exchange and mechanical repairs, totaling 22 h 21 min. Missio included first exoatmospheric observation of a comet, Kohoutek.
Dec. 18-26, 1973	Soyuz 13 Pyotr I. Klimuk Valentin V. Lebedev	14,400 lb — 169 mi	7 days 20 h 55 min	Similar to Soyuz 12, with addition of Orion 2 telescope system for sola stellar, and cometary studies. Also included were biomedical studies such as Levka brain-bloodflow device.
June 25, 1974	Salyut 3 Crewless	40,000 lb — 172 mi	214 days	Third Soviet space station. Military type. New solar panel and othe systems.
July 3–19, 1974	<i>Soyuz 14</i> Pavel R. Popovich Yuri P. Artukhin	14,400 lb 169 mi	15 days 17 h 39 min	Crew docked and transferred to Salyut 3 space station. Purpose wa to test the new space station. Stay on board was $14^{1}/_{2}$ days.
Aug. 26–28, 1974	<i>Soyuz 15</i> Gennady V. Sarafanov Lev S. Demin	14,400 lb 32 rev 183 mi	48 h 12 min	Intended as second crew transfer to <i>Salyut 3</i> , but unable to doc because remotely controlled docking system failed.
Dec. 2-8, 1974	Soyuz 16 Anatoly V. Filipchenko Nikolai Rukavishmikov	14,400 lb 140 mi	5 days 22 h 24 min	Soviet test and simulation of ASTP mission, including new compatibl docking system and reduction of spacecraft pressure from 14.7 t 10 lb/in. ² in flight. To approximate the United States docking system for docking tests, a dummy structural ring was taken along and separate in orbit.
Dec. 26, 1974	Salyut 4 Crewless	40,000 lb	770 days	Fourth Soviet space station. Civilian type.

		A. Weight			
Dates launched and recovered	Designation and crew	B. Revolutions C. Max. distance from Earth	Duration	Remarks	
Jan. 11-Feb. 9, 1975	Soyuz 17 Aleksei Gubarev Georgiy Grechko	14,400 lb — 225 mi	29 days 13 h 20 min	Crew docked and transferred to Salyut 4 space station.	
Apr. 5, 1975	Soyuz 18A Vasiley Lazarev Oleg Makarov	14,400 lb — —	Approx. 20 min	Suborbital flight. Mission aborted after launch due to vehicle failure after approximately 20 min of flight.	
May 24–July 26, 1975	<i>Soyuz 18</i> Pyotr I. Klimuk Vitali I. Sevastyanov	14,400 lb — 220 mi	62 days 23 h 20 min	Second crew transfer to <i>Salyut 4</i> , which had completed 3352 revolutions at the time. Flew part of the time simultaneously with ASTP.	
July 15–24, 1975	Apollo-Soyuz Test Project (ASTP) United States: Thomas P. Stafford Vance D. Brand Donald K. Slayton Soviet Union: Aleksely A. Leonov	32,500 lb 138 rev 143 mi 15,000 lb 96 rev	9 days 1 h 28 min 5 days 22 h 31 min	First international link-up in space (docking at 31 mi max. distance from Earth for 1 day 23 h 17 min) and last flight of an <i>Apollo</i> spacecraft. <i>Apollo</i> launched on Saturn 1B $7^{1/2}$ h after <i>Soyuz</i> . First handshake between Stafford and Leonov occurred 3:19 p.m. EDT on July 17. Crews performed interspacecraft transfer four times and conducted joint activities. A total of 31 experiments were performed by American crew, some of them jointly with Soviets, and docking tests with <i>Soyuz</i> were practiced. After final undocking,	
	Valerity N. Kubasov	—		each spacecraft continued on own mission. Soyuz returned July 21.	
June 22, 1976	Salyut 5 Crewless	40,000 lb — 169 mi	412 days	Fifth Soviet space station. Civilian type.	
July 6–Aug. 24, 1976	<i>Soyuz 21</i> Vitaly Zholobov Boris V. Volynov	14,4000 lb — 169 mi	49 days 6 h 24 min	Crew docked on July 7 and transferred to <i>Salyut 5</i> space station, launched June 22. On-board activities included earth resources photography, weather observations; formation of crystals and metals-melting in zero g; plant and insect growth for genetic studies. Crew reportedly suffered sensory deprivation effects. Flight had been preceded by uncrewed <i>Soyuz 20</i> (launched Oct. 17, 1975, docked to <i>Salyut 4</i> on Oct. 19).	
Sept. 15–23, 1976	Soyuz 22 Valery F. Bykovsky Vladimir Aksenov	14,400 lb 127 rev 175 mi	7 days 21 h 54 min	Earth surveillance mission. Test of a special East German camera system for use on space stations.	
Oct. 14–16, 1976	Soyuz 23 Vyacheslav Zudov Valery Rozhdestvenskiy	14,400 lb 32 rev 172 mi	2 days 6 min	Attempt to dock with <i>Salyut 5</i> space station failed. Crew returned to Earth.	
Feb. 7-25, 1977	So <i>yuz 24</i> Viktor V. Gorbatko Yuri Glazkov	14,400 lb 285 rev 176 mi	17 days 17 h 23 min	Crew transferred to <i>Salyut 5</i> on Feb. 8. Space station had mostly military mission with special photo reconnaissance runs.	
Sept. 29, 1977	Salyut 6 Uncrewed	42,000 lb — 221 mi	-	Uncrewed launch of new Soviet space station. Reentered July 29, 1982.	
Oct. 9–11, 1977	Soyuz 25 Vladimir Kovalyonok Valery Ryumin	14,400 lb 32 rev 221 mi	2 days 46 min	Attempt to dock with Salyut 6 failed. Crew returned to Earth.	
Dec. 10–Jan. 16, 1977	Soyuz 26 Yuri Romanenko Georgiy Grechko	14,400 lb 590 rev 221 mi	37 days 10 h 6 min Crew duration: 96 days 10 h	Crew transferred to <i>Salyut 6</i> for a 96-day mission, breaking United States record of 84 days on Mar. 4, 1978. Crew returned on Mar. 16, 1978, with <i>Soyuz 27</i> .	
Jan. 10–16, 1978	Soyuz 27 Vladimir Dzhanibekov Oleg G. Makarov	14,400 lb 1024 rev 219 mi	64 days 22 h 53 min Crew duration: 5 days 22 h 59 min	Crew transferred to <i>Salyut</i> 6 to join <i>Soyuz</i> 26 crew. Resupplied by first crewless logistics flight, <i>Progress</i> 1 (launched Jan. 20). On return to Earth, crew used the "older" <i>Soyuz</i> 26 spacecraft, leaving <i>Soyuz</i> 27 for the long-duration crew (returned Mar. 16).	
Mar. 2–10, 1978	Soyuz 28 Aleksei Gubarev Vladimir Remek	14,400 lb 125 rev 221 mi	7 days 22 h 17 min	Crew transferred to Salyut 6 on Mar. 4 to join Soyuz 26 crew. First non-Soviet cosmonaut: Remek (Czechoslovakia).	
June 16–Sept. 3, 1978	<i>Soyuz 29</i> Vladimir Kovalyonok Alexander Ivanchenko	14,400 lb 125 rev 230 mi	79 days 15 h 23 min Crew duration: 140 days	Crew transferred to <i>Salyut 6</i> on June 17 to establish new duration record (140 days). Returned on Nov. 2 with <i>Soyuz 31</i> spacecraft. Resupplied by <i>Progress 2</i> (launched July 7), <i>Progress 3</i> (launched Aug. 7), <i>Progress 4</i> (launched Oct. 3).	

		A. Weight		
Dates launched and recovered	Designation and crew	B. Revolutions C. Max. distance from Earth	Duration	Remarks
June 27–July 5, 1978	<i>Soyuz 30</i> Pyotr I. Klimuk Miroslav Hermaszewski	14,400 lb 125 rev 225 mi	7 days 22 h 4 min	Crew transferred to Salyut 6 to join Soyuz 29 crew. Second non-Soviet cosmonaut: Hermaszewski (Poland).
Aug. 26–Nov. 2, 1978	Soyuz 31 Valery F. Bykovsky Sigmund Jaehn	14,400 lb 1070 rev 221 mi	67 days 20 h 14 min Crew duration: 7 days 20 h 49 min	Crew transferred to <i>Salyut 6</i> to join <i>Soyuz 29</i> crew. Third non-Soviet cosmonaut: Jaehn (East Germany). Crew returned with <i>Soyuz 29</i> spacecraft.
Feb. 25–June 13, 1979	<i>Soyuz 32</i> Vladimir Lyakhov Valery Ryumin	14,400 lb 1711 rev 232 mi	108 days 4 h 24 min Crew duration: 175 days	Crew transferred to <i>Salyut 6</i> to establish new endurance record (175 days). Returned on Aug. 19 with <i>Soyuz 34</i> . <i>Soyuz 32</i> returned uncrewed. Resupplied by <i>Progress 5</i> (launched Mar. 12) and <i>Progress 6</i> (launched May 13).
Apr. 10–12, 1979	<i>Soyuz</i> 33 Nikolai Rukavishnikov Georgi Ivanov	14,400 lb 31 rev 221 mi	1 day 23 h 1 min	Docking with Salyut 6 failed. Crew returned to Earth. Fourth non-Soviet cosmonaut: Ivanov (Bulgaria).
June 6–Aug. 19, 1979	Soyuz 34	14,400 lb 1152 rev 257 mi	73 days 18 h 17 min	Launched crewless and docked to Salyut 6. Returned Soyuz 32 crew. Resupplied by Progress 7 (launched June 28). New type of crewless Soyuz (Soyuz T1) launched Dec. 16 on flight test to dock with Salyut 6.
Apr. 9–June 3, 1980	<i>Soyuz 35</i> Leonid Popov Valery Ryumin	14,400 lb 869 rev 220 mi	55 days 22 min Crew duration: 184 days 20 h 12 min	Crew transferred to <i>Salyut</i> 6 for record-duration mission (184 days). Returned Oct. 11 with <i>Soyuz</i> 37 spacecraft. Resupplied by <i>Progress</i> 8 (launched Mar. 27), <i>Progress</i> 9 (launched Apr. 27), <i>Progress</i> 10 (launched June 29), <i>Progress</i> 11 (launched Sept. 28).
May 26–July 31, 1980	S <i>oyuz 36</i> Valery Kubasov Bertalan Farkas	14,400 lb 1040 rev 220 mi	65 days 20 h 54 min	Crew transferred to Salyut 6 to join with Soyuz 35 crew. Returned June 3 with Soyuz 35 spacecraft. Fifth non-Soviet cosmonaut: Farkas (Hungary).
June 5–9, 1980	Soyuz T2 Yuriy Malyshev Vladimir Aksenov	14,400 lb 62 rev 220 mi	3 days 22 h 41 min	First crewed test of new type spacecraft. Crew transferred to Salyut 6 on June 6. Tenth crewed visit to Salyut 6.
July 23–Oct. 11, 1980	Soyuz 37 Viktor V. Gorbatko Pham Tuan	14,400 lb 1258 rev 220 mi	79 days 15 h 17 min Crew duration: 7 days 20 h 42 min	Crew transferred to <i>Salyut</i> 6 on July 25 to join with <i>Soyuz</i> 35 crew. Returned July 31 with <i>Soyuz</i> 36 spacecraft. Sixth non-Soviet cosmonaut: Pham Tuan (Vietnam).
Sept. 18-26, 1980	<i>Soyuz</i> 38 Yuri Romanenko Arnaldo T. Mendez	14,400 lb 124 rev 220 mi	7 days 20 h 43 min	Crew transferred to <i>Salyut 6</i> to join with <i>Soyuz 35</i> crew. Returned Sept. 26 with <i>Soyuz 38</i> . Seventh non-Soviet cosmonaut: Mendez (Cuba).
Nov. 27–Dec. 10, 1980	<i>Soyuz T3</i> Leonid Kizim Oleg G. Makarov Gennady Strekalov	14,400 lb 202 rev 220 mi	12 days 19 h 8 min	Crew transferred to <i>Salyut 6</i> . Thirteenth crewed visit to <i>Salyut</i> 6. First three-person mission since ill-fated <i>Soyuz 11</i> in 1971. Returned Dec. 10. <i>Salyut 6</i> resupplied by <i>Progress 12</i> (launched Jan. 24).
Mar. 12–May 26, 1981	<i>Soyuz T4</i> Vladimir Kovalyonok Viktor Savinykh	14,400 lb 1183 rev 220 mi (estimated)	75 days	Crew transferred to <i>Salyut 6</i> on Mar. 14. Savinykh became 100th human to go into space. To date: 43 U.S. astronauts, 57 East-bloc cosmonauts.
Mar. 22–30, 1981	<i>Soyuz 39</i> Vladimir Dzhanibekov Jugderdemidiyn Gurragcha	14,400 lb 125 rev 220 mi	7 days 20 h 43 min	Crew transferred to <i>Salyut 6</i> on Mar. 23 to join with <i>Soyuz T4</i> crew. Returned on Mar. 30 with <i>Soyuz 39</i> . Eighth non-Soviet cosmonaut: Gurragcha (Mongolia).
Apr. 12–14, 1981	STS 1 John Young Robert Crippen	205,000 lb 36 rev 172 mi	2 days 6 h 20 min 59 s	First crewed orbital test flight of the U.S. space shuttle <i>Columbia</i> , for a full, successful orbital check of the revolutionary reusable spacecraft and its many subsystems. Return to Earth, featuring atmospheric entry at Mach 25, hypersonic/supersonic maneuvering flight and subsonic unpowered landing at Edwards AFB, California, was celebrated by millions around the world. Time of landing on Apr. 14: 1:20 p.M. EST.

		A. Weight		
Dates launched and recovered	Designation and crew	B. Revolutions C. Max. distance from Earth	Duration	Remarks
May 14–22, 1981	<i>Soyuz 40</i> Leonid Popov Dumitru Prunariu	14,400 lb 124 rev 220 mi	7 days 20 h 41 min	Crew transferred to <i>Salyut 6</i> on May 15 to join with <i>Soyuz T4</i> crew. Returned May 22 with <i>Soyuz 40</i> . Ninth non-Soviet cosmonaut: Prunariu (Romania).
Nov. 12–14, 1981	STS 2 Joe Engle Richard Truly	210,000 lb 36 rev 166 mi	2 days 6 h 13 min	Second crewed orbital test flight of the U.S. space shuttle <i>Columbia</i> . Launch at Cape Canaveral after two postponements and a liftoff delay due to hardware problems. Mission duration reduced from the planned 124 h to 54 h due to malfunction of on- board fuel cell. Despite shortened mission, crew achieved all ob- jectives. Landing on Rogers Dry Lake, Edwards AFB.
Mar. 22–30, 1982	STS 3 Jack R. Lousma C. Gordon Fullerton	210,000 lb 130 rev 150 mi	8 days 4 min	Third crewed orbital test of space shuttle <i>Columbia</i> . Launch at Cape Canaveral after 1 h delay. Fully successful. Flight extended 1 day to enable landing at Northrup Strip, New Mexico, after rains rendered Rogers Dry Lake, California, too soft.
Apr. 19, 1982	Salyut 7 Crewless	42,000 lb — 163 mi	-	Crewless launch of new Soviet space station, replacing <i>Salyut 6</i> (deorbited July 29, 1982). Eleventh anniversary of launch of <i>Salyut 1</i> . <i>Salyut 7</i> and an attached module <i>Cosmos 1686</i> remained in orbit for almost 9 years, reentering on Feb. 6, 1991, after about 50,200 orbits of the Earth.
May 13-Dec. 10, 1982	<i>Soyuz T5</i> Anatoliy Berezovoy Valentin Lebedev	14,400 lb Approx. 3380 r ev (crew) 216 mi	106 days Crew duration: 211 days	Crew transferred to <i>Salyut 7</i> on May 14 and went on to establish new duration record of 211 days. Was host to two visiting crews, <i>T</i> 6 and <i>T7</i> , and returned with <i>Soyuz T7</i> spacecraft.
June 25–July 2, 1982	<i>Soyuz T</i> 6 Vladimir Dzhanibekov Aleksandr Ivanchenkov Jean-Loup Chrétien	14,400 lb 125 rev 194 mi	7 days 21 h 51 min	Crew transferred to <i>Salyut 7</i> on June 25 to join with <i>T5</i> crew. Returned July 2 with <i>T6</i> spacecraft. Tenth non-Soviet cosmonaut: Jean-Loup Chrétien (France).
June 27–July 4, 1982	STS 4 Thomas K. Mattingly Henry W. Hartsfield	210,000 lb 112 rev 150 mi	7 days 1 h 10 min	Fourth and last orbital test flight of <i>Columbia</i> , after 74 workdays of refurbishment. Mission accomplished and flight test program concluded with great success. Landing at Edwards AFB.
Aug. 19–27, 1982	<i>Soyuz T7</i> Leonid Popov Aleksandr Serebrov Svetlana Savitskaya	14,400 lb 125 rev (crew) 194 mi	113 days Crew duration: 7 days 21 h 52 min	Crew transferred to <i>Salyut 7</i> on Aug. 20 to join <i>75</i> crew. Returned Aug. 27 with <i>75</i> spacecraft. Second woman in space: Svetlana Savitskaya.
Nov. 11–16, 1982	STS 5 Vance D. Brand Robert Overmyer Joseph Allen William Lenoir	210,000 lb 81 rev 150 mi	5 days 2 h 14 min	First operational mission of shuttle <i>Columbia</i> successfully delivers reimbursable cargo of two commercial communications satellites (comsats) to orbit. First four-person crew. After landing at Edwards, <i>Columbia</i> returns to manufacturer for changes necessitated by <i>Spacelab</i> payload (<i>STS</i> 9).
Mar. 10–Aug. 14, 1983	Kosmos 1443 Crewless	42,000 lb — 214 mi	157 days	Salyut 7 temporarily enlarged by docking of crewless "Kosmodul" Kosmos 1443 resupply/reentry vehicle.
Apr. 4–9, 1983	STS 6 Paul Weitz Karol Bobko Donald Peterson Story Musgrave	250,000 lb 80 rev 183 mi	5 days 24 min	First flight of space shuttle <i>Challenger</i> featuring first EVA by shuttle astronauts (Musgrave and Peterson) after 9 years. Cargo of 2-ton <i>Tracking and Data Relay Satellite (TDRS)</i> deployed successfully but attained geostationary orbit (GEO) later as planned after major rescue program by NASA (58 days) involving 39 maneuvers by mini-thrust-control jets.
Apr. 20–22, 1983	Soyuz T8 Vladimir Titov Gennady Strekalov Alexandr Serebrov	14,400 lb 32 rev 194 mi	48 h 18 min	Mission aborted. Crew returned after failure of link-up with Salyut 7 space station when rendezvous radar deployed incorrectly.
June 18–24, 1983	STS 7 Robert Crippen Frederick Hauck John Fabian Sally Ride Norman Thagard	250,000 lb 97 rev 187 mi	6 days 2 h 24 min	Shuttle <i>Challenger</i> with first five-person crew including first American woman in space (Ride). Successful deployment of two commercial comsats (<i>Anik C-2, Palapa B-1</i>). First deployment/retrieval of free-flying experiment-carrying platform, <i>Shuttle Palette Satellite</i> (<i>SPAS 01</i>) of West Germany, with Canadian-built robot-arm remote manipulator system.
June 27–Nov. 24, 1983	<i>Soyuz T</i> 9 Vladimir Liakov Aleksandr Aleksandrov	14,400 lb 2390 rev 214 mi	149 days 9 h 46 min	Crew transferred to <i>Salyut 7/Kosmos 1443</i> on June 28, conducted close to 300 experiments during their stay—which was extended unexpectedly when a planned follow-on mission did not come because of explosion of carrier rocket during the launch attempt

TABLE 1. Crewee	d space flights 1961–2006 (con	t.)		
Dates launched and recovered	Designation and crew	A. Weight B. Revolutions C. Max. distance from Earth	Duration	Remarks
June 27–Nov 24, 1983 (<i>cont.</i>)				on Sept. 26. Its crew, Titov and Strekalov, already ill-fated on <i>T8</i> , escaped serious injury. <i>T9</i> crew added solar arrays, delivered by <i>Kosmos 1443</i> , to <i>Salyut 7</i> , and returned on Nov. 24 with <i>T9</i> spacecraft.
Aug. 30–Sept. 5, 1983	STS 8 Richard Truly Daniel Brandenstein Dale Gardner William Thornton Guion Bluford	250,000 lb 97 rev 193 mi	6 days 1 h 8 min	Third flight of <i>Challenger</i> , featuring first night launch (Canaveral) and first night landing (Edwards). First black NASA astronaut in space (Bluford). Successful delivery of Indian comsat <i>Insat 1-B</i> , and test of robot-arm remote manipulation system with 4-ton proof mass. First inflight investigations into space motion sickness phenomenon (space adaptation syndrome; SAS) by physician Thornton.
Nov. 28–Dec. 8, 1983	STS 9/SL 1 John Young Brewster Shaw Byron Lichtenberg Ulf Merbold Owen Garriott Robert Parker	250,000 lb (<i>Spacelab</i> : 26,000) 166 rev 156 mi	10 days 7 h 47 min	First flight of European-built (ESA) <i>Spacelab</i> aboard shuttle <i>Columbia</i> . First six-person crew. First foreign shuttle astronaut: Merbold (West Germany). Work aboard <i>Spacelab</i> on more than 70 experiments in many scientific-technical disciplines went so well that mission was extended by 1 day. Landing at Edwards, after mission had conclusively proved the practical worth of humans in space.
Feb. 3–11, 1984	STS 41-B* Vance D. Brand R. Gibson Bruce McCandless Robert Stewart Ron McNair	250,000 lb 127 rev 185 mi	7 days 23 h 16 min	Tenth flight of shuttle and fourth flight of <i>Challenger</i> , carrying two reimbursable commercial comsats (<i>Westar 6, Palapa B-2</i>). Satellites deployed successfully, but payload assist module (PAM) upper stage failed in both cases. Comsats retrieved later by mission 51-A. First free (untethered) EVA performed by McCandless, wearing manned maneuvering unit (MMU) back-pack system, to 300 ft from orbiter. Second black astronaut (McNair). First landing at Kennedy Space Center, Florida.
Feb. 8–Oct. 2, 1984	Soyuz T10 Leonid Kizim Vladimir Solovyov Oleg Atkov	14,400 lb 3790 rev (crew) 200 mi	63 days Crew duration: 236 days 21 h 50 min	Crew transferred to <i>Salyut 7</i> on Feb. 9 and went on to set new endurance record of 237 days. During the 34-week stay, Kizim and Solovyov made six EVAs and spent almost 23 h outside the station, mainly for major repair work on <i>Salyut 7</i> . Crew took more than 25,000 photographs of Earth and was joined twice by visiting crews (<i>T11</i> and <i>T12</i>). Resupplied by <i>Progress 19</i> on Feb. 23, <i>Progress 20</i> on Apr. 17, <i>Progress 21</i> on May 8, <i>Progress 22</i> on May 30, and <i>Progress 23</i> on Aug. 15. Returned on Oct. 2 with <i>Soyuz T11</i> spacecraft.
Apr. 2–11, 1984	<i>Soyuz T11</i> Juri Malishev Grennady Strekalov Rakesh Sharma	14,400 lb 125 rev (crew) 200 mi	183 days Crew duration: 7 days 21 h 42 min	Crew transferred to <i>Salyut 7</i> to join <i>T10</i> crew. Returned with more than 1000 photographs of India and other Earth areas with <i>Soyuz T10</i> spacecraft. Eleventh non-Soviet cosmonaut: Sharma (India).
Apr. 6–13, 1984	STS 41-C Robert Crippen Richard Scobee George Nelson Terry Hart James van Hoften	250,000 lb (LDEF: 21,000) 107 rev 313 mi	6 days 23 h 41 min	Fifth flight of <i>Challenger</i> . Flew new "direct" ascent to higher- than-usual altitude for rendezvous with solar telescope satellite <i>Solar Maximum Mission (SMM)</i> , inoperative since end of 1980. Crew deployed large Long Duration Exposure Facility (LDEF), then captured and repaired <i>SMM</i> with full success, setting new EVA record of 7 h 18 min.
July 17–19, 1984	<i>Soyuz T12</i> Vladimir Dzhanibekov Sventlana Savitskaya Igor Volk	14,400 lb 196 rev 221 mi	12 days 7 h 14 min	Crew transferred to <i>Salyut 7</i> on July 18 to join <i>T10</i> crew. Second space mission of Savitskaya, who also became world's first female spacewalker (3 h 35 min), performing experiments with electron-beam tool for space assembly and repair. Crew returned with <i>T12</i> spacecraft.
Aug. 30–Sept. 5, 1984	STS 41-D Henry W. Hartsfield Michael Coats Richard Mullane Stephen Hawley Judith Resnik Charles Walker	250,000 lb 96 rev 185 mi	6 days 56 min	First flight of space shuttle <i>Discovery</i> . Cargo: three reimbursable comsats (<i>SBS 4, Telstar 3-C, Leasat 1</i>) successfully deployed. Also: OAST-1 experiment pallet. Second U.S. woman in space (Resnik). First "passenger" from industry (Walker, of McDonnell Douglas).
Oct. 5–13, 1984	STS 41-G Robert Crippen Dave Walker Kathryn Sullivan Sally Ride David Leestma Marc Garneau Paul Scully-Power	250,000 lb 132 rev 220 mi	8 days 5 h 24 min	Sixth flight of <i>Challenger</i> . First seven-person crew, including two women, one Canadian (Garneau), one Australian (Scully-Power). Payload: <i>Earth Radiation Budget Satellite (ERBS)</i> , successfully deployed; OAST-3 experiments. First U.S. woman to walk in space (Sullivan). Second landing at Kennedy Space Center.
1 lb = 0.45 kg.	1 mi =1.6 km. 1 ft = 0.3 m. 1	yd = 0.9 m. 1 fathom =	= 1.8 m. 1 lb/in. ² = 6	6.9 kPa.

*Beginning with 41-B, shuttle fights were no longer numbered consecutively. In the new system, the first digit indicates the last digit of the current fiscal year (begins Oct. 1 of previous calendar year), that is, 4 = 1984. The second digit indicates the launch site: 1 = Kennedy Space Center; 2 = Vandenberg Air Force Base. The letter indicates which serial flight in the fiscal year; for example, B = second flight.

TABLE 1. Crewed space flights 1961-2006 (cont.) A. Weight B. Revolutions Dates launched Designation C. Max. distance from Earth and recovered and crew Duration Remarks Nov. 8-16, 1984 STS 51-A 250,000 lb 7 days 23 h Second flight of Discovery. Two reimbursable comsats (Telesat-H, Leasat/Syncom IV-1) successfully deployed. Two others, Palapa B-1 and Westar 6 from mission 41-B, successfully retrieved and Frederick Hauck 126 rev 45 min Dave Walker 220 mi returned to Earth for repair. Fifth U.S. woman in space (Fisher). Joseph Allen Anna Fisher Third landing at Kennedy Space Center. Dale Gardner Jan. 24-27, 1985 Third flight of Discovery. First dedicated Department of Defense STS 51-C 3 days 1 h Thomas Mattingly mission, classified secret. Fourth landing at Kennedy Space Center. 47 rev 34 min Loren Shriver Gary Payton James Buchli Ellison Onizuka Apr. 12-19, 1985 STS 51-D 250,000 lb 6 days 23 h Fourth flight of Discovery. Payload of two reimbursable comsats (Telesat-I/Anik, Syncom IV-3) successfully deployed, but one Karol Bobko 109 rev 55 min Donald Williams 283 mi (Syncom) failed to ignite its upper stage despite heroic attempts by Margaret Rhea Seddon crew with remote manipulator system. First flight of a U.S. politician Jeffrey Hoffman (Senator Garn, Utah). Second flight of industry representative David Griggs Walker. Sixth U.S. woman in space (Seddon). Landing at Kennedy Charles Walker Space Center resulted in damaged tires. E.J. "Jake" Garn Apr. 29-May 6, 1985 STS 51-B/Spacelab 3 250,000 lb 7 days 8 min Second flight of Spacelab pressurized module. Crew conducted over a dozen major experiments and tested research-animal Robert Overmyer 110 rev Frederick Gregory 220 mi holding facility with rats and two squirrel monkeys. Mission accomplished with great success. Landing of Challenger at Edwards Don Lind Norman Thagard AFB. William Thornton Lodewijk van den Berg Taylor Wang June 6-Sept. 26, Soyuz T13 Vladimir Dzhanibekov 14,400 lb 112 days/ Dzhanibekov: Crew transferred to Salyut 7 space station on June 8, after testing 1985 new flight control systems in Soyuz spacecraft. Performed exten-Viktor Savinykh 168 days 3 h 200 mi sive repair/maintenance of power system on station. Resupplied 51 min/Savinykh by Progress 24 (launched June 21), Cosmos 1669, a new type of free-flying, crewless platform (docked July 21, separated Aug. 29). On Sept. 26, Dzhanibekov returned with "visiting" Georgi Grechko in Soyuz T13, leaving Savinykh with Vasyutin and Volkov in Salyut until Nov. 21. June 17-24, 1985 STS 51-G 250,000 lb 7 days 1 h Fifth flight of Discovery, with cargo of three reimbursable comsats Daniel Brandenstein 111 rev 40 min (Morelos-A/Mexico; Arabsat 1B/Arabia; Telstar 3D/AT&T). Crew John Creighton 240 mi also deployed and later retrieved free-flying instrument platform Shannon Lucid SPARTAN 1 (Shuttle Pointed Autonomous Research Tool for As-Steven Nagel tronomy). Payload included six scientiflc-technological experi-John Fabian ments from U.S., France, and Germany. Two foreign astronauts: Patrick Baudrey Prince Al-Saud (Saudi Arabia), Baudry (France). Seventh U.S. Sultan Salman Al-Saud woman (Lucid). July 29-Aug. 6, 1985 STS 51-F/Spacelab 2 252,800 lb 7 days 22 h Eighth flight of Challenger, first flight of Spacelab pallets-only Charles G. Fullerton 126 rev 45 min version. Lift-off after 1 h 37 min delay due to computer problem. Roy D. Bridges, Jr. 197 mi One of three orbiter engines shut down prematurely (after 5 min Karl G. Henize 45 s) but not early enough to cause mission abort. Flight extended Anthony W. England 1 day tohelp recover from on-board experiment problems. Mission F. Story Musgrave accomplished. Landing at Edwards AFB. Loren W. Acton Aug. 27-Sept. 3, 1985 STS 51-I 337,000 lb 7 days 2 h Sixth flight of Discovery. Liftoff after two launch delays, one Joe H. Engle 17 min (Aug. 24) caused by weather, the other (Aug. 25) by backup 111 rev Richard O. Covey 280 mi computer problem. Payload of three reimbursable comsats [ASC James D. A. van Hoften AUSSAT 1 (Australia), Leasat (Syncom) IV-4] successfully John M. Lounge deployed, the first two on the same day-a first. Crew performed rendezvous with Leasat (Syncom) IV-3, deployed Apr. 13 by STS William F. Fisher 51-D and drifting lifeless, and successfully conducted salvage and repair operation in space on Aug. 31 and Sept. 1, redeploying the comsat for normal operation. Landing at Edwards AFB Crew transferred to Salyut 7/Soyuz T13 on Sept. 18. Veteran Sept. 17-Nov. 21, Soyuz T14 14,400 lb 9 days/Grechko; 1985 Vladimir Vasyutin 64 days 21 h cosmonaut Grechko returned with Dzhanibekov on Sept. 26 in Georgy Grechko 52 min/Vasyutin Soyuz T13, making this the first successful space station crew 200 mi Alexander Valkov and Volkov; rotation. Vasyutin, Volkov, and Savinykh returned suddenly on 168 days 3 h Nov. 21 with Soyuz T14 because of illness of Vasyutin, requiring 51 min/Savinykh hospitalization.

 $1 \text{ lb} = 0.45 \text{ kg}. \quad 1 \text{ mi} = 1.6 \text{ km}. \quad 1 \text{ ft} = 0.3 \text{ m}. \quad 1 \text{ yd} = 0.9 \text{ m}. \quad 1 \text{ fathom} = 1.8 \text{ m}. \quad 1 \text{ lb/in.}^2 = 6.9 \text{ kPa}.$

Dates launched and recovered	Designation and crew	A. WeightB. RevolutionsC. Max. distance from Earth	Duration	Remarks
Oct. 3–7, 1985	STS 51-J Karol Bobko Ronald J. Grabe Robert Stewart William A. Pailes David C. Hilmers	— 63 rev 320 mi	4 days 1 h 45 min	First flight of fourth shuttle orbiter <i>Atlantis</i> , setting new altitud record (after <i>41-C</i>). Payload: military (classifled). Landing a Edwards AFB.
Oct. 30–Nov. 6, 1985	STS 61-A/Spacelab D1 Henry W. Hartsfield Steven R. Nagel Guion Bluford Bonnie J. Dunbar James F. Buchli Reinhard Furrer Ernst Messerschmid Wubbo Ockels	250,000 lb (<i>Spacelab</i> : 26,000) 111 rev 202 mi	7 days 32 min	Ninth flight of <i>Challenger</i> with the West German <i>D</i> -1 mission, the first foreign-chartered <i>Spacelab</i> flight. Crew included three foreign astronauts: Furrer, Messerschmid (both W. Germany), and Ockel (Netherlands). Eighth U.S. woman (Dunbar). First eight-member crew. Mission concluded on Nov. 6 with great success. Landing at Edwards AFB with test of new nose wheel steering.
Nov. 26–Dec. 3, 1985	STS 61-B Brewster H. Shaw Bryan D. O'Connor Sherwood C. Spring Mary L. Cleave Jerry L. Ross Rudolfo Neri Vela Charles D. Walker	261,455 lb 109 rev 235 mi	6 days 22 h 54 min	Second flight of <i>Atlantis</i> , carrying three commercial comsats <i>Morelos-B</i> (Mexico), <i>Aussat 2</i> (Australia), and <i>Satcom KU-2</i> . Ross and Spring conducted major outboard activities demonstrating space-station-type assembly and handling of large structures ir space (payload designations: EASE/ACCESS). One foreign payload specialist: Neri (Mexico). Third flight of industry representative Walker. Ninth U.S. woman (Cleave). First flight of <i>PAM-D2</i> upper stage. Heaviest PAM payload (Satcom). Landing at Edwards AFE after full mission accomplishment.
Jan. 12–18, 1986	STS 61-C Robert L. Gibson Charles F. Bolden, Jr. George D. Nelson Steven A. Hawley Franklin E. Chang-Diaz Robert J. Cenker C. William Nelson	250,000 lb 96 rev 201 mi	6 days 2 h 4 min 9 s	First flight of <i>Columbia</i> after major modifications since its last flight (<i>STS 9</i>) in Nov. 1983. Originally scheduled for Dec. 18, 1985, launch was delayed seven times due to systems problems and inclement weather. Crew included first Hispanic American in space (Chang-Diaz, Costa Rica), a payload specialist from industry (Cenker, RCA), and the second U.S. congressman in space (Florida U.S. Rep. Bill Nelson). Crew deployed advanced <i>Satcom K-1</i> communications satellite, took photographs of Comet Halley, and performed numerous on-board experiments. Landing at Edwards AFB after two postponements due to bad weather at Kennedy Space Center in Florida. Second night landing.
Jan. 28, 1986	STS 51-L Francis R. Scobee Michael J. Smith Judith A. Resnik Ellison S. Onizuka Ronald E. McNair Gregory B. Jarvis Christa McAuliffe	250,000 lb 	1 min 13 s	<i>Challenger</i> was launched at 11:38 a.m. EST after a night of below- freezing temperatures. Its right solid-rocket booster developed a catastrophic leak due to malfunction of a joint seal between two motor elements. The shuttle was destroyed 73 s after liftoff, killing its crew of seven and bringing the United States space program to a halt for 32 months.
Feb. 19, 1986	<i>Mir</i> Crewless	235,000 lb 210 rev —	_	Crewless launch of new Soviet space station <i>Mir</i> (Peace), similar in size and design to <i>Salyut</i> 7, with estimated length of 57 ft. Represents a new-generation station, with six docking ports for future buildup with add-on modules and docking of <i>Soyuz</i> and <i>Progress</i> logistics flights.
March 13–July 16, 1986	Soyuz T15 Leonid Kizim Vladimir Solovyev	14,400 lb 2000 rev (est.) 210 mi	125 days	Crew docked at <i>Mir</i> on Mar. 15 about 49 h after liftoff and transferred to the space station. First resupply flight followed on Mar. 19 with launch of crewless <i>Progress</i> 25 which docked to <i>Mir</i> on Mar. 21. <i>Progress</i> 26 was launched on Apr. 23, and the tanker-transport docked with <i>Mir</i> on Apr. 26. On May 5, Kizim and Solovyev transferred from <i>Mir</i> to <i>Salyut</i> 7 using <i>Soyuz</i> 715. Their stay in the <i>Salyut</i> 7/ <i>Soyuz</i> 715/ <i>Cosmos</i> 1686 complex, trailing <i>Mir</i> in orbit by 15 min, included 3 h 50 min spacewalk on May 28 to demonstrate the ability to work with a 15-m-long triangular lattice. A crewless <i>Soyuz</i> 7 <i>M</i> , launched on May 21, docked with <i>Mir</i> on May 23 and returned to Earth on May 30—a test of a modified version of the <i>Soyuz</i> with improved computers, rendezvous-docking systems, and landing parachute. A second EVA on May 31, 5 h long, also demonstrated structural deployment. <i>Progress</i> 26 undocked or June 22 and deorbited on June 23. On June 26, the crew returned to Earth. Kizim had set a new record of 373 days of cumulative time in space by a human. Landing was 34 mi northeast of Arkalyk

TABLE 1. Grewed space	- mgms 1901–2006 (Cont.)			
Dates launched and recovered	Designation and crew	A. Weight B. Revolutions C. Max. distance from Earth	Duration	Remarks
Feb. 6–July 30, 1987	Soyuz TM-2 Yuri Romanenko Alexander Laveikin	14,400 lb Approx. 145 rev 230 mi	174 days 1 h 22 min/ Laveikin; 8 days 20 h 1 min/Faris and Viktorenko	First flight to <i>Mir</i> in 1987 was the crewless <i>Progress</i> 27 tanker/transport, launched on Jan. 16, and docked to <i>Mir</i> on Jan. 18, to provision the station for the subsequent crew visit. After liftoff in a new, more advanced <i>Soyuz</i> version, crew docked at <i>Mir</i> on Feb. 8 for a long-duration attempt. After an initial zero-g adaptation problem experienced by Laveikin, crew settled down to conduct materials processing work. On Mar. 5, <i>Progress 28</i> , a new tanker/transport, docked with <i>Mir's</i> rear port and was unloaded. On Mar. 9, its engines boosted <i>Mir</i> to increase its orbit altitude.
March 31, 1987	Kvant Crewless	43,000 lb — 250 mi		An astrophysics module, <i>Kvant</i> , launched on Mar. 31, failed to dock to the <i>Mir</i> core on Apr. 5. After a second fruitless attempt on Apr. 9, the crew corrected the problem—a piece of cloth left by the ground crew—in a 3 h 40 min EVA on Apr. 12, and docking was accomplished. <i>Kvant</i> 's propulsion module was jettisoned Apr. 13, making the module's rear docking port accessible. Resupply ship <i>Progress 29</i> , launched Apr. 21, docked to this port on Apr. 23. After unloading and taking on stored waste, it burned up in reentry on May 11. On May 13, <i>Progress 30</i> was launched with fresh supplies and mail. In a second and third EVA, on June 12 and 16, the crew installed a third solar power array on the station.
July 21–Dec. 29, 1987	<i>Soyuz TM-3</i> Alexander Viktorenko Alexander P. Alexandrov Mohammed Faris	14,400 lb Approx. 2576 rev 230 mi	161 days/ Alexandrov; 326 days/ Romanenko; 8 days/ Levchenko	After launch from Baikonur, the crew, which included the first Syrian cosmonaut (Faris), docked with <i>Mir</i> on July 24. On July 29, Viktorenko and Faris, accompanied by Laveikin who had shown an abnormal electrocardiogram reading, undocked <i>Soyuz TM-2</i> and landed in Kazakhstan early on July 30. Romanenko and Alexandrov continued the mission on <i>Mir.</i> Resupply: <i>Progress 31</i> , docked Aug. 7, separated Sept. 22; <i>Progress 32</i> , launched Sept. 22, docked Sept. 26. On Sept. 30, Romanenko broke the current record of 237 days. After separation of <i>Progress 32</i> , resupply ship <i>Progress 33</i> docked on Nov. 23. When Yuri Romanenko returned Dec. 29, 1987, in <i>TM-3</i> with Alexandrov and Anatoliy Levchenko (of <i>TM-4</i>), he had set a new record for the longest crewed space flight to date–326 days.
Dec. 21, 1987–June 17, 1988	Soyuz TM-4 Vladimir Titov Musa Manarov Anatoliy Levchenko	14,400 lb — 230 mi	366 days 22 h 39 min/Titov and Manarov; 7 days 21 h 2 min/ Levchenko	Docked on Dec. 23, after 33 revolutions, to <i>Mir</i> . When, on Dec. 29, Levchenko (a shuttle test pilot), Alexandrov, and Romanenko returned to Earth in <i>TM-3</i> , landing at 12:16 PM. about 50 mi from Arkalyk in Kazakhstan. Romanenko had set a new world record of 326 days in space. Titov and Manarov went on for a long-duration stay in <i>Mir</i> , slightly over 1 year. Resupply ship <i>Progress 34</i> was launched Jan. 21, 1988, and docked early on Jan. 23, <i>Progress 35</i> followed on Mar. 24 (undocked May 5), and <i>Progress 36</i> on May 15.
June 7–Sept. 7, 1988	<i>Soyuz TM-5</i> Anatoly Solovyov Viktor Savinykh Alexander Alexandrov	14,400 lb 162 rev 225 mi	10 days 8 h 5 min (crew)	Docked. to <i>Mir</i> on June 9. The 10-day mission was the 63d crewed flight since Gagarin. Alexandrov was the second Bulgarian (after Georgi Ivanov) and the 13th foreign national flown by the Soviets since 1978. Countdown and launch were widely televised. Visitor crew returned to Earth on June 17 in <i>TM-4</i> , leaving the "fresh" <i>TM-5</i> docked to <i>Mir</i> . On June 18, Titov and Manarov transferred the <i>TM-5</i> from the aft to the forward docking port. During a 5-h EVA on June 30, Titov and Manarov did not succeed in replacing a faulty detector unit on a British/Dutch x-ray telescope outside <i>Mir</i> , because of a broken wrench. <i>Progress 37</i> , the next resupply ship, was launched July 19.
Aug. 29–Dec. 21, 1988	<i>Soyuz TM-6</i> Vladimir Lyakhov Valery Polyakov Abdullah Ahad Mohmand	14,400 lb 138 rev 221 mi	8 days 20 h 27 min/Lyakov and Mohmand; 241 days/ Polyakov	<i>TM-6</i> was launched for a trip to <i>Mir</i> , with a crew including a physician, Polyakov, who was to stay aboard <i>Mir</i> to observe Titov and Manarov, and an Afghan cosmonaut, Mohmand. Docking occurred on Aug. 31. Since Lyakhov had received special training as a rescuer-cosmonaut, he was able to handle <i>TM-6</i> alone. After undocking from <i>Mir</i> in <i>Soyuz TM-5</i> of the previous mission, Lyakhov and Mohmand prepared for reentry at 3 a.m. but became stranded in orbit due to malfunction of a guidance system sensor, followed by computer input errors. With oxygen supply threatening to run out after 48 h, the twice-failed retroburn was achieved 24 h later after preventive measures. <i>TM-5</i> landed on Sept. 7, 100 mi from Dzhezkazgan in Kazakhstan. On Sept. 12, the crewless resupply ship <i>Progress 38</i> docked with <i>Mir</i> after <i>Soyuz TM-6</i> had been moved, on Sept. 8, from the <i>Kvant</i> astrophysics module to the airlock of the station. <i>Progress 39</i> was launched Sept. 9, carrying French equipment for an upcoming visit by a French cosmonaut.

TABLE 1. Crewed space	e flights 1961–2006 (cont.)			
Dates launched and recovered	Designation and crew	A. Weight B. Revolutions C. Max. distance from Earth	Duration	Remarks
Aug. 29–Dec. 21, 1988 (<i>cont</i> .)				In a 4-h EVA on Oct. 20, Titov and Manarov successfully repaired a telescope in the <i>Kvant</i> module, finishing an earlier attempt of June 30.
Sept. 29-Oct. 3, 1988	STS 26 [†] Frederick H. Hauck Richard O. Covey John M. (Mike) Lounge David C. Hilmers George D. Nelson	253,700 lb 645 rev 188 mi	4 days 1 h 57 min	The United States' return to space. After a 32-month hiatus, NASA astronauts were back in space on the orbiter <i>Discovery</i> , after a flawless launch at 11:37 a.m. EDT. At 6 h 13 min into the mission Nelson deployed the prime payload, NASA's second <i>Tracking and Data Relay Satellite (TDRS)</i> for transfer to geostationary orbit. Secondary payloads were 11 experiments, including 2 by students. It was the seventh flight of <i>Discovery</i> , with over 200 engineering changes since the <i>Challenger</i> loss. Touchdown occurred at 12:37 p.M. EDT at Edwards, with a rollout distance of 7451 ft.
Nov. 15, 1988	Buran Crewless	224,000 lb 2 rev 156 mi	3 h 25 min	First flight of Soviet reusable space shuttle <i>Buran</i> (Blizzard). Launched by second flight of Energia launch vehicle at 6 a.m. local time, landed two orbits later on runway at Baikonur, 7.5 mi from launch pad.
Nov. 26, 1988–Apr. 27, 1989	Soyuz TM-7 Alexandrovitch Volkov Sergei Krikalev Jean-Loup Chrétien	14,400 lb — 230 mi	25 days/ Chrétien; 152 days/ Volkov and Krikalev	Second flight of French cosmonaut Chrétien to a Soviet space station. Mission, named "Aragatz," docked on <i>Mir</i> on Nov. 28. Chrétien became first French citizen to perform an EVA when he installed the deployable antenna structure ERA in space on Dec. 9. On Dec. 21, Chrétien returned in <i>Soyuz TM-6</i> with the long-duration crew, Titov and Manarov, who on this day completed 366 days on <i>Mir</i> . Volkov and Krikalev stayed in space with Polyakov until Apr. 27, 1989, when they returned in <i>TM-7</i> , leaving <i>Mir</i> crewless for the first time in 2 years. During its occupied period, the space station had been resupplied by a total of 17 <i>Progress</i> flights with about 40 tons of goods.
Dec. 2–6, 1988	<i>STS 27</i> Robert L. Gibson Guy S. Gardner Mike Mullane Jerry L. Ross William M. Shepherd	 70(?) rev 	4 days 9 h 6 min 19 s	Launched as second flight after the <i>Challenger</i> accident, <i>Atlantis</i> carried a radar imaging reconnaissance satellite on a classified Department of Defense mission into a high-inclination orbit. Returned to Edwards after highly successful mission.
March 13-18, 1989	STS 29 Michael L. Coats John E. Blaha James F. Buchli Robert C. Springer James P. Bagian	263,290 lb 79 rev 206 mi	4 days 23 h 39 min	<i>Discovery</i> , on this third flight after the <i>Challenger</i> disaster, de- ployed its main payload, the third <i>Tracking and Data Relay Satellite</i> , <i>TDRS-D</i> , completing the constellation of NASA's new communi- cations system. After conducting a number of physical, biological, medical, and materials experiment, <i>Discovery</i> returned to Edwards AFB.
May 4–8, 1989	STS 30 David M. Walker Ronald J. Grabe Norman E. Thagard Mary L. Cleave Mark C. Lee	217,513 lb 64 rev 185 mi	4 days 56 min 38 s	Originally scheduled for Apr. 28, the launch of <i>Atlantis</i> was aborted at T-31 due to failure of the hydrogen recirculation pump of a main engine and rescheduled for May 4. On that day, <i>Atlantis</i> lifted off and, $6^{1}/_{2}$ h later, placed the Venus radar mapper probe <i>Magellan</i> on its trajectory to the planet Venus, the first interplanetary probe launched by the United States in nearly 11 years. <i>Atlantis</i> landed at Edwards after a highly successful mission.
Aug. 8–13, 1989	STS 28 Brewster H. Shaw, Jr. Richard N. Richards James C. Adamson Mark N. Brown David C. Leestma	 80 rev 192 mi	5 days 1 h 56 s	Fifth flight after the <i>Challenger</i> accident. The extensively refurbished <i>Columbia</i> , on its first trip after over 3 years of inactivity, carried a classified Department of Defense payload into a 57° inclination orbit. Landing at Edwards took place after a successful mission.
Sept. 6, 1989–Feb. 19, 1990	Soyuz TM-8 Alexander Viktorenko Alexander Serebrov	14,400 lb Approx. 2662 rev 230 mi	166 days 7 h 58 min	After a pause of almost 5 months in its crewed program, the Soviet Union launched <i>TM-8</i> to <i>Mir</i> , the 67th crewed flight of the Soviet Union, carrying <i>Mir</i> 's 20th and 21st occupants. Docking on <i>Mir</i> occurred on Sept. 8, executed manually after a last-minute techni- cal problem with the automatic rendezvous/docking system of the <i>TM-8</i> . The arrival of the crew had been preceded by the docking of <i>Progress M1</i> , the first of a new series of crewless resupply craft, able to carry 5952 lb of cargo (882 lb more than the earlier version), and equipped with a recoverable entry capsule (not used on this flight).

1 lb = 0.45 kg. 1 mi = 1.6 km. 1 ft = 0.3 m. 1 yd = 0.9 m. 1 fathom = 1.8 m. 1 lb/in.² = 6.9 kPa. [†]After the *Challenger* accident, shuttle mission designations returned to consecutive numbering. Unavoidable changes in flight assignments, however, resulted in numbered flights not being flown in sequence.

Dates launched and recovered	Designation and crew	A. Weight B. Revolutions C. Max. distance from Earth	Duration	Remarks
Oct. 18–23, 1989	STS 34 Donald E. Williams Michael J. McCulley Shannon W. Lucid Franklin R. Chang-Diaz Ellen S. Baker	264,775 lb 79 rev 206 mi	4 days 23 h 40 min	Launch of <i>Atlantis</i> with the <i>Galileo</i> Jupiter probe as primary payload occurred after two launch postponements, one from Oct. 12 due to a main engine controller problem, the other from Oct. 11 due to inclement weather. At 7:15 PM, the 38,500-lb payload was successfully deployed; at 8:15 PM, the Inertial Upper Stage (IUS) first stage was fired, followed by second-stage ignition a 8:20. The 2-ton <i>Galileo</i> separated from the IUS at 9:05 with deployed booms, was spun up to 2.8 revolutions per min a 9:30 PM, and began its 6-year voyage to Jupiter, which includes one Venus and two Earth flybys for gravity assist. After conducting in-space experiments on material processing, plant life, ozon monitoring, ice crystals, and human physiology, <i>Atlantis</i> returnee to Earth two orbits early due to adverse wind at the landing site.
Nov. 22–27, 1989	STS 33 Frederick D. Gregory John E. Blaha Manley L. (Sonny) Carter Kathryn C. Thornton F. Story Musgrave	 78(?) rev	5 days 7 min	On the seventh shuttle flight after the <i>Challenger</i> accident, the <i>Discovery</i> lifted off in darkness, carrying a classified military payload into a low-inclination orbit. After a successful mission the vehicle landed at Edwards. It was the third night launch of the shuttle program.
Nov. 26, 1989	Kvant 2 Crewless	43,000 lb — 250 mi	_	<i>Kvant 2</i> , the second of four large building blocks of the Soviet <i>Mir</i> complex, was launched on a Proton. It docked to <i>Mir</i> on Dec. 6, after a lengthy delay caused by a multistep trouble-shooting program to deploy a stuck solar array and subsequent rendezvous problems. After the successful docking. <i>Kvant 2</i> , which carried systems and equipment to expand <i>Mir</i> 's operational capability and a crewed maneuvering unit for extravehicular activity, generated 3.5 kW of power from each of its two solar arrays.
Jan. 9–20, 1990	STS 32 Daniel C. Brandenstein James D. Wetherbee Bonnie J. Dunbar Marsha S. Ivins G. David Low	229,848 lb 172 rev 224 mi	10 days 21 h 52 s	Longest shuttle mission to date. Originally scheduled for Dec. 18, 1989, but delayed due to unfinished launch pad work and one weather-caused slip (on Jan. 8). Launched on its mission to deploy the Navy communications satellite <i>Syncom IV-F5</i> (on Jan. 10) and retrieve the Long Duration Exposure Facility (LDEF), a 12-sided open-grid aluminum satellite, 30 ft long, 14 ft in diameter, and 10.5 tons in mass, carrying 57 experiments. LDEF was originally orbited by <i>Challenger</i> (<i>STS 41C</i>) on Apr. 6, 1984, but its planned retrieva 1 year later was delayed by schedule slips and the <i>Challenger</i> accident. After <i>Columbia</i> 's rendezvous with LDEF on Jan. 12. Mission Specialist Dunbar used the Remote Manipulator System arm to grapple the research facility, 2093 days and 32,419 orbits after LDEF's launch. <i>Columbia</i> landed at Edwads, delayed 2 days by ground fog. Heaviest landing weight (by 5 tons) to date.
Feb. 11–Aug. 9, 1990	<i>Soyuz TM-9</i> Anatoliy Solovyev Alexander Balandin	14,400 lb 230 rev —	179 days	Docked with <i>Mir</i> on Feb. 13 to relieve Viktorenko and Serebrov Earlier in February, they had conducted several spacewalks (EVA) to test a crewed maneuvering unit named <i>Ikar</i> (Icarus) Serebrov on four flights on Feb. 1 during a 5-h EVA, Viktorenko several times on a 3-h 45-min EVA on Feb. 5. The cosmonauts also tested a modified space suit and portable life-support system. The Soviet crewed maneuvering unit weighed 440 lb and was propelled by 36 compressed-air thrusters. Solovyev and Balantin became the 22d and 23d occupants of <i>Mir</i> . Viktorenko and Serebrov returned to Earth on Feb. 19, landing in <i>Soyuz TM-8</i> near Arkalyk 205 mi northeast of Baikonur, in good health after 166 days in space, including five spacewalks and dozens of experiment programs.
Feb. 28–Mar. 4, 1990	STS 36 John O. Creighton John H. Casper Michael Mullane David C. Hilmers Pierre J. Thuot	71 rev	4 days 10 h 19 min	Launched as ninth flight since <i>Challenger</i> loss, after five delays due to head cold of the Commander, bad weather, and a ground computer glitch. <i>Atlantis</i> deployed a secret intelligence-gathering satellite on a Department of Defense mission in a highly inclined orbit (62°). Returned to Edwards AFB after completing a fully successfur mission.
Apr. 24–29, 1990	STS 31 Loren J. Shriver Charles F. Bolden Steven A. Hawley	259,229 lb 76 rev 381 mi	5 days 1 h 15 min	The long-awaited launch of the <i>Hubble Space Telescope</i> 20 years in planning and development, took place after an ear- lier launch attempt, on Apr. 10, was scrubbed due to malfunc- tion of a valve in one of <i>Discovery</i> 's auxiliary power units. The

TABLE 1. Crewed space flights 1961–2006 (cont.)					
Dates launched and recovered	Designation and crew	A. Weight B. Revolutions C. Max. distance from Earth	Duration	Remarks	
Apr. 24–29, 1990 (cont.)	Bruce McCandless Kathryn D. Sullivan			12.5-ton spacecraft, 42.5×14 ft, was placed into space by the manipulator arm on Apr. 25 at 8:45 a.m. EDT and released, after deployment of its solar arrays and a checkout, at 3:38 p.m. For the remainder of the mission, the crew carried out medical and engineering tests in the shuttle middeck. <i>Discovery</i> landed at Edwards AFB.	
May 31, 1990	Kristall Crewless	43,000 lb 250 rev 	_	As third major addition to <i>Mir</i> , the technology module <i>Kristall</i> docked to the Soviet space station on June 10. Nearly identical to the previous module, <i>Kvant 2</i> , with length of 45 ft, width of 14.3 ft, and volume of 2000 ft ³ , <i>Kristall</i> forms the second of four spokes of the <i>Mir</i> complex. Equipped with furnaces for materials experiments in space, it also has a docking port for the <i>Buran</i> space shuttle. Docking took two attempts, using the backup system to a failed maneuvering jet. On June 11, <i>Kristall</i> was moved by the Ljappa pivot arm from forward docking port to a lateral port opposite the <i>Kvant 2</i> module. With core, <i>Kvant 1</i> , <i>Kvant 2</i> , <i>Kristall</i> , and <i>Soyuz-TM</i> , <i>Mir</i> now weighed 183,000 lb. <i>Kristall</i> also carried equipment to repair loose insulation panels on <i>Soyuz TM-9</i> . Repair took place on July 17, but Solovyev and Balandin encountered difficulties in closing the <i>Kvant 2</i> outer airlock door, damaged on exit, and exceeded the life-support capability of their spacesuits. Crew used emergency entry method (depressurizing <i>Kvant 2</i>), conducted 3-h EVA to repair hatch on July 26, and returned to Earth on Aug. 9, one week after arrival of <i>TM-10</i> , landing at Arkalyk in Kazakhstan with 286 lb of material.	
Aug. 1–Dec. 10, 1990	<i>Soyuz TM-10</i> Gennaidy Manakov Gennaidy Strekalov	14,400 lb Approx. 2096 rev 235 mi	131 days	Launched to replace <i>TM-9</i> , it docked with <i>Mir</i> on Aug. 3 to continue work of previous crew (which had completed 506 of 520 planned science experiments), concentrating mostly on materials manufacturing with the <i>Kristall</i> module and switching solar panels from <i>Kristall</i> to <i>Kvant 1</i> . Crewless <i>Progress M-4</i> was launched Aug. 15 and docked Aug. 17, to bring supplies and boost <i>Mir</i> again up to 248 mi altitude. <i>Progress M-4</i> was used to check out a new feature, a returnable capsule for returning material to Earth from <i>Mir</i> , prior to its use on <i>Progress M-5</i> . Manakov and Strekalov returned on Dec. 10.	
Oct. 6–10, 1990	STS 41 Richard N. Richards Robert D. Cabana Bruce E. Meinick William M. Shepherd Thomas D. Akers	256,330 lb 65 rev 206 mi	4 days 2 h 10 min	Replacing the postponed <i>STS 35</i> (<i>Columbia</i>), <i>Discovery</i> was launched with the European-built out-of-the-ecliptic solar probe Ulysses plus additional scientific experiments on board. <i>Ulysses</i> was deployed during the fifth orbit and launched by its own IUS/PAM rocket system. The fastest artificial object, it started its 5-year mission via a gravity-assist slingshot maneuver around Jupiter in Feb. 1992. It passed out of the Sun in summer 1994 and spring 1995, respectively.	
Nov. 15–20, 1990	STS 38 Richard O. Covey Frank L. Culbertson Carl J. Meade Robert C. Springer Charles D. ("Sam") Gemar	— 79 rev 130 mi	4 days 21 h 54 min 32 s	On this fifth night launch, <i>Atlantis</i> , with a classified DOD payload, lifted off at 6:48 p.m. EST on it seventh flight. The planned landing at Edwards on Nov. 19 was canceled due to high winds and look place 1 day later at Kennedy at about 4:43 p.m. EST. It was the sixth Kennedy landing, $5^{1}/_{2}$ years after the last one (<i>Discovery</i> , on Apr. 5, 1985).	
Dec. 2–11, 1990	STS 35 Vance D. Brand Guy S. Gardner John M. Lounge Robert A. R. Parker Ronald A. Parise Jeffrey A. Hoffman Samuel T. Durrance	267,513 lb 143 rev 220 mi	8 days 23 h 6 min 5 s	The first shuttle mission in 7 years dedicated to a single science discipline: astrophysics. After 6 months delay due to a faulty cooling system valve, hydrogen leaks, and payload electronics repair, <i>Columbia</i> 's tenth flight took off with ASTRO 1 (two <i>Spacelab</i> pallets with three ultraviolet telescopes on a pointing system, and one x-ray telescope) to explore the hottest, albeit invisible parts of the universe from the Orbiter cargo bay. The flight featured more than 80 h of observations (about 394) of about 135 science targets in the universe, starting 23 h after orbit insertion and ending 4 h before deorbit burn. Also, an experiment of <i>Columbia</i> in voice and data mode (SAREX) was conducted by Parise. During orbit 39 on Dec. 4, <i>Columbia</i> and <i>Mir</i> were only 31 mi apart, and <i>Mir</i> was sighted twice from the shuttle. <i>STS</i> 35 landed at Edwards one day earlier than planned because of bad weather that was forecast for later, after a mission troubled by computer, pointing, and plumbing problems.	

Dates launched and recovered	Designation and crew	A. Weight B. Revolutions C. Max. distance from Earth	Duration	Remarks
Dec. 2–May 26, 1990	<i>Soyuz TM-11</i> Viktor Afanasyev Musa Manarov Toyohiro Akiyama	14,400 lb Approx. 2740 rev 230 mi	8 days/Akiyama; 131 days/ Manakov and Strekalov	Lifting off with the first Japanese in space on board, a journalist, Soyuz TM-11 docked at Mir 2 days later, making up a record num- ber of 12 humans in space at the same time (5 in Mir, 7 in STS 35). Cosmonauts Manakov, Strekalov, and Akiyama returned on Dec. 10, in TM-10 with Afanasyev and Manarov taking over on Mir, Resupply ship Progress M7, launched Mar. 19, failed twice to dock to Mir's Kvant module. On the second attempt on Mar. 23, ground controllers barely avoided a collision between Mir and M7. On March 26, the crew relocated Soyuz TM-11 from the main port to the Kvant 1 module and docked M7 successfully to that port on Mar. 28.
Apr. 5–11, 1991	STS 37 Steven Nagel Kenneth D. Cameron Jerry L. Ross Jerome Apt Linda M. Godwin	240,809 lb 92 rev 281 mi	5 days 23 h 32 min	Atlantis lifted off with the 35,000-lb Gamma Ray Observatory (GRO), the heaviest NASA science satellite ever carried by a shuttle. Equipped with four instruments to observe celestial high- energy sources with 10 times the sensitivity of prior experiments, GRO was set free on April 7 after an unplanned emergency EVA by Ross and Apt to deploy GRO's stuck high-gain antenna—the first United States spacewalk since Dec. 1985. On Apr. 8, Ross and Apt performed a scheduled 6-h EVA to test various crew equipment translation aids such as carts on a track, for use on the exterior of the space station.
Apr. 28–May 6, 1991	STS 39 Michael L. Coats L. Blaine Hammond, Jr. Gregory J. Harbaugh Donald R. McMonagle Guion S. Bluford C. Lacy Veach Richard J. Hieb	236,623 lb 133 rev 162 mi	8 days 7 h 22 min	Discovery made the first unclassified DOD-dedicated shut- tle mission. The mission was highlighted by around-the-clock observations of the atmosphere, gas release from subsatellites, shuttle engine firings, and the shuttle's orbital environment in wavelengths ranging from infared to the far-ultraviolet, including the aurora australis.
May 18-Oct. 10, 1991	<i>Soyuz TM-12</i> Anatoly Artsebarsky Sergei Krikalev Helen Sharman	14,400 lb Approx. 2265 rev 230 mi	7 days 19 h 5 min/ Sharman; 175 days 2 h 48 min/ Afanasyev and Manarov	With the first British space traveler, food technologist Helen Sharman, as part of Britain's commercial project Juno, <i>Soyuz</i> docked at <i>Mir</i> on May 20, the ninth crew to visit the station. Earlier, on Apr. 25, the <i>Mir</i> crew, during a 5-h EVA, inspected the automatic docking system antenna that had nearly caused the collision with the <i>Progress M7</i> automated transport on Mar. 23. <i>M7</i> was discarded to burn up in the atmosphere. Sharman, Afanasyov, and Manarov returned on May 26 in <i>TM-11</i> , landing in Kazakhstan. In the first 3 months of their stay, Artsebarsky and Krikalev made six spacewalks totalling a record 32 h, for such tasks as repairing a docking antenna and a television camera, deploying instruments, and assembling a 46-ft-tall towerlike structure called Sofora, to carry reaction jets, on <i>Mir</i> . On Aug. 16, <i>Progress M8</i> was separated from <i>Mir</i> , and on Aug. 20 new supplies were brought up on <i>M9</i> in a night launch. <i>Soyuz TM-12</i> returned on Oct. 10 with Artsebarsky, Aubakirov, and Viehboeck.
June 5–14, 1991	STS 40 Bryan D. O'Connor Sidney M. Gutierrez Tamara E. Jernigan M. Rhea Seddon James P. Bagian F. Drew Gaffney Millie Huges-Fulford	241,175 lb 145 rev 185 mi	9 days 2 h 15 min	Columbia carried Spacelab Life Sciences 1 (SLS 1), with three medical doctors, one biochemistry researcher, and one physi- cist on board, to perform 18 major experiments on the cardio- vascular, renal/endocrine, blood, immune, musculoskeletal, and neurovestibular systems. Test objects included white rats and jellyfish.
Aug. 2–11, 1991	STS 43 John E. Blaha Michael A. Baker Shannon W. Lucid G. David Low James C. Adamson	247,965 lb 140 rev 207 mi	8 days 21 h 21 min	The Atlantis flight put NASA's fourth Tracking and Data Relay Satellite (TDRS-E) into space, boosted successfully to a geosynchronous orbit by an Inertial Upper Stage (IUS). Other payloads included an atmospheric ultraviolet calibration instrument to measure the Earth's ozone layer, a heat pipe experiment, and an optical communications experiment. Atlantis landed at Kennedy.
Sept. 12–18, 1991	STS 48 John O. Creighton Kenneth S. Reightler Mark N. Brown James F. Buchli Charles D. Gemar	224,141 lb 81 rev 356 mi	5 days 8 h 27 min	Discovery carried the giant Upper Atmosphere Research Satellite (UARS) into a 57°-inclination orbit, the first major flight element of NASA's Mission to Planet Earth program. The 14,400-lb satellite carried nine complementary scientific instruments to provide a more complete understanding of the chemistry, dynamics, and energy input of the upper atmosphere, including the ozone layer, plus a radiometer measuring the Sun's energy output. Other on-board work included microgravity research on crystal growth, fluids and structures behavior. Ilfe sciences, and radiation effects.

Dates launched and recovered	Designation and crew	A. Weight B. Revolutions C. Max. distance from Earth	Duration	Remarks
Oct. 2–Mar. 25, 1991	<i>Soyuz TM-13</i> Alexander Volkov Toktar Aubakirov Franz Viehboeck	14,400 lb Approx. 2740 rev 230 mi	8 days/ Viehboeck and Aubakirov; 144 days 19 h/ Artsebarsky	<i>Soyuz</i> docked with <i>Mir</i> on Oct. 4. Cosmonauts Artsebarsky Viehboeck (first space traveler from Austria), and Aubakirov (firs Kazakh crew member) returned on Oct. 10 in <i>TM-12</i> , leavin Volkov and Krikalev in control of the space station.
Nov. 24–Dec. 1, 1991	STS 44 Frederick D. Gregory Terence T. Henricks Story Musgrave Mario Runco, Jr. James S. Voss Thomas J. Hennen	247,087 lb 108 rev 195 mi	6 days 22 h 51 min	Atlantis, on a DOD-dedicated mission, carried a Defense Suppor Program (DSP) satellite for detection, from geosynchronou orbit, of nuclear detonations, and missile and space launches After its deployment on the first day of the mission, the crew worked with a variety of secondary payloads, including naked eye Earth observations, radiation measurements, and medical in vestigations. Due to malfunction of one of three Inertial Measure ment Units (IMUs), the flight had to be shortened by 3 days.
Jan. 22–30, 1992	STS 42 Ronald J. Grabe Stephen S. Oswald Norman E. Thagard William F. Readdy David C. Hilmers Roberta L. Bondar Ulf D. Merbold	231,497 lb 128 rev 188 mi	8 days 1 h 16 min	Discovery carried the International Microgravity Laboratory (IML 1), dedicated to worldwide research in the behavior of materials and life in weightlessness. 225 scientists from 1 countries, experimenting in 42 research disciplines, cooperated i this fight. Along with IML 1, the payload consisted of 12 Get Awa Special (GAS) containers, the large-format IMAX camera, and tw student experiments. The crew included foreign astronauts Bonda (Canada) and Merbold (Germany).
Mar. 17–Aug. 10, 1992	<i>Soyuz TM-14</i> Alexander Viktorenko Alexander Kaleri Klaus-Dietrich Flade	14,400 lb Approx. 2280 rev 245 mi	312 days 6 h 1 min/Krikalev; 175 days 3 h 52 min/Volkov; 7 days 21 h 57 min/Flade	The crew, Including the German guest cosmonaut Flade, docker with <i>Mir</i> on Mar. 19. Flade conducted a crowded science prograr and returned with former <i>Mir</i> occupants Volkov and Krikale on Mar. 25, to land in <i>Soyuz TM-13</i> in Kazakhstan, leaviny Viktorenko and Kaleri in <i>Mir</i> . In space since May 18, 1991 Krikalev was originally slated to return in late 1991, but th schedule was revised, leaving him in orbit for five extra months On Feb. 20–21 he conducted his seventh spacewalk, fetching experiments left outside during previous EVAs, and setting a new record.
Mar. 24–Apr. 2, 1992	STS 45 Charles Bolden Brian Duffy C. Michael Foale David Leestma Kathryn Sullivan Dirk Frimout Byron Lichtenberg	222,086 lb 142 rev 182 mi	8 days 22 h 9 min 20 s	The multinational flight of <i>Atlantis</i> , including a Belgian payload specialist (Frimout), carried the first ATLAS (Atmospheric Laborator for Applications and Science) associated with NASA's Mission t Planet Earth initiative. The ATLAS payload of 14 instruments re mained attached to the Orbiter, most of them mounted on tw. <i>Spacelab</i> pallets. A Far Ultraviolet Space Telescope (FAUST) wa also employed.
May 7–16, 1992	STS 49 Daniel C. Brandenstein Kevin P. Chilton Richard J. Hieb Bruce E. Melnick Pierre J. Thuout Kathryn C. Thornton Thomas D. Akers	246,008 lb 140 rev 181 mi	8 days 21 h 18 min 36 s	This was the first mission featuring four EVAs, and the maide voyage of the orbiter <i>Endeavour</i> . The main mission was reboost ing the 11.7-ft-diameter, 17.5-ft-high communications satellit <i>INTELSAT VI</i> (F-3) launched 2 years earlier by a Titan rocket but stranded in a useless orbit when the Titan's second stag failed to separate. After rendezvous with the satellite, Thuot an Hieb, on two spacewalks on flight days 4 and 5, failed to cap ture the 8960-lb satellite. Joined by Akers on May 13, howeve they succeeded in grasping it with their hands, berthing it in th shuttle cargo bay with the remote manipulator arm, attaching to a new perigee kick motor, and redeploying it. This first three person EVA in history lasted 8 h 36 min, the longest spacewal ever conducted. The solid-fuel motor was fired remotely from the ground on May 14, taking <i>INTELSAT VI</i> to its final destination in a geosynchronous orbit. In a fourth EVA on May 14 Thornton and Akers evaluated space-station assembly equipmer and techniques as well as several prototype crew self-rescuidevices.
June 25–July 9, 1992	STS 50 Richard N. Richards Kenneth D. Bowersox Bonnie J. Dunbar Carl J. Meade Ellen S. Baker Lawence J. DeLucas Eugene H. Trinh	245,902 lb 220 rev 185 mi	13 days 19 h 30 min 30 s	The first extended-duration (EDO) shuttle flight. <i>Columbia</i> , specially modified for 2-week missions, carried the first U.S. Microgravity Laboratory 1 (USML 1), a <i>Spacelab</i> modul equipped with 31 experiments for advanced research in zerg. The longest shuttle flight to date, it supported investigation of the effects of weightlessness on plants, humans, and materials. The mission also carried investigations into polymer membrane manufacturing, the Space Shuttle Amateur Radio Experiment II (SAREX-II), and numerous other cargo bay, middeck, an secondary payloads.

Dates launched and recovered	Designation and crew	A. Weight B. Revolutions C. Max. distance from Earth	Duration	Remarks
July 27–Feb. 1, 1992	Soyuz TM-15 Anatoliy Solovyev Sergei Avdeyev Michel Tognini	14,400 lb 2946 rev 245 mi	146 days/Kaleri and Viktorenko; 14/ Tognini	The third Russian-French mission since 1982, <i>Soyuz</i> carried Frence researcher Tognini, to conduct 10 biomedical, fluid and material science and technological experiments on board <i>Mir</i> . Tognin returned on Aug. 10 in <i>Soyuz TM-14</i> along with Viktorenko an Kaleri. At touchdown in Kazakhstan, the spherical <i>TM</i> 14 rolle over, briefly trapping but not injuring the crew. Solovyev an Avdeyev remained in <i>Mir</i> with numerous tasks such as installin new gyrodynes and rocket thrusters for attitude control. <i>Progress</i> <i>M-13</i> was launched on Aug. 14, carrying supplies and medica gear to <i>Mir</i> .
July 31–Aug. 8, 1992	STS 46 Loren J. Shriver Andrew M. Allen Jeffrey A. Hoffman Franklin R. Chang-Diaz Marsha S. Ivins Claude Nicollier Franco Malerba	241,797 lb 126 rev 266 mi	7 days 23 h 14 min 12 s	Atlantis successfully deployed (on Aug. 2) the first Europea Retrievable Carrier (EURECA) platform, carrying seeds, spores crystals, and other experimental materials for research in zero g to be retrieved about 9 months later. Due to data errors, EURECA' orbital self-transfer to its final altitude of 313 mi occurred later that planned but was accomplished on Aug. 7, Deployment of the mai payload, the 1142-Ib Italian Tethered Satellite System (TSS 1) o a 12.5-mi-long cable from the <i>Atlantis</i> , however, was terminate when a breakdown in the reel mechanism halted the tether afte only 850 ft. The crew included two foreign astronauts, Malerba (Italy) and Nicollier (Switzerland) for ESA.
Sept. 12–20, 1992	STS 47 Robert L. Gibson Curtis L. Brown, Jr. Mark C. Lee Jay Apt N. Jan Davis May C. Jemison Mamoru Mohri	232,661 lb 126 rev 192 mi	7 days 22 h 30 min	Endeavour carried the Japan-chartered Spacelab SL-J with 4 materials and life sciences experiments and nine Get Away Specia (GAS) payload canisters. The cargo also included the Israeli Spac Agency Investigation About Hornets (ISAIAH) experiment an another flight of the Shuttle Amateur Radio Experiment (SAREX, The crew included the first Japanese shuttle astronaut (Mohri).
Oct. 22–Nov. 1, 1992	STS 52 James D. Wetherbee Michael A. Baker C. Lacy Veach William M. Shepherd Tamara E. Jernigan Steven G. MacLean	239,178 lb 158 rev 189 mi	9 days 20 h 56 min	Columbia carried the U.S. Microgravity Payload (USMP 1 as main cargo. The crew, with the third Canadian astronau (MacLean), performed materials and science experiments in micro g with the U.S. Microgravity Payload 2 (USMP 2), deployed th second Italian/U.S. <i>Laser Geodynamic Satellite (LAGEOS 2)</i> , and conducted other physical experiments including the Canadia Experiment #2 (CANEX 2), and space-station-related investiga- tions.
Dec. 2–9, 1992	STS 53 David M. Walker Robert D. Cabana Guion S. Bluford James S. Voss Michael Richard Clifford	230,731 lb 116 rev 237 mi	7 days 7 h 20 min	As the last dedicated DOD mission, <i>Discovery</i> carried the satellit DOD 1 as primary and two major experiments, the Glow Cryogenic Heat Pipe Experiment (GCP) and Orbital Debris Rada Calibration Spheres (DDERACS), as secondary payloads. Wit GCP, remotely commanded special sensors observed orbite surfaces, contamination events, and airglow, and measured th zero-g performance of liquid oxygen heat pipes. ODERACS had t be canceled when its container in the cargo bay failed to open.
Jan. 13–19, 1993	STS 54 John H. Casper Donald R. McMonagle Mario Runco, Jr. Susan J. Helms Gregory J. Harbaugh	248,338 lb 188 rev 95 mi	5 days 23 h 39 min	Endeavour carried the sixth <i>Tracking and Data Relay Satelliti</i> (<i>TDRS</i>), its Inertial Upper Stage (IUS), and the Diffuse X-Ra Spectrometer (DXS), consisting of two instruments attached to the sides of the cargo bay. After successful deployment of <i>TDRS F</i> , crew activities included several microgravity experiments or materials and rodents, and the first in a series of test spacewalk leading up to space-station construction. Harbaugh and Runcc assisted by Helms from the cabin, tested EVA training procedure for the space station and the <i>Hubble</i> repair mission (STS 67) Using children's toys, the crew responded to students from various elementary schools asking questions about physics in zero <i>g</i> .
Jan. 24–July 22, 1993	<i>Soyuz TM-16</i> Gennadiy Manakov Alexander Polishchuk	14,400 lb Approx. 2790 rev 242 mi	189 days/ Solovyev and Avdeyev	Equipped with the new androgynous docking system APAS-89, the tested for use in the newly planned <i>Mir/space</i> shuttle docking mission, <i>Soyuz</i> docked with <i>Mir</i> , for the first time, at the <i>Krista</i> module's docking port. The previous crew, Solovyev and Avdeyev returned on Feb. 1 in <i>Soyuz TM-15</i> after 189 days in space. On Feb. 4, <i>Progress M-15</i> was undocked from <i>Mir's</i> aft port and used the deploy, by centrifugal force, a 66-ft-diameter solar sail made from superthin aluminum foil. The purpose of the test, which reflected a visible light beam to Earth, was to assess the force of photon from the Sun for propulsive application.

TABLE 1. Crewed	TABLE 1. Crewed space flights 1961-2006 (cont.)						
Dates launched and recovered	Designation and crew	A. Weight B. Revolutions C. Max. distance from Earth	Duration	Remarks			
Apr. 8–17, 1993	STS 56 Kenneth D. Cameron Stephen S. Oswald Michael Foale Kenneth D. Cockrell Ellen Ochoa	225,597 lb 147 rev 185 mi	9 days 6 h 8 min	Discovery carried ATLAS 2, the second in NASA's series of Atmospheric Laboratory for Applications and Science missions. The remote sensing instruments of ATLAS studied the Sun's energy output and Earth's middle-atmosphere chemical makeup, particularly with regard to ozone depletion. The crew on Apr. 11 deployed Spartan-201, a free-flying platform carrying instruments to study the Sun's corona and the solar wind, and recovered the platform on Apr. 13.			
Apr. 26–May 6, 1993	STS 55 Steven R. Nagel Terence T. Henricks Jerry L. Ross Charles J. Precourt Bernard A. Harris Ulrich Walter Hans Wihelm Schlegel	244,156 lb 158 rev 187 mi	9 days 23 h 40 min	Columbia carried the Spacelab module on its eleventh flight, chartered for the second time by the German space agency DARA, with two German payload specialists (Walter, Schlegel) among its crew. The Spacelab D2 mission was devoted to zero- g-relevant investigations in a wide range of science disciplines, including materials research; life sciences; technology experiments in optics, robotics, automation, and medical applications; Earth observations; atmospheric physics; and astronomy, for a total of about 90 different projects. Many investigations were aimed at developing, testing, and verifying future operations/utilization in permanent space-station laboratories.			
June 21–July 1, 1993	STS 57 Ronald Grabe Brian Duffy David Low Nancy Sherlock Peter (Jeff) J. K. Wisoff Janice Voss	239,319 lb 158 rev 289 mi	9 days 23 h 46 min	Highlights of the <i>Endeavour's</i> flight were the first flight of the privately developed middeck module <i>Spacehab</i> and the retrieval of the European space platform EURECA, in orbit since its deployment from <i>STS</i> 46 on Aug. 1, 1992. <i>Spacehab</i> experiments included 13 commercial payloads, ranging from drug improvement to feeding plants, cell splitting, the first soldering experiments inspace by United States astronauts, and high-temperature melting of metals; and six NASA experiments involving biotechnology, human factors, and water recycling for the future space station. Other payloads included SHOOT (Superfluid Helium On-Orbit Transfer) to demonstrate the resupply of liquid helium containers in space. A 5½-h EVA by Low and Wisoff on June 25 served to latch two EURECA antennas and to refine procedures for the planned <i>Hubble</i> repair mission.			
July 1–Jan. 14, 1993	Soyuz TM-17 Vasili Tsibliyev Alexander Serebrov Jean-Pierre Haignere	14,400 lb Approx. 2790 rev 242 mi	179 days/ Polishchuk and Manakov; 21 days/ Haignere	Carrying the fourth French cosmonaut since 1982, <i>Soyuz</i> docked with <i>Mir</i> on July 3. Resupply craft <i>Progress M-18</i> , docked on May 24, was undocked on July 4 and reentered destructively, after separating a small ballistic Raduga ("Rainbow") reentry capsule carrying experiments and recordings for a soft landing in Russia. During his stay at the station, Haignere conducted 11 major Altair experiments, then returned to Earth on July 22 with the previous crew, Polishchuk and Manakov, in <i>TM-16</i> .			
Sept. 12–22, 1993	STS 51 Frank Culbertson William Readdy Jim Newman Dan Bursch Carl Walz	250,023 lb 157 rev 296 mi	9 days 20 h 11 min	Discovery carried two payloads. An Advanced Communication Technology Satellite (ACTS), a testbed for technology leading to a new generation of comsats, was deployed and injected by a Transfer Orbit Stage (TOS) toward its geosynchronous location. The German Orbiting and Retrievable Far and Extreme Ultraviolet Spectrometer–Shuttle Pallet Satellite (ORFEUS–SPAS) was set free by the robot arm and was retrieved 6 days later, the first of a series of such missions. Other mission tasks included a 7-h EVA by Walz and Newman on Sept. 16 as part of planning for the Hubble repair mission.			
Oct. 18–Nov. 1, 1993	STS 58 John E. Blaha Richard A. Searfoss M. Rhea Seddon William S. McArthur, Jr. David A. Wolf Shannon W. Lucid Martin J. Fettman	244,860 lb 224 rev 179 mi	14 days 13 min 40 s	<i>Columbia</i> carried SLS-2, the second <i>Spacelab</i> dedicated to life sciences research. NASA's longest shuttle mission to date conducted, in a 39° inclination orbit, a series of experiments to gain more knowledge on how the human body adapts to the weightless environment of space. The 14 experiments were performed on the crew, which included two medical doctors (Seddon, Wolf), a physicist/biochemist (Lucid), and a veterinarian (Fettman), and on laboratory animals (48 rats).			
Dec. 2-13, 1993	STS 61 Richard O. Covey Kenneth D. Bowersox Tom Akers Jeffrey A. Hoffman F. Story Musgrave Claude Nicollier Kathryn Thornton	236,670 lb 162 rev 370 mi	10 days 19 h 58 min 35 s	Endeavour made the first in a series of planned visits to the orbiting Hubble Space Telescope (HST). The crew included four astronauts with considerable EVA experience (Hoffman, Musgrave, Thornton, Akers) and Swiss astronaut Nicollier (ESA). On Dec. 4, Nicollier grappled the HST with the shuttle's robot arm, berthing it in the cargo bay. In five space walks during the following 5 days, the as- tronauts installed the Wide Field/Planetary Camera II and the Corrective Optics Space Telescope Axial Replacement (COSTAR)			
1 lb = 0.45 kg.	1 mi = 1.6 km. 1 ft = 0.3 m.	1 yd = 0.9 m. 1 f	athom = 1.8 m. 1 lb/in. ²	= 6.9 kPa.			

TABLE 1. Crewed	space flights 1961-2006 (c	ont.)		
Dates launched and recovered	Designation and crew	A. Weight B. Revolutions C. Max. distance from Earth	Duration	Remarks
Dec. 2–13, 1993 (cont.)				mechanism, both designed to compensate for the erroneous spherical aberration of the primary mirror. They also replaced the spacecraft's solar arrays and installed new gyroscopes, additional computer memory, and other systems to increase <i>HST</i> reliability. Total EVA time accumulated by all crews: 35 h 28 min. <i>HST</i> was set free on Dec. 10.
Jan. 8–July 9, 1994	Soyuz TM-18 Viktor Afanasyev Yuri Usachov Valery Polyakov	14,400 lb 3072 rev 242 mi	197 days 3 h 2 min/ Tsibliyev and Serebrov; 437 days 17 h 59 min/ Polyakov	This was the fifteenth mission to the <i>Mir</i> station. Docking occurred on Jan. 10. Polyakov, a physician cosmonaut who had spent 241 days in space in 1988 (<i>TM</i> -6), intended to conduct the longest human space flight in history to date to test the effects of long periods of weightlessness on the human body. He returned on <i>TM-20</i> on Mar. 22, 1995, after 14.6 months in space. During the mission, three <i>Progress</i> supply ships were to be sent to <i>Mir</i> , which was repaired and maintained by the crews. The previous crew, Vasily Tsibliyev and Alexander Serebrov, returned to Earth on Jan. 14 after 6 months, despite a last-minute slight collision of their <i>Soyuz TM-17</i> with <i>Mir</i> .
Feb. 3–11, 1994	STS 60 Charles F. Bolden Kenneth S. Reightler, Jr. N. Jan Davis Ronald M. Sega Franklin R. Chang-Diaz Sergei K. Krikalev	233,290 lb 130 rev 220 mi	8 days 7 h 9 min	This <i>Discovery</i> mission included the first Russian cosmonaut (Krikalev). The mission also carried the commercial <i>Spacehab</i> facility (on its second flight) with 12 biotechnology and materials processing payloads, and the free-flying Wake Shield Facility, a 12- ft-diameter disk designed to generate an ultravacuum environment in space in its wake, within which to grow thin semiconductor crystal films of gallium arsenide. WSF 1 deployment, however, did not take place due to problems with a horizon sensor on the satellite; crystals were grown with WSF attached to the robot arm. The mission also included deployment of the German satellite <i>BREMSAT</i> .
Mar. 4–18, 1994	STS 62 John H. Casper Andrew M. Allen Charles D. Gemar Marsha S. Ivins Pierre J. Thuot	245,457 lb 223 rev 188 mi	13 days 23 h 17.5 min	This <i>Columbia</i> mission was dedicated to microgravity research with the U.S. Microgravity Payload 2 (USMP 2) and the Office of Aeronautics and Space Technology 2 (OAST 2) payloads. The four experiments on USMP 2 were designed to study solidification techniques, crystal growth, and gas behavior in zero g. OAST 2 carried five experiments for various other effects of materials and environmental phenomena in space. The Orbiter also carried special acceleration measurement systems, and the Shuttle Solar Backscatter Ultraviolet (SSBUV) instrument.
Apr. 9–20, 1994	STS 59 Sidney M. Gutierrez Kevin P. Chilton Jerome Apt Michael R. Clifford Linda M. Godwin Thomas D. Jones	237,048 lb 181 rev 139 mi	11 days 5 h 50 min	Endeavour, in a 57° orbit, carried the first Space Radar Laboratory (SRL 1) for internationally conducted studies of the Earth's global environment. SRL consisted of the Spaceborne Imaging Radar-C (SIR-C) and the X-Band Synthetic Aperture Radar (X-SAR). Other payloads included microgravity experiments (CONCAP-IV) and medical investigations.
July 1–Nov. 4, 1994	<i>Soyuz TM-19</i> Yuri Malenchenko Talgat Musabayev	14,400 lb 1960 rev 242 mi	125 days 20 h 53 min; 182 days/ Afanasyev and Usachov	This mission docked with <i>Mir</i> on July 3, carrying a Kazakhstan cosmonaut (Musabayev) besides a Russian. The two crew members replaced Afanasyev and Usachev, who returned to Earth in <i>Soyuz TM-18</i> on July 9, after 6 months in space. Valery Polyakov continued his orbital stay. Malenchenko and Musabayev returned in <i>TM-19</i> on Nov. 4, 1994 (with Ulf Merbold of <i>TM-20</i>).
July 8–23, 1994	STS 65 Robert D. Cabana James Donald Halsell Richard J. Hieb Carl E. Walz Leroy Chiao Donald A. Thomas Chiaki Naito-Mukai	228,640 lb (landing) 235 rev 188 mi	14 days 17 h 55 min	Columbia carried the second mission of the International Micro- gravity Laboratory (IML 2) with furnaces and other facilities to study the behavior of materials and life in the weightless environment, and to produce a variety of material structures, from crystals to metal aloys. More than 200 scientists contributed 82 investigations, of which 81 were successfully completed in 19 IML 2 facilities, plus other middeck experiments. The crew included the first Japanese woman in space (Naito-Mukai).
Sept. 9–20, 1994	STS 64 Richard N. Richards L. Blaine Hammond Jerry Linenger Susan J. Helms Carl J. Meade Mark C. Lee	210,916 lb (landing) 176 rev 162 mi	10 days 22 h 50 min	Discovery was launched for a mission of atmospheric research, robotic atmospheric research, robotic operations, and an EVA. LITE (Lidar In-Space Technology Experiment) was flown and used in space. Operating from the shuttle payload bay, the neodymium YAG laser, an optical radar, probed the Earth's atmosphere to collect measurements on clouds and airborne dust. Other payloads were the Robot-Operated Processing System

 $1 \ \text{lb} = 0.45 \ \text{kg}. \quad 1 \ \text{ml} = 1.6 \ \text{km}. \quad 1 \ \text{ft} = 0.3 \ \text{m}. \quad 1 \ \text{yd} = 0.9 \ \text{m}. \quad 1 \ \text{fathom} = 1.8 \ \text{m}. \quad 1 \ \text{lb}/\text{in}.^2 = 6.8 \ \text{kPa}.$

Dates launched and recovered	Designation and crew	A. Weight B. Revolutions C. Max. distance from Earth	e Duration	Remarks
Sept. 9–20, 1994 (cont.)				(ROMPS), a Simplified Aid for EVA Rescue (SAFER) backpack used experimentally, and the astronomy satellite Spartan 201, deployed and retrieved after 40 h of free flight.
Sept. 30-Oct. 11, 1994	STS 68 Michael A. Baker Terrence W. Wilcutt Thomas D. Jones Steven L. Smith Daniel W. Bursch Peter (Jeff) J. K. Wisoff	223,040 lb (landing) 182 rev 139 mi	11 days 5 h 46 min	Endeavour carried the Space Radar Laboratory into space for the second time. An international team used its instruments to study how the Earth's global environment is changing.
Oct. 4–Mar. 22, 1994	<i>Soyuz TM-20</i> Alexander Viktorenko Elena Kondakova Ulf Merbold	14,400 lb 2637 rev 247 mi	31 days 11 h 36 min/Merbold; 169 days 5 h 22 min/ Viktorenko and Kondakova	This mission docked at <i>Mir</i> on Oct. 6, its crew including the third Russian woman (Kondakova) since inception of the Soviet/Russian space program, and ESA-astronaut Merbold on his third space trip. <i>TM-20</i> was the first of two ESA Euro-Mir missions, and its launch had depended on successful docking of <i>Progress</i> <i>M 24</i> , accomplished by <i>Mir</i> commander Malenchenko on Sept. 2 manually. Malenchenko and Musabayev then performed two EVAs, on Sept. 9 and 13, to inspect the (undamaged) orbital station where it had been struck by <i>M 24</i> on Aug. 30 and by <i>Soyuz TM-17</i> after its undocking in January, as well as <i>Mir's</i> Safora control mast. On Oct. 11, the crew had to cope with a serious power and attitude control failure of the station's core module. After performing experiments in life sciences, materials processing, and technology development, though limited by a failure of the CSK 1 materials furnace, Merbold, with Malenchenko and Musabayev, returned to <i>TM-19</i> on Nov. 14, leaving Kondakova and Viktorenko in space with Polyakov (to return on Mar. 22, 1995).
Nov. 3–14, 1994	STS 66 Donald R. McMonagle Curtis L. Brown Ellen Ochoa Joseph R. Tanner Jean-François Clervoy Scott E. Parazynski	209,857 lb (landing) 174 rev 193 mi	10 days 22 h 35 min	Atlantis carried the Atmospheric Laboratory for Applications and Science (ATLAS) on its third flight for Mission to Planet Earth. The crew, with a French astronaut (Clervoy) for ESA, used the remote- sensing laboratory for studying the Sun's energy output, the middle atmosphere's chemical makeup, and the effects of these factors on global ozone levels. On Nov. 4, they deployed the German- built reusable satellite <i>CRISTA-SPAS</i> , carrying cryogenic infrarec spectrometers, telescopes, and a high-resolution spectrograph fo atmosphere research. They recovered it on Nov. 12, using a new rendezvous method in preparation for later rendezvous/docking flights to the Russian <i>Mir</i> station.
Feb. 3–11, 1995	STS 63 James D. Wetherbee Eileen M. Collins Bernard A. Harris C. Michael Foale Janice Voss Vladimir G. Titov	211,318 lb (landing) 129 rev 250 mi	8 days 6 h 28 min 15 s	<i>Discovery</i> lifted off with the second Russian cosmonaut (Titov and the first woman shuttle pilot (Collins) within a narrow launch window of 5 min, as required by its mission: rendezvous and fly around of the Russian space station <i>Mir</i> at 51.6° inclination. The rendezvous, a dress rehearsal of docking missions to follow, took place on Feb. 6, validating new flight techniques required for those missions, though no docking was performed as yet. Also on board <i>Spacehab</i> on its third flight and <i>Spartan 204</i> , a satellite carrying various science experiments. It was deployed by Titov on Feb 7 and retrieved with the robot arm on Feb. 9. Foale and Harris performed a 4.5-h EVA on Feb. 9.
Mar. 2–18, 1995	STS 67 Stephen S Oswald William G. Gregory Tamara E. Jernigan John M. Grunsfeld Wendy B. Lawrence Samuel T. Durrance Ronald A. Parise	217,683 lb (landing) 262 rev 220 mi	16 days 15 h 9 min	Setting a shuttle duration record of about $16^{1}/_{2}$ days in orbit, <i>Endeavour</i> carried the second flight of the ASTRO observation platform with three unique telescopes for ultraviolet astronomy. Its tasks included mapping areas of space still largely uncharted in ultraviolet and measuring intergalactic gas abundance. Ultraviolet spectra were taken of about 300 celestial objects, including Jupiter and the Moon. Other experiments concerned large-structure control dynamics and materials processing.
Mar. 14–Sept. 11, 1995	<i>Soyuz TM-21</i> Vladimir Dezhurov Gennady Strekalov Norman Thagard	14,400 lb 2820 rev 246 mi	115 days 9 h 44 min	This eighteenth crew visit to <i>Mir</i> carried the first American (Thagard) on a Russian spacecraft. Rendezvous and docking took place on Mar. 16. With the current <i>Mir</i> crew of 3 cosmonauts and the 7 astronauts on <i>STS 67/Endeavour</i> , the number of humans simultaneously in space was 13, a record (until the <i>STS 67</i> landing on Mar. 18). <i>Mir 17</i> crew Viktorenko, Kondakova, and Polyakov returned on Mar. 22 in <i>TM</i> -20, with a new long-duration record of 437 days 17 h 59 min set by Polyakov, and longest

TABLE 1. Crewed	space flights 1961–2006 (con	t.)		
Dates launched and recovered	Designation and crew	A. Weight B. Revolutions C. Max. distance from Earth	Duration	Remarks
Mar. 14–Sept. 11, 1995 (cont.)				space flight by a woman (Kondakova) of 169 days 5 h 22 min. After a 3-month stay, Thagard, Dezhurov, and Strekalov were to be returned to Earth by the shuttle <i>Atlantis</i> (<i>STS 71</i>). Known as <i>Mir 18</i> , Thagard's mission was part of Phase One of joint activities leading to the <i>International Space Station</i> . At the end of his mission, Thagard had established a new United States record for time spent in space of 115 days 9 h 44 min.
May 20, 1995	Spektr Crewless	43,000 lb — 257 mi	_	Russia's new four-solar-array research module, launched on a heavy-lift Proton, carried extensive atmospheric science instru- mentation and about 1760 lb of United States equipment for car- diovascular and other biomedical research by astronaut Thagard, later by his replacement, Bonnie Dunbar. <i>Spektr</i> docked automat- ically to <i>Mir</i> on June 1, requiring five spacewalks by Dezhurov and Strekalov between May 12 and June 2 and the repositioning of the older module, <i>Kristall</i> . This prepared the station for the arrival and docking of the shuttle <i>Atlantis</i> (<i>STS 71</i>).
June 27–July 7, 1995	STS 71 Robert L. (Hoot) Gibson Charles J. Percourt Ellen S. Baker Gregory J. Harbaugh Bonnie Dunbar Anatoly Y. Solovyev (up) Nikolai M. Budarin (up) Vladimir Dezhurov (down) Gennady Strekalov (down) Norman Thagard (down)	214,700 lb (landing) 153 rev 246 mi	9 days 19 h 22 min 17 s; 75 days 11 h 19 min/ Solovyev and Budarin	Atlantis opened Phase One of the International Space Station program as the first of a series of shuttle– <i>Mir</i> linkups, 20 years after the joint United States–Russian ASTP mission. Docking took place on June 29, forming the largest human-made structure in space. Joint scientific (mostly medical) investigations were carried out inside the Spacelab module in the Atlantis cargo bay. Solovyev and Budarin then took over as new occupants of <i>Mir</i> (<i>Mir 19</i>), while the <i>Mir</i> 18 crew of Dezhurov, Strekalov, and Thagard returned to Earth with Atlantis. Undocking was on July 4, 4 days 21 h 35 min after linkup.
July 13–22, 1995	STS 70 Terrence (Tom) Henricks Kevin Kregel Donald Thomas Nancy Currie Mary Ellen Weber	195,195 lb (landing) 142 rev 206 mi	8 days 22 h 21 min 2 s	Discovery carried the seventh <i>Tracking and Data Relay Satellite</i> (<i>TDRS 7</i>). It was released out of the payload bay and boosted into geosynchronous orbit with its Inertial Upper Stage (IUS). Other payloads included a Commercial Protein Crystal Growth (CPCG) facility, a Bioreactor Demonstrator, and various biological experiments.
Sept. 3–Feb. 29, 1995	<i>Soyuz TM-22</i> Yuri Gidzenko Sergei Avdeyev Thomas Reiter	14,400 lb 2789 rev 246 mi	179 days	EUROMIR'95 docked to <i>Mir</i> on Sept. 5, after <i>Progress M-28</i> was undocked and jettisoned. ESA astronaul Reiter (Germany) served as flight engineer, marking the first time for Europe in Russia's space program. On Oct. 20, he (with Avdeyev) installed a European experiment on the outside of <i>Spektr</i> . Two cassettes exposed to natural and human-made dust and debris were returned by Reiter and Gidzenko during another EVA (the third) on Feb. 8, 1996. The <i>Mir</i> 19 crew (Solovyev and Budarin) returned in <i>TM-21</i> on Sept. 11. On-board activities included experiments in life sciences and micro-g materials processing. Most of ESA's equipment had been brought to <i>Mir</i> by <i>Progress M-28</i> in July and by <i>Spektr</i> . Another load arrived on <i>Progress M-29</i> (launch Oct. 8, docking Oct. 10, at aft end of <i>Kvant 1</i>), including the Munich Space Chair, set up Oct. 11 in <i>Spektr</i> . In November, the crew hosted the visiting shuttle <i>Atlantis/STS 74</i> . The crew returned 6 days after arrival of <i>TM-23</i> with the <i>Mir 21</i> crew.
Sept. 7–18, 1995	STS 69 David M. Walker Kenneth D. Cockrell James S. Voss Jim Newman Michael L. Gernhardt	21,971 lb (landing) 170 rev 232 mi	10 days 20 h 29 min	Endeavour carried the Spartan 210 free-flyer on its third flight and the Wake Shield Facility (WSF 2) on its second flight as main payloads, Spartan released on Sept. 8 and retrieved 2 days later, observed the solar wind and the Sun's outer atmosphere in conjunction with the <i>Ulysses</i> probe while it passed over the Sun's north polar region. The WSF, deployed by the robot arm on Sept. 11 and retrieved on Sept. 14, was partially successful in growing thin semiconductor films in the ultravacuum created in its wake. Other payloads included the International Extreme Ultraviolet Hitchhiker (IEH 1) and the Capillary Pumped Loop 2/Gas Bridge Assembly (CAPL 2/GBA) for microgravity studies of fluid systems. Voss and Gernhardt on an EVA tested construction tools, an arm-sleeve strap-on computer, and suit modifications for space- station assembly.

 $1 \ \text{lb} = 0.45 \ \text{kg}. \qquad 1 \ \text{mi} = 1.6 \ \text{km}. \qquad 1 \ \text{ft} = 0.3 \ \text{m}. \qquad 1 \ \text{yd} = 0.9 \ \text{m}. \qquad 1 \ \text{fathom} = 1.8 \ \text{m}. \qquad 1 \ \text{lb}/\text{in}.^2 = 6.9 \ \text{kPa}.$

Dates launched and recovered	Designation and crew	A. Weight B. Revolutions C. Max. distance from Earth	Duration	Remarks
Oct. 20–Nov. 5, 1995	STS 73 Kenneth D. Bowersox Kent Rominger Kathryn Thornton Catherine Coleman Michael Lopez-Alegria Fred Leslie Albert Sacco	230,158 lb (landing) 255 rev 174 mi	15 days 21 h 53 min	Columbia carried the second U.S. Microgravity Laboratory (USMI 2) Spacelab with science and technology experiments in areas such as fluid physics, materials science, biotechnology, combustion science, and commercial space processing technologies. Student interacted with the crew, discussing and comparing on-board microgravity experiments with similar ground-based experiments
Nov. 12–20, 1995	STS 74 Kenneth Cameron James Halsell Chris Hadfield Jerry Ross William McArthur	205,000 (landing) 128 rev 185 mi	8 days 4 h 31 min	Atlantis carried a Russian-built Docking Module (DM), two new solar arrays, and supplies for <i>Mir</i> on the second shuttle- <i>Mir</i> docking mission. After attaching the DM to <i>Mir's Kristal</i> module with the robot arm and linking the shuttle's airlock to it on Nov. 15, the <i>Atlantis</i> crew, including a Canadian (Hadfield), was welcomed by Gidzenko, Avdeyev, and Reiter. During 3 days of joint operation, the crews transferred the United States biomedical and microgravity science samples and data collected by the <i>Mir</i> 18, 19, and 2 <i>L</i> resident crews from the station to the shuttle, and fresh supplies to <i>Mir</i> . On Nov. 18, <i>Atlantis</i> undocked, leaving the DM attached to the station for future linkups.
Jan. 11–20, 1996	STS 72 Brian Duffy Brent W. Jett, Jr. Leroy Chiao Winston E. Scott Koichi Wakata Daniel T. Barry	217,000 lb (landing) 141 rev 290 mi	8 days 22 h 1 min	<i>Endeavour</i> retrieved, on Jan. 13, the Japanese satellite <i>SFU</i> (Space Flyer Unit) launched by an H-2 rocket in March 95. On Jan. 14, the Japanese crew member (Wakata) released the <i>OAST-Flyer</i> , a free flying platform with four autonomous experiments. On Jan. 16, the satellite was retrieved. On Jan. 14, Chiao and Barry performed a spacewalk to evaluate a portable work platform. A second EVA by Chiao and Scott, took place on Jan. 17 for further testing o spacesuit modifications and space station assembly tools and techniques.
Feb. 21–Sept. 2, 1996	<i>Soyuz TM-23</i> Yuri Onufrienko Yuri Usachev	14,400 lb 3027 rev 247 mi	193 days 19 h 11 min	The <i>Mir</i> 21 crew docked to the <i>Mir</i> station on Feb. 23, joining the three crew members of <i>Mir</i> 20, who returned in <i>TM-22</i> or Feb. 29. In the first month, on-board activities focused on national research, followed by the United States science program NASA 2 when astronaut Shannon Lucid joined the crew on Mar. 23, On May 21 and 24, Onufrienko and Usachev performed their second and third spacewalk to install and unfurl the Mir Cooperative Solar Array brought up by <i>Atlantis</i> (<i>STS</i> 74) and a fourth on May 30 to install a European earth resources camera on the <i>Priroda</i> module. On June 13, they installed a truss structure called Rapana to <i>Kvant</i> 1, taking the place of the similar Strela. Later, they worked on the French research program Cassiopeia when <i>Mir</i> 22 arrived with French astronaut Andre-Deshays. Onufrienko and Usachev returned with her on Sept. 2.
Feb. 22–Mar. 9, 1996	STS 75 Andrew Allen Scott Horowitz Franklin Chang-Diaz Jeffrey Hoffman Claude Nicollier Maurizio Cheli Umberto Guidoni	229,031 lb (landing) 250 rev 185 mi	15 days 17 h 39 min	Columbia carried three foreign astronauts among its crew (Cheli/ESA-Italy, Nicollier/ESA-Switzerland, Guidoni/ASI-Italy). Or Feb. 25, the Tethered Satellite System, on a reflight from STS 46 in July 1992, was deployed on the end of its thin 12.5-mi-long tethe line. Experiments could not be completed, however, since, with 12 mi of tether deployed, the cable broke close to the shuttle, and the satellite was lost in orbit. On-board activities then focused on the operation of the third U.S. Microgravity Payload (USMP 3), con tinuing research aimed at improving basic knowledge of materials under zero-g conditions.
Mar. 22–31, 1996	STS 76 Kevin Chilton Richard Searfoss Ronald Sega M. Richard Clifford Linda Godwin Shannon Lucid	246,335 lb (landing) 144 rev 246 mi	9 days 5 h 15 min 53 s	Atlantis lifted off for the third linkup with Russia's <i>Mir</i> station as part of Phase 1 of the <i>International Space Station</i> program After docking on Mar. 23 and hatch opening, the two crews of eight conducted joint operations and transferred 5000 lb o supplies and water to <i>Mir</i> . The first spacewalk by United States astronauts (Godwin, Clifford) with a shuttle attached to the Russian station took place on Mar. 27, to mount four experiments to its outside for assessing its environment. After concluding science and technology experiments in the <i>Spacehab</i> module in the carge bay, <i>Atlantis</i> undocked on Mar. 28, leaving Lucid, the first United States astronaut to be ferried to <i>Mir</i> by shuttle, aboard the station to join Onufrienko and Usachev for the next 188 days.
Apr. 23, 1996	Priroda Crewless	43,000 lb 	_	The <i>Priroda</i> module of <i>Mir</i> was launched on a Proton rocket Docking occurred on the first attempt on Apr. 26, despite the failure of two of the module's patteries on Apr. 25. On Apr. 27. <i>Prirod</i>

TABLE 1. Crewed space flights 1961–2006 (cont.)					
Dates launched and recovered	Designation and crew	A. WeightB. RevolutionsC. Max. distance from Earth	Duration	Remarks	
Apr. 23, 1996 (cont.)				was rotated from an axial to a radial docking port. The module carries a set of remote-sensing instruments and some microgravity experiments.	
May 19-29, 1996	STS 77 John H. Casper Curtis L. Brown Andrew S. W. Thomas Daniel W. Bursch	254,879 lb (landing) 161 rev 177 mi	10 d 39 min 18 s	Endeavour carried into orbit the commercial Spacehab module (fourth flight), with nearly 3000 lb of experiments and support equipment, an Inflatable Antenna Experiment (IAE) on the free- flying Spartan 207 platform, and four technology experiments called TEAMS. The crew included a Canadian (Garneau). The Spartan was released on May 20, and the 50-ft-diameter IAE inflated on its three 92-ft-long struts, the most complex and precise inflatable space structure ever. It was separated from Spartan later and burned up in the atmosphere. The satellite was retrieved on May 21, followed by further experiments which included rendezvous with the passive aerodynamically stabilized, magnetically damped satellite PAMS/STU (Satellite Test Unit).	
June 20–July 7, 1996	STS 78 Terence Henricks Kevin Kregel Richard M. Linnehan Susan J. Helms Charles E. Brady, Jr. Jean-Jacques Favier Robert Brent Thirsk	256,170 lb (landing) 271 rev 177 mi	16 d 21 h 48 min	Columbia carried the Life and Microgravity Sciences (LMS) Spacelab, equipped to focus on both living organisms in the low- gravity environment and the processing of various substances when gravity effects are greatly reduced. Crew included a French and a Canadian payload specialist (Favier, Thirsk). Mission set a shuttle duration record.	
Aug. 17, 1996– Mar. 2, 1997	<i>Soyuz TM-24</i> Valeriy Korzun Aleksandr Kaleri Claudie André-Deshays	14,400 lb 243 rev —	32 d 18 h 27 min/André- Deshays; 196 d 17 h 26 min/Korzun and Kaleri	The <i>Mir</i> 22 crew included a French female researcher-cosmonaut (André-Deshays). The three cosmonauts joined the <i>Mir</i> 21 crew on board <i>Mir</i> on Aug. 19. Prior to their arrival, on Aug.1, the crewless cargo ship <i>Progress M</i> -31 was undocked and deorbited over the Pacific; on Aug. 2, <i>Progress</i> 32 arrived. It was undocked on Aug. 18 to free up a docking port, and linked up again with <i>Mir</i> on Sept. 3.	
Sept. 16–26, 1996	STS 79 William F. Readdy Terrence W. Wilcutt Jay Apt Tom Akers Carl E. Watz John E. Blaha (up) Shannon Lucid (down)	249,352 lb (landing) 159 rev 239 mi	10 d 3 h 18 min; 188 d 5 h/ Lucid	Atlantis took off for its fourth docking with Russia's <i>Mir</i> , the sec- ond flight to carry a commercial <i>Spacehab</i> module in its cargo bay, and the first to carry a double version of it. Main mission purposes: to replace Lucid on <i>Mir</i> with Blaha for a 4-month stay; to carry 4600 lb of logistics, including food, clothing, experiment supplies, and spare equipment, to the space station; and to return 2200 lb of science samples and hardware to Earth. Five different experiments were turned off on the shuttle and rapidly transferred to <i>Mir</i> , the first such transfer of powered scientific apparatus. The two vehicles docked on Sept. 18, and separated on Sept. 23. The mission began a new era of permanent space occupancy by U.S. astronauts. Lucid set records for longest space flight, both for women and for U.S. astronauts (188 days 5 hours).	
Nov. 19–Dec. 7, 1996	STS 80 Kenneth D. Cockrell Kent V. Rominger Tamara E. Jernigan Thomas D. Jones Story Musgrave	226,954 lb (landing) 278 rev 220 mi	17 d 15 h 53 min	Columbia carried the 7876-lb (3600-kg) reusable German <i>ORFEUS-SPAS II</i> ultraviolet observatory and the 4650-lb (2100-kg) Wake Shield Facility (WSF), on its third flight, with its 4750-lb (2155-kg) carrier system. The <i>ORFEUS-SPAS</i> satellite was released on Nov. 19 and recovered with the RMS-robot arm on Dec. 3. The free-flying WFS was deployed on Nov. 22 and retrieved on Nov. 25. Both satellites were highly successful in their missions. The crew conducted biomedical and genetic experiments in weightlessness. Astronaut Musgrave tied NASA astronaut John Young's record for most space flights by any human being (6), and at age 61, he became the oldest person to date to fly in space. The flight set a new duration record for the space shuttle.	
Jan. 12–22, 1997	STS 81 Michael A. Baker Brent W. Jett, Jr. Peter (Jeff) Wisoff John M. Grunsfeld Marsha S. Ivins Jerry M. Linenger (up) John Blaha (down)	249,936 lb (landing) 160 rev 245 mi	10 d 4 h 56 min	Atlantis on the fifth mission (of 9) to <i>Mir</i> replaced astronaut John Blaha with Jerry Linenger as U.S. <i>Mir</i> resident and delivered various supplies. After docking on Jan. 14, the largest transfer of cargo items to date took place: 1400 lbs (635 kg) of water, 1137 lb (516 kg) of U.S. science equipment, 2206 lb (1000 kg) of Russian logistics, and 268 lb (122 kg) of miscellaneous material.	
Feb. 10-Aug. 14, 1997	Soyuz TM-25 Vasiliy Tsibliev Aleksandr Lazutkin Reinhold Ewald (Mir '97)	14,400 lb 2885 rev 246 mi	19 d 16 h 35 min/ Ewald; 184 d 22 h 6 min/ Tsibliev and Lazutkin	The <i>Mir 23</i> crew joined the three station occupants, which included U.S. astronaut Jerry Linenger. On Feb. 23, a small flash fire broke out in the <i>Kvant</i> module when a lithium perchlorate "candle," a secondary oxygen-generating system, malfunctioned. The crew extinguished the flames within 90 seconds, but had to wear	

Dates launched and recovered	Designation and crew	A. WeightB. RevolutionsC. Max. distance from Earth	Duration	Remarks
Feb. 10–Aug. 14, 1997 (<i>cont</i> .)				breathing apparatus and subsequently filtering masks because of smoke. A few hours later, operations returned to normal. Ewall returned to Earth with Korzun and Kaler; in <i>Soyuz TM-24</i> on Mar. 2 The two Elektron primary oxygen systems (based on electrolysi of waste water) having failed previously, only one of the secondar systems burning chemicals remained until arrival of the crewless resupply ship <i>Progress M-34</i> on Apr. 8. On June 25, the automate <i>Progress M-34</i> collided with <i>Mir</i> when Tsibliev attempted to lin it to the <i>Spektr</i> module during a redocking exercise trying ou the TORU manual guidance system. The collision damaged on of <i>Spektr</i> 's four solar arrays, a radiator and its hull, resulting i its depressurization—a major emergency. The crew was able to disconnect power and data cables and close the airtight hatcl in time. During the subsequent weeks, on-board operations were severely curtailed because of the power shortage and were furthe beset by heart problems of the <i>Mir</i> commander (Tsibliev) and various subsystems malfunctions. Tsibliev and Lazutkin returned to Earth on <i>Soyuz TM-25</i> , one week after the arrival of the replacemen crew (<i>Soyuz TM-26, Mir-24</i>).
Feb. 11–21, 1997	STS 82 Kenneth D. Bowersox Scott J. Horowitz Joseph R. Tanner Steven A. Hawley Gregory J. Harbaugh Mark C. Lee Steven L. Smith	213,312 lb (landing) 150 rev 370 mi	9 d 23 h 36 min	Discovery caught up with and retrieved the Hubble Space Tele scope (HST) on Feb. 13 for its second scheduled servicing mission since its launch in April 1990. After installing the observatory with the manipulator arm on a turntable in the cargo bay, Lee, Smith Harbaugh, and Tanner performed critical maintenance of the HST and significantly upgraded its scientific capabilities by exchang ing eight degraded or failed components and replacing two majo instruments, the Goddard High Resolution Spectrometer (GHRS and the Faint Object Spectrograph (FOS) with the more advancee Space Telescope Imaging Spectrograph (STIS) and the Near In frared Camera and Multi-Object Spectrometer (NICMOS), respect tively. The HST was deployed and released in perfect condition on Feb. 19.
April 4–8, 1997	STS 83 James D. Halsell Susan L. Still Janice Voss Michael L. Gernhardt Donald A. Thomas Roger K. Crouch Gregory T. Linteris	259,705 lb (landing) 63 rev 185 mi	3 d 23 h 33 min 30 s	Columbia carried the Spacelab module configured as the Micro gravity Science Laboratory-1 (MSL-1) with 33 investigations from four space agencies in the areas of crystallography, combustion science, materials science, and life sciences. The mission was to test some of the hardware, facilities, and procedures to be used or the International Space Station (ISS). In its international complex ity, the flight was intended to mirror the future work aboard the ISS bringing together academic, industrial, and governmental partners from around the world. One of the important new facilities designed to enable research in zero-g was the EXPRESS Rack (EXpedite Pro cessing of Experiments to the Space Station). However, on Apr. 6 a malfunction of one of the three power-generating fuel cells of the shuttle necessitated shutdown of the unit and early mission termi nation. Columbia had to return on Apr. 8 without accomplishing its mission.
May 15–24, 1997	STS 84 Charles J. Precourt Eileen M. Collins C. Michael Foale (up) Jerry Linenger (down) Carlos I. Noriega Edward T. Lu Jean-François Clervoy Elena V. Kondakova	221,087 lb (landing) 145 rev 247 mi	9 d 5 h 21 min	Atlantis successfully conducted the sixth <i>Mir</i> docking mission of the <i>ISS</i> /Phase I program, and remained linked to <i>Mir</i> for 5 days unloading 1025 lb (465 kg) water, 845 lb (383 kg) U.S. science equipment, 2576 lb (1168 kg) Russian logistics including a much needed Elektron oxygen generator unit and a vacuum cleaner, and 393 lb (178 kg) miscellaneous material. The mission also server to exchange U.S. astronaut Jerry Linenger, who had spent ove 132 days in <i>Mir</i> , against Michael Foale for another 4 months o continuing U.S. presence in space.
July 1–17, 1997	STS 94 James D. Halsell Susan L. Still Janice Voss Michael L. Gernhardt Donald A. Thomas Roger K. Crouch Gregory T. Linteris	259,705 lb (landing) 251 rev 185 mi	15 d 16 h 45 min	Repeat of mission STS 83 of Apr. 4–8, 1997, with same crew and payload. Columbia carried the Microgravity Science Laboratory (MSL-1), a continuation of NASA's micro-g research program in the fields of materials science, protein crystal growth and physics, with 33 investigations developed by scientists from four space agencies. Main focus: combustion and othe processes important to advanced industrial operations and engine development.
Aug. 5, 1997– Feb. 19, 1998	Soyuz TM-26 Anatoly Solovyev Pavel Vinogradov	14,400 lb ~3093 rev 244 mi	197 d 15 h 35 min/ Solovyev and Vinogradov	Soyuz TM-26 carried a specially trained crew for repairing the damaged space station. The <i>Mir 24</i> crew worked the next 6 ¹ / ₂ months to make repairs and reestablish proper function to al onboard systems, except for the <i>Spektr</i> module. When Solovyev and Vinogradov returned to Earth on Feb. 19 <i>Mir</i> was again in satisfactory condition for safe human occupancy.

Dates launched and recovered	Designation and crew	A. WeightB. RevolutionsC. Max. distance from Earth	Duration	Remarks
Aug. 7–19, 1997	STS 85 Curtis L. Brown Kent V. Rominger N. Jan Davis Robert L. Curbeam, Jr. Stephen K. Robinson Bjarni V. Tryggvason	217,910 lb (landing) 189 rev 185 mi	11 d 20 h 27 min	Discovery lifted off with the Cryogenic Infrared Spectrometers and Telescopes for the Atmosphere-Shuttle Pallet Satellite-2 (<i>CRISTA-SPAS-2</i>) on its second flight (and the fourth mission in a cooperative venture between the German Space Agency DARA and NASA). The 7724-Ib (3500-kg) satellite was released in orbit to measure infrared radiation emitted by the atmosphere; on Aug. 16 it was retrieved with the shuttle's mechanical arm. Other payloads included the Japanese Manipulator Flight Demonstration (MFD), that is, three separate tests of a robot arm being developed by Japan for its module JEM of the <i>International</i> <i>Space Station</i> (<i>ISS</i>); the Technology Applications and Science-1 (TAS-1) package with seven separate experiments for studies of atmospheric and topographic physics, solar energy, and testing of new thermal control devices; and the International Extreme Ultraviolet Hitchhiker-02 (IEH-02) with four experiments to research UV radiation from the stars, the Sun, and other sources in the solar system.
Sept. 25–Oct. 6, 1997	STS 86 James D. Wetherbee Michael J. Bloomfield Vladimir G. Titov Scott E. Parazynski Jean-Loup Chrétien Wendy B. Lawrence David A. Wolf (up) C. Michael Foale (down)	251,730 lb (landing) 169 rev 243 mi	10 d 19 h 21 min	Atlantis performed the seventh docking mission with <i>Mir</i> . On Oct. 2, Parazynski and Titov performed a 5 h 1 m EVA to retrieve four suitcase-sized Mir Environmental Effects Payload (MEEP) experiments from the exterior of the docking module where they had been attached by <i>STS 76</i> crew members in March 1976. They also left a large (121 lb or 55 kg) solar array cover cap at the outside for planned use in a future EVA as part of Russia's attempts to repair the damaged <i>Spektr</i> module, and evaluated the small jet-backpack called SAFER (Simplified Aid for EVA Rescue).
Nov. 19-Dec. 5, 1997	STS 87 Kevin R. Kregel Steven W. Lindsey Kalpana Chawla Winston E. Scott Takao Doi Leonid K. Kadenyuk	226,449 lb (landing) 251 rev 174 mi	15 d 16 h 34 min 4 s	<i>Columbia</i> lifted off with the U.S. Microgravity Payload-4 (USMP-4), the <i>Spartan 201-04</i> free-flyer, a Collaborative Ukrainian Experiment (CUE) for plant space biology research, and several other hitchhiker payloads. The launch, for the first time, incorporated a new flight maneuver by which the shuttle, at about 6 min into the flight, rolled to a "heads up" position, allowing early acquisition of the Tracking and Data Relay Satellite System (TDRSS) in geostationary orbit in lieu of the Bermuda tracking station and resulting in cost savings. On Nov. 21, the <i>Spartan</i> satellite was deployed with the RMS, Intended for investigations of physical conditions and processes of the solar corona and the hot outer layers of the Sun's atmosphere, the spacecraft did not self-activate and began to tumble on a failed attempt by Chawla to regrapple it with the RMS. It was manually recaptured on Nov. 24 by Scott and Doi during a 7 h 43 m spacewalk. On Dec. 3, Scott and Doi douring a 7 h 43 m spacewalk. On bec. 3, Scott and Doi during a row fue <i>SS</i> , including a telescoping (to 17.5 ft) crane for moving large Orbital Replacement Units (ORUS), and a prototype free-flying TV camera (AERCam/Sprint) for exterior inspections.
Jan. 22–31, 1998	STS 89 Terrence W. Wilcutt Joe F. Edwards, Jr. James F. Reilly II Michael P. Anderson Bonnie J. Dunbar Salizhan S. Sharipov Andrew S. Thomas (up) David A. Wolf (down)	251,612 lb (landing) 138 rev 247 mi	8 d 19 h 47 min	Endeavour completed the eighth docking mission to <i>Mir</i> . The crew transferred 774 items of cargo including a new computer and air-conditioning equipment. Wolf replaced Thomas (who became the seventh U.S. astronaut, NASA 7, to stay on Mir for a long-duration mission). Wolf had spent 126 days aboard <i>Mir</i> , bringing the total U.S. crew time on <i>Mir</i> to 843 days. Other payloads on <i>STS</i> 89 included the <i>Spacehab</i> double module with science experiments and new research equipment for NASA 7.
Jan. 29–Aug. 25, 1998	Soyuz TM-27 Talgat Musabayev Nikolai Budarin Leopold Eyharts	14,400 lb ~3290 rev 239 mi	207 d 51 min/ Musabayev and Budarin	<i>Soyuz TM-27</i> launched with the new <i>Mir 25</i> crew and a researcher (Eyharts) from France's space agency CNES. After docking to <i>Mir</i> on Jan. 31, the space station had five occupants until Feb. 19, when the <i>Mir 24</i> crew (Solovyev and Vinogradov), along with Eyharts, departed in <i>Soyuz TM-26</i> .
Apr. 17–May 3, 1998	STS 90 Richard A. Searfoss Scott D. Altman Richard M. Linnehan Kathryn P. Hire Dafydd Rhys Williams Jay Clark Buckey, Jr. James A. Pawelczyk	232,500 lb (landing) 255 rev 178 mi	15 d 21 h 50 min	<i>Columbia</i> 's prime mission for the <i>Spacelab</i> module (on its sixteenth flight) was to conduct microgravity research in the human nervous system, including brain, spinal cord, nerves, and sensory organs. Acting both as subjects and operators, the crew performed 26 experiments organized in eight groups: Adult Neuronal Plasticity, Mammalian Development, Aquatic, Neurobiology, Autonomic Nervous System, Sensory Motor and Performance, Vestibular, and Sleep. Included were studies on a variety of species, such as rats, mice, toad fish, zebra fish, crickets, and snails. Secondary objectives included measurement of shuttle vibration forces, demonstration of the Bioreactor system used for cell growth, and three Getaway Special (GAS) payloads.

Dates launched and recovered	Designation and crew	A. WeightB. RevolutionsC. Max. distance from Earth	Duration	Remarks
June 2–12, 1998	STS 91 Charles J. Precourt Dominic L. Pudwill Gorie Franklin R. Chang-Diaz Wendy B. Lawrence Janet Lynn Kavandi Valery V. Ryumin Andrew Thomas (down)	259,834 lb (landing) 154 rev 237 mi	9 d 19 h 48 min	The last of nine docking missions with <i>Mir</i> ended Phase of the <i>International Space Station (ISS)</i> program as <i>Discover</i> linked up with the space station on June 4. Payloads flying of <i>Discovery</i> included the Alpha Magnetic Spectrometer (AMS) of the Department of Energy to study high-energy particles from dee space, several experiments of advanced technology, human lift sciences, <i>ISS</i> risk mitigation, and microgravity physics, four Get Away Specials (GAS) experiments, and two Space Experiment Module (SEM) canisters with experiments from U.S. middle schoo high school, and college students. U.S. astronauts performed the majority of science research on <i>Mir</i> while their Russian hosts primarily concerned themselves with station upkeep and maintenance. This international cooperation provided a wealth of new knowledge about space flight, particularly with regard to space station design, operation, and maintenance, and was the first of it kind on this magnitude.
Aug. 13, 1998– Feb. 28, 1999	<i>Soyuz TM-28</i> Gennady Padaika Sergei Avdeev Yuri Baturin	14,400 lb ~3109 rev 225 mi	11 d 19 h 41 min/ Baturin; 198 d 15 h 30 min/ Padalka	The <i>Mir 26</i> crew docked to <i>Mir</i> , using the manual backup system TORU due to prior failure of one of two automatic KURS systems on <i>Mir</i> . Baturin returned with the <i>Mir 25</i> crew on Aug. 25.
Oct. 29–Nov. 7, 1998	STS 95 Curt Brown, Jr. Steven W. Lindsey Scott Parazynski Stephen Robinson Pedro Duque Chiaki Mukai John Glenn	227,783 lb (landing) 134 rev 348 mi	8 d 21 h 45 min	Discovery lifted off with a Japanese woman astronaut on he second flight (Mukai), and payload specialist Senator John Glenn at 77 years the oldest human ever to fly into space and also the first American to fly into Earth orbit. To take scientific advantage of Glenn's advanced age and his unbroken medica history records with NASA over the last 40 years, about 10 o the flight's 83 experiments and payloads investigated questions pertaining to geriatrics in space, that is, the phenomena of natura aging vs. weightlessness-induced effects closely resembling them The other experiments made up one of the most diverse sets of commercial and scientific research investigations addressing biotechnology, life sciences, solar physics, and astronomy. The payload included the <i>Spartan 201-04</i> free-flyer on its reflight after its failed deployment on mission STS 87. It was deployed on Nov. and successfully retrieved 2 days later. The <i>Discovery</i> mission was the last U.S. flight prior to the new era of the <i>International Space Station</i> .
Nov. 20, 1998	FGB/Zarya Crewless	44,088 lb 240 mi	-	Funded by the U.S. and built by Russia (Khrunichev State Research and Production Center, KhSC, in Moscow), the power and contro module <i>FGB</i> (<i>Funktionalnyi-grusovoi blok</i>) named <i>Zarya</i> (Dawn was launched on a Proton rocket from Russia's cosmodrome in Kazakhstan as the first component of the <i>International Space</i> <i>Station</i> .
December 4–15, 1998	STS 88 Robert D. Cabana Frederick W. Sturckow Jerry L. Ross Nancy J. Currie James H. Newman Sergei K. Krikalev	263,927 lb (orbit) 185 rev 243 mi	11 d 19 h 18 min	<i>Endeavour</i> lifted off with the connecting module Node 1/Unity the first U.Sbuilt station element for the <i>ISS</i> . Rendezvous and connection with <i>ISS/Zarya</i> occurred on Dec. 6. On Dec. 14, the crew deployed two satellites: an Argentinian satellite (<i>SAC-A</i>) or Dec. 13 and a U.S. Air Force test satellite (<i>MightySat</i>). Additiona payloads were an IMAX camera for filming from the cargo bay an educational Space Experiment Module (SEM-07), and othe experiments.
Feb. 20–Aug. 27, 1999	<i>Soyuz TM-29</i> Viktor Afanasyev Jean-Pierre Haignere Ivan Bella	14,400 lb ~2950 rev 227 mi	7 d 21 h 56 min/ Bella; 187 d 22 h 18 min/ Afanasyev and Haignere; 379 d 14 h 52 min/ Avdeev	The new <i>Mir 27</i> crew docked on Feb. 22, and began some 100 experiments, with an incubator with 60 quail eggs hatching a few days after docking, to study effects of space mission on embryonic and postembryonic developments. Fledglings returned to Earth (only two survived) with Padalka and Bella in <i>Soyuz TM-28</i> or Feb. 28. On July 18, the Progress M-42 cargo drone docked to deliver a backup computer intended to provide motion contror redundancy during a forthcoming period of approx. 6 months of crewless (mothballed) operation of <i>Mir</i> , for the first time in 10 years.
May 27–June 6, 1999	STS 96 Kent V. Rominger Rick D. Husband Tamara E. Jernigan Ellen Ochoa Daniel T. Barry Julie Payette (Canada) Valery Tokarev (Russia)	262,035 lb (orbit) 153 rev 234 mi	9 d 19 h 13 min	The first visit in history to the <i>International Space Station</i> occurred when <i>Discovery</i> docked flawlessly to the Unity/ <i>Zarya</i> complex on May 29. The crew began a spacewalk to transfer two cranes (one Russian, one U.Smade) from the shuttle's cargo bay to the station's outside, installed two new portable foot restraints, and attached three bags with tools and handrails for future assembly operations. The shuttle's hatch to the <i>ISS</i> mating adapter PM/2 was opened on May 29 for leak and pressurization tests. Or May 30 the crew members began transferring supplies, equipment

Dates launched and recovered	Designation and crew	A. WeightB. RevolutionsC. Max. distance from Earth	Duration	Remarks
May 27–June 6, 1999 (<i>cont.</i>)				and water. They also replaced 18 recharge controller units for the si Zarya batteries, installed mufflers over noisy FGB fans, and replace a power distribution unit and transceiver of the Node communicatio system. At work conclusion, the crew had transferred a total or 3567 lb (1618 kg) material to the ISS, including 686 lb (311 kg) wate After releasing the student satellite Starshine on June 5, Discover returned.
July 23–27, 1999	STS 93 Eileen M Collins Jeffrey S. Ashby Catherine G. Coleman Steven A. Hawley Michel Tognini	219,980 lb (landing) 177 rev 79 mi	4 d 22 h 49 min 37 s	Columbia launched, with the Chandra X-Ray Observatory (AXAF the third in NASA's series of Great Observatories. The observator was successfully deployed and subsequently boosted itself to final orbit of 86,992 by 6034 mi (139,188 by 9655 km), with a perio of 63 h 28 m. Secondary objectives included firing of Columbia' jet thrusters at various times during the flight to help an Air Forc satellite gather data on the characteristics of jet plumes in space Also, the Southwest Ultraviolet Imaging System (SWUIS) collected data on ultraviolet light originating from a variety of planetar bodies.
Dec. 19–27, 1999	STS 103 Curtis L. Brown Scott J. Kelly Steven L. Smith Jean-François Clervoy John M. Grunsfeld Michael Foale Claude Nicollier	211,129 lb (landing) 119 rev 367 mi	7 d 23 h 11 min	Discovery conducted the third maintenance/repair mission to th Hubble Space Telescope (HST). The mission, originally planned for 2000, became necessary when the HST had to be put in "safe (dormant) mode with only two of its six gyroscopes functioning The crew installed six new gyros and six voltage/temperatur improvement kits; replaced the main computer, the Fine Guidanc Sensor, and a radio transmitter; and installed a new digital recorder as well as new insulation on two equipment bay doors.
Feb. 11–22, 2000	STS 99 Kevin R. Kregel Dominic L. P. Gorie Gerhard P. J. Thiele Janet L. Kavandi Janice Voss Mamoru Mohri	225,669 lb (landing) 181 rev 146 mi	11 d 5 h 38 min	Shuttle Radar Topography Mission (SRTM), with primary objectiv to acquire a high-resolution topographic map of Earth's lan mass between 60°N and 56°S, and to test new technologies for deployment of large rigid structures and measurement of the distortions to extremely high precision. The 87 cube-shaped "bays of the carbon fiber-reinforced plastic, stainless steel, alpha titanium and Invar radar mast carried active radar antennas for C-band an X-band wavelengths. SRTM will produce topographic maps of Eart 30 times as precise as the best global maps to date.
Apr. 4–June 16, 2000	Soyuz TM-30 Sergei Zalyotin Alexander Kaleri	14,400 lb 1139 rev 210 mi	72 d 19 h 43 min	First crew return to <i>Mir</i> after the 14-year-old station had floate for 223 days without occupants in "mothballed" condition. Prior t their flight, the crewless <i>Progress M1-1</i> (first of an upgraded versio with double tankage volume) docked automatically with the statio on Feb. 3. A second Progress flight to <i>Mir</i> occurred on Apr. 27 The crew had the task of returning the station to safe operatin condition. After 2 weeks of tests, they succeeded in sealing a air leak in a hatch between the core module's transfer compartmer and the airless <i>Spektr</i> module. On May 12, an EVA was conducte to test new techniques of leak repair. The <i>Mir 28</i> crew returned t Earth in <i>TM-30</i> , leaving the renovated station sealed up and agai on automatic control.
May 19–29, 2000	STS 101 James D. Halsell Scott J. Horowitz Mary Ellen Weber Jeffrey N. Williams James S. Voss Susan J. Helms Yuri V. Usachev	225,504 lb (landing) 239 rev 155 mi	9 d 20 h 9 min 9 s	The primary objectives of the second crewed visit to the <i>Internationa</i> Space Station (ISS) were to remove and replace FGB/Zary hardware that had failed, such as battery systems, and to reboos the ISS to an altitude consistent with the projected Servic Module/Zvezda launch and docking. Atlantis docked with ISS o May 21, and Williams and Voss reinstalled a loose U.S. crane assembled and installed a Russian "Strela" cargo crane, replace a communications antenna outside the Node Unity, and installe eight handrails on the outside wall of Unity. Repair and replacemer activities extended over the following 4 days. In three separat maneuvers, ISS altitude was increased 28.7 mi (45.9 km) to 237.7 : 229.4 mi (380.3 × 367.0 km). The crew transferred over 3300 I (1500 kg) of gear from Atlantis to ISS and returned other cargo to the shuttle.
July 12, 2000	Service Module (SM) Zvezda Crewless	42,000 lb 	-	The Service Module (SM), named Zvezda (Star), the first Russian funded and -owned element of the <i>ISS</i> , was launched on a Proto rocket from Baikonur Cosmodrome in Kazakhstan. The docking or Zvezda with the <i>ISS</i> on July 25, executed very gently at 8 in./ (0.2 m/s) by the <i>FGB/Zarya</i> as active partner, marked the arrival of the first human living quarters for the new outpost. The new <i>IS</i> stack spanned 119 ft (36 m) in length, and its mass had grown t almost 60 tons.

Dates launched and recovered	Designation and crew	A. Weight B. Revolutions C. Max. distance from Earth	Duration	Remarks
Sept. 8–20, 2000	STS 106 Terrence W. Wilcutt Scott D. Altman Daniel C. Burbank Edward T. Lu Richard A. Mastracchio Yuri I. Malenchenko Boris V. Morukov	221,271 lb (landing) 185 rev 200 mi	11 d 19 h 11 min 2 s	Atlantis flew the third logistics/outfitting flight to the space station Lu and Malenchenko conducted a space walk to prepare the ISS exterior for future operations. During a week of docked operations the crew worked as movers, cleaners, plumbers, electricians, and cable installers, stowing more than 6600 lb (3 metric tons) of supplies and filling the disposable <i>Progress</i> M1-3 ship (which had docked at <i>Zvezda</i> 's aft port on August 8) with trash and discarded materials.
Oct. 11–22, 2000	STS 92 Brian Duffy Pamela A. Melroy Leroy Chiao William S. McArthur Peter J. K. Wisoff Michael E. Lopez-Alegria Koichi Wakata	204,455 lb (landing) 202 rev 200 mi	12 d 21 h 43 min	Discovery brought the first truss structure segment (called Z1 for its zenith port location on Node Unity), and the third docking adapter (PMA-3, pressurized mating adapter) to the space station. The Z1 truss was installed on the Node, with necessary electrica connections being hooked up and other deployment activities being accomplished during a space walk on Oct. 15. PMA-3 was installed on the Node's nadir port during a second space walks for this mission.
Oct. 31, 2000–Mar. 20, 2001	<i>Soyuz TM-31</i> , 2R/1S William Shepherd Yuri Gidzenko Sergei Krikalev	14,400 lb — 236 mi	140 d 23 h 38 min	The first resident crew (Expedition One) was launched to the International Space Station. After liftoff in Baikonur, the Soyuz-U carrier injected the Soyuz TM-31 (ISS mission 2R) into orbit. Docking occurred on Nov. 2. The second crewless tanker ship Progress M1-4 (2P) arrived in November—followed by STS 97 in December with the P6 photovoltaic power module; STS 98 (5A) with the U.S. Laboratory Destiny in February 2001; a third crewless logistics flight, Progress M-244 (3P), in the same month; and the Lab outfitting mission 5A.1/STS 102 in March.
Nov. 30–Dec. 11, 2000	STS 97 Brent W. Jett Michael J. Bloomfield Joseph R. Tanner Marc Garneau Carlos I. Noriega	197,879 lb (landing) 170 rev 234 mi	10 d 19 h 58 min	Endeavour flew Mission 4A to the <i>ISS</i> , the first shuttle visit to the station with an <i>ISS</i> crew awaiting it. Garneau used the remote manipulator system (RMS) to "park" the 17-ton ITS (integrated truss segment) P6 above the cargo bay to allow for its temperature adjustment. Then Garneau attached P6, attached to the Z1 ITS, or top of the Node, assisted by Noriega and Tanner on an EVA. Jett deployed the two solar array wings, jointly spanning 240 ft (73 m). Deployment of the second wing was delayed one day when the tensioning of the two array blankets of the first wing malfunctioned Two additional EVAs were conducted, the latter to correct the starboard wing's tension and to install an electrical potential sensor The P6 provided the <i>ISS</i> with up to 62 kW of power.
Feb. 7–20, 2001	STS 98 Kenneth D. Cockrell Mark L. Polansky Robert L. Curbeam Marsha S. Ivins Thomas D. Jones	254,694 lb (orbiter at liftoff) 201 rev 239 mi	12 d 21 h 20 min	Atlantis/5A carried U.S. Laboratory module Destiny to the Interna- tional Space Station. The 28 \times 14 ft (8.5 \times 4.3 m) Lab was installed at the Node by Ivins using the (RMS) robot arm, assisted by space walkers Curbeam and Jones. Two more EVAs followed as the ISS crew completed setup and activation of critical and noncritical Lab systems.
Mar. 8–21, 2001	STS 102 James D. Wetherbee James M. Kelly Andrew S. W. Thomas Paul W. Richards James S. Voss (up) Susan J. Helms (up) Yury V. Usachev (up) William Shepherd (down) Sergei Krikalev (down) Yuri P. Gidzenko (down)	198,507 lb (orbiter at liftoff) 201 rev 242 mi	12 d 19 h 49 min 166 d 6 h 41 min/Usachev, Voss, Helms	<i>Discovery</i> /5A.1, the eighth shuttle mission to the <i>ISS</i> , carried resupply and systems/science racks for outfitting the U.S. Lab Destiny, all packed in Leonardo, the first of three Italian-built multipurpose logistics modules (MPLMs) used as "moving vans". During the docked period, the MPLM was transferred to the <i>ISS</i> Node's nadir port, unloaded, filled with trash and discarded equipment, and returned to the shuttle. In two space walks, crew members worked outside to support <i>ISS</i> assembly. In the first crew rotation, Usachev, Voss, and Helms became the new station crew replacing Shepherd, Krikalev, and Gidzenko who, after 141 days in space, returned to Earth.
Apr. 19–May 1, 2001	STS 100 Kent V. Rominger Jeffrey S. Ashby Chris A. Hadfield John L. Phillips Scott E. Parazynski Umberto Guidoni Yuri V. Lonchakov	228,188 lb (orbiter at liftoff) 186 rev 253 mi	11 d 21 h 30 min	Endeavour on Mission ISS-6A carried the 3960-lb (1800-kg) Canada-built Space Station Remote Manipulator System (SSRMS) the second Italian-built MPLM Raffaello (total weight loaded 9000 lb or 4100 kg), and three non-U.S. crew members. The mission featured two EVAs by Parazynski and Hadfield, one to support transfer and deployment of the SSRMS Canadarm2 from the shuttle payload bay to the <i>ISS</i> Lab module and install a UHF antenna, the second for external connections and removal of an ECS (Early Communications System) antenna from the Node. After the MPLM was berthed at the Node, the crew transferred two EXPRESS payload racks and individual components from it to the <i>ISS</i> and loaded it with return cargo. <i>Endeavour</i> undocked on April 29.

Dates launched and recovered	Designation and crew	A. WeightB. RevolutionsC. Max. distance from Earth	Duration	Remarks
Apr. 28–May 6, 2001	<i>Soyuz TM-32</i> , 2S Talgat A. Musabaev Yuri M. Baturin Dennis Tito	14,400 lb 123 rev 252 mi	7 d 22 h 4 min	The first "taxi" flight to the <i>ISS</i> brought a fresh <i>Soyuz</i> returnvehicle for the <i>ISS</i> crew. The flight also carried the first commercial space tourist, U.S. businessman Dennis Tito. The <i>Soyuz</i> capsule docked at the <i>FGB</i> nadir port on Apr. 30. On May 5, the three visitors climbed into the old <i>Soyuz TM-31</i> , docked at the Service Module aft end, and returned to Earth, landing in Kazakhstan.
July 12–24, 2001	STS 104 Steven W. Lindsey Charles O. Hobaugh Michael L. Gernhardt Janet L. Kavandi James F. Reilly	258,222 lb (orbiter at liftoff) 201 rev 250 mi	12 d 18 h 35 min	Atlantis, on mission ISS-7A, carried the 6.5-ton Joint Airlock <i>Quest</i> , which was installed at the starboard berthing port of the Node, completing Phase 2 of the <i>ISS</i> program. Also installed were four high-pressure tanks with oxygen and nitrogen. These activities were supported by three space walks by Gernhardt and Reilly, the last one from the new airlock.
Aug. 10–22, 2001	STS 105 Scott J. Horowitz Frederick W. Sturckow Patrick G. Forrester Daniel T. Barry Frank L. Culbertson (up) Vladimir N. Dezhurov (up) Mikhail Tyurin (up) Yury V. Usachev (down) James S. Voss (down) Susan J. Helms (down)	257,748 lb (orbiter at liftoff) 185 rev 250 mi	11 d 21 h 13 min; 128 d 20 h 45 min/Culbertson, Dezhurov, Tyurin	<i>Discovery</i> delivered the Expedition 3 crew (Culbertson, Dezhurov, Tyurin) to the <i>International Space Station</i> and picked up the second station crew (Usachev, Voss, Helms) on the second crew rotation mission. As ISS assembly mission 7A.1, <i>Discovery</i> also brought equipment and supplies, carried in the MPLM (multipurpose logistics module) Leonardo on its second flight, which was docked temporarily at the Node for unloading and reloading with return cargo. Barry and Forrester conducted two space walks to install an ammonia servicer, materials experiment carriers, handrails, and cables.
Sept. 14, 2001	4R (DC-1) Crewless	14,500 lb 250 mi	_	The Russian Docking Compartment (DC-1) <i>Pirs</i> was launched on a Soyuz-U rocket from Baikonur/Kazakhstan at 7:35 p.m. EDT. The Stikovochnly Otsek No. 1 (SO1) linked up with the <i>ISS</i> on Sept. 16. The new 16-ft-long (4.88-m) module, with a mass of 8000 lb (3.6 metric tons), is an additional docking port for future Soyuz and Progress vehicles arriving at the station, plus an added stowage area and an airlock for EVAs in Russian Ortan-M spacesuits. DC-1 was launched attached to the cargo ship <i>Progress M-SO1</i> , a Soyuz-derived instrumentation and propulsion section of about 3000 lb (1.36 metric tons), which was jettisoned on Sept. 26. <i>ISS</i> mass now totaled 303,500 lb (137.67 metric tons).
Oct. 21–30, 2001	So <i>yuz TM-33,</i> 3S Viktor Afanasyev Konstantin Kozeyev Claudie Haigneré (France)	14,400 lb ~153 rev 257 mi	9 d 19 h 0 min	Second <i>ISS</i> "taxi" flight, bringing a fresh Soyuz crew return vehicle (CRV) to the station, good for another 200 days. Haigneré was fly- ing commercially. Crew assignment included the French Andromedé (Andromeda) science mission. Afanasyev, Kozeyev, and Haigneré departed in the old <i>Soyuz TM-32</i> and landed in Kazakhstan under automatic control.
Dec. 5–17, 2001	STS 108 Dominic L. Gorie Mark E. Kelly Linda M. Godwin Daniel M. Tani Yuri Onufrienko (up) Carl E. Walz (up) Daniel W. Bursch (up) Frank Culbertson (down) Vladimir Dezhurov (down) Mikhail Tyurin (down)	~257,000 lb (orbiter at liftoff) 186 rev 334 mi	11 d 19 h 36 min; 196 d 20 h 39 min/ Onufrienko, Walz, Bursch	On first utilization flight (UF-1) of the <i>ISS</i> program, <i>Endeavour</i> carried MPLM <i>Raffaello</i> with approximately 3 tons (2.7 metric tons) of cargo plus the fourth expeditionary crew (Onufrienko, Bursch, and Walz) to rotate with the Expedition 3 crew (Culbertson, Dezhurov, and Tyurin). The MPLM was transferred to the Node nadir port with the shuttle RMS. Tani and Godwin conducted an EVA on Dec. 10. After cargo unloading and its replacement with discarded equipment and waste, <i>Raffaello</i> was returned to the shuttle cargo bay. Before its undocking, <i>Endeavour</i> conducted three reboosts of the station, with a total mean altitude increase of 7.6 n.mi. (14 km).
Mar. 1–12, 2002	STS 109 Scott D. Altman Duane G. Carey John M. Grunsfeld Nancy J. Currie Richard M. Linnehan James H. Newman Michael J. Massimino	257,917 lb (orbiter at liftoff) 165 rev 359 mi	10 d 22 h 10 min	Fourth "service call" to the <i>Hubble Space Telescope (HST)</i> was per- formed by <i>Columbia</i> after undergoing more than 100 improvements since its last fiight (<i>STS</i> 93). During flve space walks, <i>HST</i> received new, more durable solar arrays (SA) with their diode boxes (DBAs), a large gyroscopic Reaction Wheel Assembly (RWA) for pointing, a new power control unit (PCU), a new advanced camera for surveys (ACS), and an experimental cryocooler to restore the dormant Near Infrared Camera and Multi-Object Spectrometer (NICMOS) to oper- ation. After reboost on Mar. 8 by about 4 mi (6.4 km), <i>HST</i> was deployed from <i>Columbia</i> on Mar. 9.
Apr. 8–19, 2002	STS 110 Michael J. Bloomfleld Stephen N. Frick Rex J. Walheim Ellen Ochoa Lee M.E. Morin Jerry L. Ross Steven L. Smith	257,079 lb (orbiter at liftoff) 171 rev 250 mi	10 d 19 h 43 min	Atlantis flew mission 8A to the <i>ISS</i> , carrying the S0 truss and the Mobile Transporter (MT). Four EVAs were conducted to install the S0 and the MT. On Apr. 14 the MT rail cart traversed a distance of 72 ft (22 m) during this first operation of a "railroad in space." This was also the first flight during which the SSRMS robot arm of the <i>ISS</i> was used to maneuver spacewalkers around the station, and the first where all EVAs of a shuttle crew were performed from the station's own airlock, <i>Quest</i> .

Dates launched and recovered	Designation and crew	A. Weight B. Revolutions C. Max. distance from Earth	Duration	Remarks
Apr. 25–May 4, 2002	Soyuz TM-34 Yuri Gidzenko Roberto Vittori (Italy) Mark Shuttleworth (South Africa)	14,400 lb ~154 rev 256 mi	9 d 21 h 25 min	Third "taxi" flight to deliver a fresh CRV to <i>ISS</i> . Shuttleworth was the second "space tourist." A science program, sponsored by Italy and South Africa, included 11 experiments. Gidzenko, Vittori, and Shuttleworth returned in <i>Soyuz TM-33</i> , which had been docked to the <i>ISS</i> since October 2001.
June 5–19, 2002	STS 111 Kenneth D. Cockrell Paul S. Lockhart Franklin Chang-Diaz Philippe Perrin (France) Valery Korzun (up) Peggy Whitson (up) Sergei Treschev (up) Yuri Onufrienko (down) Carl E. Walz (down) Daniel W. Bursch (down)	256,889 lb (orbiter at liftoff) 217 rev 240 mi	13 d 21 h 35 min; 184 d 22 h 14 min 23 s/Korzun, Whitson, Treschev	Endeavour, on ISS Mssion UF-2, carried the flfth ISS resident crew (Korzun, Whitsun, and Treschev), returning with the fourth station crew (Onufrienko, Bursch, and Walz). Its cargo comprised MPLM <i>Leonardo</i> on its third flight, loaded with 5600 lb (2500 kg) of equipment and consumables; the Mobile Base System (MBS); a replacement wrist roll (WR) joint for the SSRMS; meteoroid/orbital debris protection shields (ODPs) for the <i>Zvezda</i> Service Module; and new science payloads. During three spacewalks, astronauts installed the MBS on the MT (mobile transporter), temporarily stowed the ODPs, and replaced the SSRMS WR joint.
Oct. 7–18, 2002	STS 112 Jeffrey S. Ashby Pamela A. Melroy David A. Wolf Sandra H. Magnus Piers J. Sellers Fyodor N. Yurchikhin (Russia)	256,917 lb (orbiter at liftoff) 172 rev 252 mi	10 d 19 h 58 min	Atlantis flew Mission 9A to the <i>ISS</i> . On Oct. 10, its main payload, the 14.5-ton (13.1-metric-ton) S1 truss element, was transferred with the Canadarm2 from the shuttle cargo bay to the <i>ISS</i> and firmly attached to the S0 center truss segment atop the Lab module. The installation for the first time featured the SSRMS operating from its new MBS. The crew also transferred about 1000 lb (450 kg) of supplies, equipment and other cargo to the station. Three spacewalks were performed to connect the S1 and S0 trusses with electrical, fiber-optical, data, and coolant umbilicals, and to install external systems.
Oct. 29–Nov. 9, 2002	Soyuz TMA-1, 5S Sergei Zalyotin Yuri Lonchakov Frank De Winne (ESA, Belgium)	15,920 lb ~170 rev 260 mi	10 d 20 h 53 min	Fourth taxi flight to the <i>ISS</i> , to replace the previous CRV, <i>Soyuz TM-34</i> . <i>TMA-1</i> was the first of a new version of the Soyuz, with modifications financed by NASA for improved safety and widened crewmember size range. A science program sponsored by Belgium and the Russian space agency Rosaviakosmos (RSA) included 23 experiments. The crew returned on the "old" <i>Soyuz TM-34</i> .
Nov. 23–Dec. 7, 2002	STS 113 Jim Wetherbee Paul Lockhart Michael Lopez-Alegria John Herrington Kenneth Bowersox (up) Nikolai Budarin (up) Donald Pettit (up) Valery Korzun (down) Peggy Whitson (down) Sergei Treschev (down)	256,747 lb (orbiter at liftoff) 215 rev 247 mi	14 d 6 h 47 min; 161 d 1 h 14 min 38 s/ Bowersox, Budarin, Petit	Mission <i>ISS</i> -11A on <i>Endeavour</i> carried the Expedition 6 crew (Bow- ersox, Budarin, and Pettit), the fourth truss element, P1 (of 11 total), and other equipment and resupply. The 14.5-ton (13.1-metric-ton) P1 was transferred and installed on the S0 truss portside, and three spacewalks were conducted to connect the P1 and install other gear. After <i>Endeavour</i> undocked on Dec. 2, it launched two tethered mini- satellites called MEPSI (microelectromechanical systems-based Pi- cosat Inspector), and returned with the Expedition 5 crew (Korzun, Whitson, and Treschev).
Jan. 16-Feb. 1, 2003	STS 107 Rick D. Husband William C. McCool David M. Brown Kalpana Chawla Michael P. Anderson Laurel B. Clark Ilan Ramon (Israel)	263,706 lb (orbiter at liftoff) 255 rev 177 mi	15 d 22 h 21 min	<i>Columbia</i> flew on a research mission with 32 payloads and 59 sepa- rate investigations. Both in the shuttle middeck and in the <i>Spacehab</i> Research Double Module (RDM), on its first flight in the cargo bay, the crew worked on a mixed complement of competitively selected and commercially sponsored research in the space, life, and physi- cal sciences. <i>Columbia</i> was lost with its crew during reentry on Feb. 1, 16 min before planned landing at Kennedy Space Center, when it violently disintegrated over Texas. Debris from the orbiter fell in Texas and other states. The Columbia Accident Investigation Board (CAIB) later concluded that one of the left wing's leading edge RCC (reinforced carbon-carbon) elements or an associated filler strip of the heat shield had been compromised during ascent to orbit by a piece of debris, possibly a suitcase-sized chunk of foam insulation blown off the external tank by the supersonic air stream, hitting the leading edge and rendering the wing unable to withstand reentry heating longer than about 8 min after entry interface.
Apr. 25–Oct. 27, 2003	<i>Soyuz TMA-2</i> , 6S Yuri Malenchenko Ed Lu	15,920 lb ~2889 rev 256 mi	184 d 22 h 47 min	The first crew rotation flight by a Soyuz carried Expedition 7, a two-man "caretaker" crew, to stretch out <i>ISS</i> consumables during the shuttle stand-down. <i>Soyuz TMA-2</i> docked to the <i>ISS</i> on April 28, replacing the previous CRV, <i>Soyuz TMA-1</i> . The Expedition 6 crewmembers (Bowersox, Budarin, and Pettit) returned to Earth in <i>Soyuz TMA-1</i> , experiencing an unexpected switch of their onboard computer to the backup reentry mode of pure ballistic descent, which almost doubled their peak deceleration (~9 g), and missed the primary landing site by an "undershoot" of ~300 mi (480 km), landing on May 3. Recovery forces reached the landing site after a delay of 4.5 h, finding the crew in good health outside the capsule.
Dates launched and recovered	Designation and crew	A. Weight B. Revolutions C. Max. distance from Earth	Duration	Remarks
-----------------------------------	---	---	--	--
Oct. 15–16, 2003	Shenzhou 5 ("Divine Vessel 5") Yang Liwei	17,000 lb 14 rev 213 mi	21 h 23 min	Orbital launch of the first Chinese "taikonaut" on the new huma rated Long March (Chang Zheng, CZ) 2F rocket. Landing was abo 600 mi (1000 km) east of the launch area, the Jinquan Satelli Launch Center in the Gobi Desert.
Oct. 18, 2003-Apr. 29, 2004	<i>Soyuz TMA-3</i> , 7S Michael Foale Alexander Kaleri Pedro Duque (ESA, Spain)	15,920 lb ~3053 rev 249 mi	9 d, 21 h 3 min/ Duque; 194d 18 h 35 min/ Foale and Kaleri	Second <i>ISS</i> crew rotation flight by a Soyuz carried Expeditio 8 (Foale and Kaleri), the second "caretaker" crew, and visitir crewmember Duque. On Oct. 27, the previous CRV, <i>Soyuz TMA</i> -undocked from the of the <i>FGB/Zarya</i> module nadir port, where it ha stayed for 182 days, and landed in Kazakhstan with Malenchenk Lu, and Duque.
Apr. 18–Oct. 24, 2004	<i>Soyuz TMA-4</i> , 8S Gennady Padalka Michael Fincke André Kuipers (ESA, Netherlands)	15,920 lb ~2951 rev 241 mi	10 d 20 h 52 min/ Kuipers; 187 d 21 h 17 min/ Padalka and Fincke	Third crew rotation flight by a Soyuz carried Expedition 9 cre (Padalka and Fincke) plus visiting cosmonaut/researcher Kuiper <i>TMA-4</i> docked to the <i>ISS</i> on Apr. 21, replacing the previous CF <i>Soyuz TMA-3</i> , which, when it undocked on Apr. 29, had reached a in-space time of 194 days. In it, Expedition 8 crewmembers Foa and Kaleri plus Kuipers landed in Kazakhstan.
Det. 13, 2004–Apr. 24, 2005	<i>Soyuz TMA-5</i> , 9S Leroy Chiao Salizhan Sharipov Yuri Shargin (Russia)	15,920 lb ~3031 rev 241 mi	9 d 21 h 30 min/ Shargin; 192 d 19 h 35 min/ Chiao and Sharpov	Fourth <i>ISS</i> crew rotation flight by a Soyuz carried Expedition 1 (Chiao and Sharipov), the fourth "caretaker" crew, plus visiting comonaut/researcher Shargin. On Oct. 16, <i>TMA-3</i> docked to the <i>IS</i> at the DC-1 docking compartment's nadir (downward)-pointing poi On Oct. 23, the previous CRV, <i>Soyuz TMA-4</i> , undocked from the <i>FG</i> nadir port, where it had stayed for 186 days (187 days in space), an landed in Kazakhstan with Padalka, Fincke, and Shargin.
Apr. 14–Oct. 10, 2005	<i>Soyuz TMA-6</i> , 10S Sergei Krikalev John Phillips Roberto Vittori (ESA, Italy)	15,920 lb ~2818 rev 235 mi	9 d 21 h 22 min/ Vittori; 179 d 23 min/ Krikalev and Phillips	Fifth crew rotation flight by a Soyuz carried Expedition 11 cre (Krikalev and Phillips) plus visiting crewmember Vittori. TMA- docked to the <i>ISS</i> on Apr. 16, replacing the previous CRV, Soyu <i>TMA-5</i> , which, when it undocked on Apr. 24, had reached an it space time of 192 days. In it, Chiao, Sharipov, and Vittori landed Kazakhstan.
July 26–Aug. 9, 2005	STS 114 Eileen M. Collins James M. Kelly Andrew S. W. Thomas Charles J. Camarda Wendy B. Lawrence Stephen K. Robinson Soichi Noguchi (Japan)	267,825 lb (orbiter at liftoff) 219 rev 218 mi	13 d 21 h 32 min 22 s	Return-to-flight mission of <i>Discovery</i> . Prior to final approach to th <i>ISS</i> , <i>Discovery</i> performed an R-bar or rendezvous pitch maneuw (RPM) at ~600 ft (180 m) distance under the <i>ISS</i> , a 360° backfl to allow digital imagery of its thermal protection system from th <i>ISS</i> . About 15,000 lb (6800 kg) of cargo was transferred from th shuttle's flight container, the MPLM <i>Raffaelo</i> , to the <i>ISS</i> and approximately 8600 lb (3900 kg) from the station to the shuttle for return to Earth. Crewmembers performed three EVAs, tested new safe equipment and repair procedures, and carried out a first-of-its kin spacewalking heat shield repair.
Sept. 30, 2005–Apr. 8, 2006	Soyuz TMA-7, 11S William McArthur Valery Tokarev Gregory Olsen (United States)	15,920 lb ~2992 rev 229 mi	9 d 21 h 14 min/ Olsen; 189 d 19 h 53 min/ McArthur, Tokarev	Sixth <i>ISS</i> crew rotation flight by a Soyuz carried Expedition 12 cre (McArthur, Tokarev) and the third "space tourist" (Olsen). <i>TMA</i> - docked to the <i>ISS</i> on Oct. 3. On Oct. 10, <i>Soyuz TMA</i> -6 undocke from the <i>FGB</i> nadir port, where it had stayed for 177 days, an landed in Kazakhstan with Krikalev, Phillips, and Olsen.
Oct. 12–17, 2005	<i>Shenzhou 6</i> Fei Junlong Nie Haisheng	17,000 lb 75 rev 209 mi	115 h 32 min	Second orbital flight of a crewed spacecraft on China's Long Marc 2F rocket, this time carrying two "taikonauts." Both crewmembe left their seats, floated in space, and at one time took off their spac suits.
Mar. 29–Sept. 28, 2006	Soyuz TMA-8, 12S Pavel V. Vinogradov Jeffrey Williams Marcos Cezar Pontes (Brazil)	15,920 mi ~2885 rev 229 mi	9 d 22 h 18 min./ Pontes; 182 d 22 h 43 min/ Vinogradov and Williams	Seventh <i>ISS</i> crew rotation flight by a Soyuz carried Expedition 1 crewmembers Vinogradov and Williams, and visiting crewmember Pontes. On Apr. 8 the previous CRV, <i>Soyuz TMA-7</i> , undocked fro the <i>ISS</i> and returned to Kazhakstan, carrying McArthur, Tokarev, ar Pontes.
July 4–17, 2006	STS 121 Steven Lindsey Mark Kelly Stephanie Wilson Michael Fossum Piers Sellers Lisa Nowak Thomas Reiter (ESA, Germany) [up]	266,962 lb (orbiter at liftoff) 202 rev 220 mi	12 d 18 h 37 min; 171 d 3 h 54 min/ Reiter	<i>Discovery</i> , on Mission ULF1.1 to the <i>ISS</i> , flew the first of tw test flights after return to flight. The joint crews transferred th fully loaded MPLM <i>Leonardo</i> , conducted three EVAs, reberthed th loaded MPLM in the shuttle cargo bay, performed external inspe- tions of the orbiter heat shielding, and completed cargo transfer Reiter remained on the <i>ISS</i> when <i>Discovery</i> returned to Earth, joinin Expedition 13.

TABLE 1. Crewed space flights 1961–2006 (cont.)					
Designation and crew	A. Weight B. Revolutions C. Max. distance from Earth	Duration	Remarks		
STS 115 Brent Jett Chris Ferguson Dan Burbank Heidemarie Stefanyshyn-Piper Joe Tanner Steve MacLean (Canada)	269, 840 lb (orbiter at liftoff) 187 rev 217 mi	11 d 19 h 6 min	Atlantis, on Mission 12A to the <i>ISS</i> , flew Shuttle Flight Test 2. Prior to final approach, <i>Atlantis</i> performed an RPM, similar to that of <i>STS 114</i> . Preliminary views of the thermal protection system did not indicate any signs of damage. Crews installed the P3/P4 truss structure with the second set of solar arrays (of four), deployed the two solar array wings and a huge radiator panel from the P4 truss, conducted three EVAs, and resupplied the station.		
<i>Soyuz TM</i> A-9, 13S Michael Lopez-Allegria Mikhail Tyurin Anousheh Ansari	15,920 lb — 216 mi	10 d 21 h 4 min 17 s/Ansari	This flight carried new Expedition 14 crewmembers Lopez-Alegria and Tyurin, replacing Expedition 13 crewmembers Vinogradov and Williams, plus Iranian-American "taxi cosmonaut" Ansari, the first female spaceflight participant to visit the space station, flying under contract with the Russian Federal Space Agency. Expedition 13 crewmember Reiter remained on the station until Dec. 20, joining Expedition 14. On Sept. 28, <i>Soyuz TMA-8</i> undocked from the <i>ISS</i> FGB nadir port and landed in Kazakhstan with Vinogradov, Williams, and Ansari. Return and landing of <i>Soyuz TMA-9</i> was scheduled for Mar. 19, 2007, which would give Lopez-Alegria and Tyurin 182 d in space.		
STS 116 Mark Polansky William (Bill) Oefelein Nicholas Patrick Robert (Bob) Curbeam Joan Higginbotham Christer Fuglesang (ESA, Sweden) Sunita Williams (up) Thomas Reiter (down)	265,466 lb (orbiter at liftoff) 202 rev 222 mi	12 d 20 h 45 min	Space. Discovery, on Mission 12A.1 to the <i>ISS</i> , delivered and installed the P5 truss segment (serving as spacer at the end of the P4 truss), reconfigured the electrical power and thermal control sys- tems to their final configurations and activated them, and transferred extra oxygen for EVA prebreathing and nitrogen to storage tanks on the outside of the airlock. Astronauts conducted four EVAs, in- cluding one that was not originally planned to support continuing attempts by ground controllers to retract the 4B solar array of the P6 truss. Williams replaced Reiter as a member of Expedition 14 on the <i>ISS</i> .		
	Strs 115 Brent Jett Chris Ferguson Dan Burbank Heidemarie Stefanyshyn-Piper Joe Tanner Steve MacLean (Canada) Soyuz TMA-9, 13S Michael Lopez-Allegria Mikhail Tyurin Anousheh Ansari STS 116 Mark Polansky William (Bill) Oefelein Nicholas Patrick Robert (Bob) Curbeam Joan Higginbotham Christer Fuglesang (ESA, Sweden) Sunita Williams (up) Thomas Reiter (down)	space flights 1961–2006 (cont.) A. Weight B. Revolutions C. Max. distance from Earth STS 115 Brent Jett Chris Ferguson Dan Burbank Heidemarie Stefanyshyn-Piper Joe Tanner Steve MacLean (Canada) 269, 840 lb (orbiter at liftoff) Soyuz TMA-9, 13S Michael Lopez-Allegria Mikhail Tyurin Anousheh Ansari 15,920 lb – STS 116 Mark Polansky William (Bill) Oefelein Nicholas Patrick Robert (Bob) Curbeam Joan Higginbotham Christer Fuglesang (ESA, Sweden) Sunita Williams (up) Thomas Reiter (down) 265,466 lb (orbiter at liftoff) 202 rev 222 mi	space flights 1961–2006 (cont.) Designation A. Weight B. Revolutions C. Max. distance and crew Common Earth Duration STS 115 269, 840 lb 11 d 19 h 6 min Brent Jett (orbiter at liftoff) 187 rev Dan Burbank 217 mi 167 rev Heidemarie Stefanyshyn-Piper 15,920 lb 10 d 21 h 4 min Soyuz 7MA-9, 13S 15,920 lb - 10 d 21 h 4 min Michael Lopez-Allegria - 216 mi 17 s/Ansari Mikhail Tyurin 216 mi 17 s/Ansari 17 s/Ansari STS 116 (orbiter at liftoff) 202 rev 222 mi 222 mi STS 116 (crobarding at the common at liftoff) 45 min Mikhail Tyurin 202 rev 222 mi 222 mi Sourt Williams (up) Thomas Reiter (down) 201 min 45 min		

reduce g loads and heat pulse, especially for higherenergy missions such as return from high altitudes and lunar or planetary trips, advanced reentry vehicles fly a lifting reentry. Some lift can even be obtained from a basically ballistic capsule. A basic feature of the Soviet Soyuz and the American Gemini and Apollo spacecraft, as well as the United States space shuttle and future crewed vehicles, is that the return trajectory is extended by applying to the vehicle an aerodynamic crosswind or lifting force L that is created by offsetting the center of gravity from the center of aerodynamic pressure of the spacecraft (located at the centerline for symmetrical bodies) in order to let the airflow hit the vehicle asymmetrically at an angle of attack. The direction and magnitude of this force is controlled by signals from an on-board guidance computer. Although often only a small fraction of the drag force D (L/D for Apollo-type bodies is typically $\frac{1}{4}$ to $\frac{1}{5}$, this lift is sufficient to modulate the trajectory for landing point control and reduction of heating (not total, or integrated, heat load) and deceleration load peaks (Fig. 1). See AEROTHERMODYNAMICS; ATMOSPHERIC ENTRY; SPACE SHUTTLE.

Subsystems. In addition to the thermal protection system and reentry and landing systems, basic features of crewed spacecraft include the following subsystems: (1) crew equipment (couches and restraints, operational aids, clothing, food and water, hygiene equipment, medical supplies, and survival equipment); (2) displays and controls; (3) electrical power system; (4) environmental control and life support system; (5) reaction control system (for attitude maneuvers); (6) on-board propulsion system (for velocity-change maneuvers); (7) telecommunications system (on-board voice and data/telemetry radio-frequency equipment and antennas); (8) guidance and navigation system, which is interrelated with (9) the stabilization and control system (inertial sensors, optical instruments, and computer); (10) mechanical systems (escape, docking, land-

		Apollo	Gem	ini
Thick stagr	ness at nation point	1.7 in. (4.3 cm	n) 1 in. (2.	5 cm)
Max. Btu J/(r	heat rate /(ft ²)(s) m ²)(s)	620 7.04×10 ⁶	120 1.36×) 10 ⁶
Heat Btu J/n	load /ft ² 1 ²	36,000 4.09×10 ⁸	12,1 1.37×	00 10 ⁸
400 300 200 100	 (2480°C) 	3000°F (1650°C) <i>Gemini</i>	Apollo / 2300°F (1260°C)	
0	400	800 120	00 1600	2000
		range, mm		

Fig. 1. Apollo and Gemini reentry conditions; 1 ft = 0.3 m; 1 nmi = 1.852 km. (NASA)

ing); and (11) extravehicular activity (EVA) systems, including space suits, life support umbilicals or backpacks, mobility aids, restraints, and handholds. *See* SPACE NAVIGATION AND GUIDANCE; SPACE TECHNOLOGY.

Environmental controls. The main system that distinguishes crewed spacecraft from other spacecraft is the environmental control and life support system. It provides the crew with a suitably controlled atmosphere. Because all atmospheric supplies must be carried into space, it is essential to recirculate and purify the spacecraft atmosphere to keep the total weight of the vehicle within reasonable limits. In future years, partially or completely closed life support systems will even recover oxygen and water from the atmosphere for reuse, reducing resupply requirements.

Soviet and, today, Russian spacecraft have used a nitrogen-oxygen environment at normal sea-level pressure. United States spacecraft have used pure oxygen at a pressure of 5 lb/in.² (34.5 kilopascals) during space flight. It provides the lightest, simplest system possible, and was adopted for this reason. Although some medical disadvantages were feared, they did not materialize. However, the technique requires complete control of oxidizable (inflammable) materials on board. For the extended mission duration of post-*Apollo*-type vehicles, such as *Skylab*, the space shuttle and the *International Space Station*, NASA has used a more Earth-like atmosphere. *See* SPACE BIOLOGY.

Supplies. All supplies for the entire flight must be carried in the spacecraft, unless special supply missions are to be flown, for example, to a station in Earth orbit. An Earth-like environment is required because of human dependence on certain gaseous constituents for breathing. There must be provision for controlling the metabolically produced carbon dioxide (CO_2) in the spacecraft atmosphere; controlling water vapor and maintaining a comfortable humidity; reducing odor and removing toxic products; maintaining a comfortable thermal environment, atmospheric pressure, sound-pressure level, and radiation level; meeting drinking water and nutritional requirements; disposing of waste; supporting the crew during emergencies; and remedying weightlessness.

Space suits. Space suits or pressure suits are mobile spacecraft or chambers that house the astronauts and protect them from the hostile environment of space. They provide atmosphere for breathing, pressurization, and thermal control; protect astronauts from heat, cold, glare, radiation, and micrometeorites; contain a communication link and hygiene equipment; and must have adequate mobility. The suit is worn during launch, docking, and other critical flight phases. During noncritical flight phases, the space suit is removed for greater comfort. This practice has also been adopted by Soviet cosmonauts.

The use of two gases at atmospheric pressure on board, while excellent for intravehicular activities (IVA), becomes somewhat disadvantageous when extravehicular activity is required, for two principal reasons. First, adequate mobility in a space suit pressurized to over 3-4 lb/in.² (21-28 kPa) has proved extremely difficult. Second, at lower suit pressures, suitable for mobility, only pure oxygen can be used for breathing. In this case, because of the lower gas density, cooling may have to be provided to the astronaut with water, circulating in tubes sewn into an undergarment.

Power. The various life-support, communications, and other systems in a spacecraft depend upon adequate power. For very short missions conventional battery supplies may be used, but for longer-duration missions battery supplies must be constantly recharged from some other source such as a solar or nuclear supply. Fuel cells, which generate electric power from the chemical combination of suitable reactants such as hydrogen and oxygen, have proved effective in Apollo, Skylab, and the shuttle. Solar power is particularly useful within the orbit of Mars. Nuclear power may be used anywhere in space, thus removing the requirement to stay close to the Sun, but requires special provision for shielding spacecraft occupants and equipment from the radiations. See FUEL CELL; NUCLEAR BATTERY; SOLAR POWER; SPACE POWER SYSTEMS.

Communications. Radio and television communications channels permit a two-way exchange of information and instructions that are important for both normal flight and emergencies. Through telemetering links, the operations center monitors the health and operation of the spacecraft and its crew, and receives various kinds of engineering, operational, and research data.

A space flight communications system requires suitable transmitters, receivers, display units, recorders, power supplies, and antennas. The antennas impose challenging design requirements. The ground antennas must be able to track the spacecraft, and should be large enough to minimize the size and complexity of the spacecraft antenna. For near-Earth orbits the spacecraft or its antenna must be sufficiently maneuverable to keep the antenna pointed at the receiving station on the ground; for deep-space missions rapid maneuverability is not required, but accurate pointing is necessary. *See* SPACE COMMUNI-CATIONS; SPACECRAFT GROUND INSTRUMENTATION.

Reliability. Because of the potential hazards of space flight to personnel, crewed spacecraft must meet stringent requirements of safety and hardware reliability. Reliability is associated with the probability that the spacecraft systems will operate properly for the required length of time and under the specified conditions. To assure survival and return in all foreseeable emergencies, the design of a crew-rated spacecraft includes standby systems (double or triple redundancy) and allows for launch escape, alternative and degraded modes of operation, contingency plans, emergency procedures, and abort trajectories to provide maximum probability of mission success. The priority order of this reliability requirement is (1) crew safety, (2) minimum achievable mission fulfillment, (3) mission data return, and (4) minimal degradation. See RELIABILITY (ENGINEERING).

Crew recovery. For launch escape, the Mercury used



Fig. 2. Mercury spacecraft, showing approximate dimensions. 1 m = 3.3 ft. (NASA)

a solid rocket mounted on a tower (Fig. 2) to provide the necessary propulsive force to escape from a malfunctioning launch vehicle. This device was jettisoned after the critical flight phase. The Gemini and early Soviet spacecraft had ejection seats with parachutes. These were used on some of the early Vostok flights to recover the crew on land, but in Gemini they were intended only for launch escape. Because a vehicle launched from the Russian sites passes over land during the extremely critical flight phase immediately after launch, the Russians provide for landing on land to effect a safe recovery from an abort in this region. It is also used as the prime recovery method after reentry. However, Mercury and Gemini were designed for launch over the Atlantic Ocean, eastward from Cape Canaveral, Florida, so that a water-landing capability was required. Also, because much of the orbital track was over water, a water-landing technique would be necessary to cope with the majority of possible random abort situations. The same logic applied to the lunar missions of Apollo. Although the United States technique of water recovery has been highly developed and has proven eminently satisfactory for non-reusable spacecraft, more advanced spacecraft, such as the space shuttle, are designed for landing on runways much like an airplane.

In-flight maintenance. The mission reliability of spacecraft can be increased by providing for inflight maintenance and repairs by the crew. The decision to use in-flight repair on crewed spacecraft, however, must take into consideration the effect on the overall probability of mission success and flight crew safety, and the associated weight penalties which result from this reliability improvement technique.

Emergencies. Both equipment and operational procedures are carefully planned and designed to minimize harmful results from emergencies. On the operational side, procedures are worked out in advance for a wide range of imaginable occurrences. On the

equipment side, ample use is made of redundancies, so that if one piece of equipment fails an alternate means of performing that equipment's functions is available. The value of redundancy was evident in the *Apollo 13* emergency. When an explosion occurred in one of the *Apollo* fuel cell oxygen tanks, substantially disabling the spacecraft, the ability of the attached lunar module to provide oxygen, power, orientation, and propulsion made it possible to return the astronauts safely to the ground after circumnavigating the Moon.

Preventive measures. As an example of the extreme complexity of crewed space flight, the Saturn 5 launch vehicle of the United States, with the Apollo spacecraft combined, had between 5 and 6 million functional parts, miles of wiring, and thousands of connections. To assure the functional reliability required for the lunar landing of humans, preventive measures had to include conservative design (with a wide margin of safety), stringent technical and administrative controls, and extensive premission simulation and testing, far beyond the required safety level, under various mission conditions. Without high-speed digital computers, a successful lunar landing program would not have been possible and the space shuttle could not fly. See HUMAN-MACHINE SYSTEMS; SERVOMECHANISM; SIMULATION.

Physiological stresses. Experience gained from both the United States and Russia's crewed space flight programs has provided valuable physiological data. *Apollo, Skylab, Salyut,* and *Mir* were especially fruitful. Voice, heart rate, respiratory rate, blood pressure, and body temperature are monitored during periods of dynamic stress (for example, launch, reentry, and extravehicular activity). Extensive ground studies and laboratory tests have been performed.

The stress imposed by long exposure to weightlessness is still a major question. Astronauts adapt fairly quickly to the sensation of weightlessness, but the complete role played by the inner ear system in this adaptation has not been determined.

Cardiovascular deconditioning was detected postflight during the Gemini program. Conditioning equipment was developed subsequently to be used in space to measure the amount of deconditioning and to act as a countermeasure to weightlessness. The most significant type of this equipment is a lower-body negative pressure device which exposes the lower half of the body to reduced pressure; engorgement of the veins in the legs and abdominal viscera results, thereby counteracting blood pooling in the chest cavity and increased pressure in the left atrium at 0 g. This effect is believed to prevent the loss in blood volume otherwise produced as engorgement of the chest veins stimulates diuresis.

Mineral balance is also markedly affected by weightlessness. Loss of the bone calcium secreted in urine and feces is observed, calling for countermeasures. Similar loss of nitrogen indicates loss of muscle tissue, which is shown clearly by circumferential measurement of the astronauts' legs. These effects are considered serious, and although *Skylab*, *Salyut*, and *Mir* have shown that humans can stay in

space for appreciable periods of time, it is the considered opinion of those experts who have reviewed the data that for crewed missions lasting much longer than a year the causes of mineral loss have to be better understood and methods developed to prevent and control these losses. Since similar mineral losses are observed during prolonged bed rest, bed-rest subjects are being used to study this problem further.

Biomedical monitoring throughout a space flight is considered important to the safety and effectiveness of the crew. Although life support and environmental parameters (for example, total pressure, oxygen partial pressure, temperature, and radiation level) can be expected to provide the earliest indication of non-nominal conditions, their significance and urgency for corrective action usually require correlation with biomedical data. The medical evaluation of potential microbial and chemical contamination aboard the spacecraft requires microbiological, pulmonary, and metabolic studies. Basic hematological and biochemical measurements are being obtained at frequent intervals during the initial portion of a space flight to obtain a database. *See* AEROSPACE MEDICINE.

Soviet/Russian Programs

In the former Soviet Union, ideas and concepts of crewed space flight date back at least to the last century, to K. E. Tsiolkovskiy, now considered one of the pioneers of space flight. After the first Sputnik launch on October 4, 1957, developments of crewed space flight capability followed in quick succession, leading from the first-generation Vostok (East) to the second-generation Voskbod (Ascent) and to the third-generation Soyuz (Union) spacecraft. Originally engaged in an aggressive program to land a cosmonaut on the Moon before the United States lunar orbit mission of Apollo 8 in December 1968, the Soviet Union redirected its aims, after four test failures of its N1 heavy-lift launcher, toward the development of permanent human presence in Earth orbit

Vostok. The *Vostok* (**Fig. 3**), a single-seater for short-duration missions and ballistic reentry from Earth orbits, consisted of a near-spherical cabin covered entirely with thermal protection system material, having three small viewports and external radio antennas. It contained a life-support system, radios, simple instrumentation, and an ejection seat. The 7-ft (2-m) sphere of the cabin was attached to a service module resembling two truncated cones base to base, with a ring of gas pressure bottles on the upper cone close to the cabin. The module contained chemical batteries, orientation rockets, the main retropropulsion system, and additional support equipment.

In all, six *Vostok* cosmonauts were launched from 1961 to 1963, including the first woman in space, V. V. Tereshkova (*Vostok 6*). The launch vehicle was a two-stage rocket with four additional strap-on booster units surrounding the first stage, forming a pyramidlike assemblage. All five propulsion units, feeding into 20 rocket nozzles, were started on the launch pad; later, the four strap-on units fell away,



Fig. 3. Vostok, a first-generation Soviet spacecraft. (NASA)

leaving the core unit to continue burning for a time. After its burnout, the single-engine second stage took over, inserting the payload into the target orbit. Payload capability to orbit was about 10,400 lb (4750 kg).

Voskhod. The second-generation Voskbod, essentially a greatly modified Vostok (for example, the heavy ejection-seat system was removed), was a short-duration multiperson craft and was designed to permit extravehicular activity or spacewalking by one of the crew. The first three-member crew flew on Voskhod 1, while on the double-seater Voskhod 2 the place of the third seat was used for an expandable airlock to permit A. A. Leonov on March 18, 1965, to egress into space without evacuating the main cabin of air, making him the first human to perform extravehicular activity. The two Voskhod spacecraft were launched in 1964-1965 by an improved version of the Vostok launch vehicle, fitted with a new upper stage, that had the capability to lift the heavier Voskhods and-upon further uprating-the subsequent Soyuz vehicles.

Soyuz. The *Soyuz* design is much heavier, larger, and more advanced in its orbital systems, permitting extended orbital stay times. It consists of three main sections or modules: a descent vehicle, an orbital module, and an instrument-assembly module (**Fig.** 4), with a total habitable volume of about 350 ft³ (10 m³). The three elements are joined together, with the descent vehicle in the middle. Shortly before atmospheric reentry, the two outer

modules are jettisoned. The spacecraft has a total length of 24.5 ft (7.5 m), and the span of its solar panel wings is 27.5 ft (8.4 m).

Descent vehicle. The descent vehicle is intended for crew location during launch, orbital flight, descent, and landing. Roughly bell-shaped, it is covered with an ablative thermal protection system and designed for lifting reentry, permitting a controlled descent with the blunt end, at an angle of attack, facing forward. There are two portholes on its side, and one opening with a protruding periscope sighting device for the navigation system. All the control organs and most of the displays are located in this module.

Orbital module. The forward orbital module, about 7 ft (2 m) in diameter, is used by the crew as scientific laboratory and living quarters for orbital work, rest, and recreation. Consisting of two hemispheres joined by a cylindrical insert, it contains a "sofa" and folding table, a "sideboard," a control panel with instrumentation and equipment of the main life support system, and scientific and photographic equipment. A docking system is located at the front end of the orbital module. Originally of the probeand-socket design, like the United States system on Apollo, for the American-Soviet Apollo-Soyuz Test Project (ASTP), this system was replaced by the peripheral "international" (androgynous) docking system designed by NASA and Soviet engineers.

Launch vehicle. The *Soyuz* launch vehicle (**Fig. 5**), capable of placing 15,000 lb (6800 kg) in orbit, has a total height of 161.7 ft (49.3 m), with 33.8 ft (10.3 m) maximum diameter. Its first stage, assisted by the four strap-on units, has a total vacuum thrust of 504 tons (4.47×10^6 newtons), generated by five engines with a total of 20 nozzles. In addition, there are 12 steering nozzles. The side units have a length of 62 ft (19 m), the central unit 92 ft (28 m). The second stage is 26 ft (8 m) in length and 8.5 ft (2.6 m) in diameter; its four-nozzle engine creates 30 tons (2.7×10^5 N) of thrust. Total lift-off weight of the *Soyuz* is about 300 metric tons. *See* ROCKET PROPULSION; ROCKET STAGING.

Flight history. Since its first, ill-fated flight in April 1967, resulting in the death of cosmonaut V. M.



Fig. 4. Soyuz, a third-generation Soviet spacecraft. 1 m = 3.3 ft.



Fig. 5. Soyuz launch vehicle SL-4. (NASA)

Komarov because of failure of the parachute landing system, the *Soyuz* spacecraft has served as the mainstay of the Soviet/Russian crewed space program in many functions, for example, as a development test bed of autonomous navigation, guidance, maneuvering, and rendezvous procedures; research laboratory for biomedical and other scientific and technical studies; and logistics vehicle for the *Salyut* space stations.

Soyuz spacecraft carried originally up to three cosmonauts. However, when an accidental explosive decompression of the descent cabin during reentry caused the death (by pulmonary embolisms) of the Soyuz 11 crew-G. T. Dobrovolskiy, V. N. Volkov, and V. I. Patsayev-on June 30, 1971, the wearing of pressure suits during launch, ascent to orbit, descent to Earth, and other critical maneuvers was reintroduced for safety. The third seat was removed to make room for the necessary additional "suit loop" of the life-support system and the emergency air supply system, and the Soyuz thereafter carried only two cosmonauts. An improved version, the Soyuz T (for "Transport") was introduced in 1980. A new version, the Soyuz-TM, with extended mission duration and new subsystems, replaced this vehicle in February 1987, after a crewless test flight in May 1986. (The third seat was reintroduced with Soyuz TM-24 in August 1996.) In October 2002, the Soyzuz TM was, in turn, replaced by the Soyuz TMA, with modifications financed by NASA for improved safety and widened crewmember size range.

By December 31, 2006, there had been 94 crewed *Soyuz* missions (up to and including *TMA-9*) with 229 cosmonauts (some of them on more than one flight, and including foreign guest cosmonauts) and two aborted launch attempts (*Soyuz 18A* and an attempt on September 26, 1983). Thirty-two of these flights docked with, and transferred crews to, *Salyut* space stations, 49 with the *Mir* space station, and 13 with the *International Space Station* (*ISS*). The first

mission to the *ISS* carried the first resident crew and the second carried the first "space tourist," Dennis Tito. The *Soyuz TMA* craft have served as lifeboats, providing assured crew return capability to the *ISS*, and from April 2003 through September 2006 they shouldered the burden of providing crew rotation for the station, after the loss of the shuttle *Columbia* temporarily brought shuttle operations to a standstill.

Salyut. These orbital stations were 20 m (65 ft) long and 4 m (13 ft) wide, and weighed 19 tons. Between 1971 and 1981, five of them operated in low Earth orbit successively for some time. By the end of that period, the most successful of them, *Salyut 6*, had seen seven host (long-stay) crews and nine visiting crews, including eight non-Soviet cosmonauts (from Czechoslovakia, Poland, East Germany, Hungary, Vietnam, Bulgaria, Cuba, Mongolia, and Romania). Also during that period, *Salyut 6* was resupplied by 12 *Progress* crewless automatic cargo carriers.

Salyut 6 was the first of a "second generation" of *Salyut* space stations. Unlike its predecessors, it had two docking ports, instead of only one. This enabled it to receive visiting crews and resupply ships. Together with other new on-board systems, this feature was the key to missions of considerably extended duration.

Its successor, Salyut 7, was launched on April 19, 1982. By the end of 1985, Salyut 7 had seen five longstay host crews and four visiting crews, which included the first cosmonauts from India and France as well as two flights of Svetlana Savitskaya, the second woman in space and the first female spacewalker. After serious systems problems arose in 1985, the crew of Soyuz T13 performed extensive repair and replacement work on the station and saved it from becoming inoperative. With the flights of Soyuz T13 and T14, Salyut 7 saw the first successful space station crew rotation, although the mission of the T14 crew had to be abruptly terminated because of illness of one of the cosmonauts. The crew of Soyuz T15 conducted the first transfer between two space stations when they visited Salyut 7 in May and June 1986 from the new permanent orbital facility, Mir.

Mir. On February 19, 1986, the Soviet Union launched the core vehicle in its new Mir (Peace) space station complex series. This complex represented a new-generation space station which evolved from the Salyut and Kosmos series vehicles. The core, an advanced version of Salyut, had six docking ports, a large number of portholes, and was 43 ft (13.13 m) long and 13.6 ft (4.15 m) wide, with a mass of 21 metric tons (23 tons). It consisted of four sections: transfer section [8.2 ft (2.2 m) diameter], work section [10.3 ft (3.15 m)], intermediate section, and equipment section [both 13.6 ft (4.15 m)], the latter containing two rocket engines of 66 lbf (294 N) thrust each. Its two winglike solar panels were larger than those of *Salyut* 7: 818 ft² (76 m²) versus 441 ft² (41 m²), with a wingspan of 97.5 ft (29.73 m).

The core could be expanded by individual laboratory modules launched separately and connected to its various docking ports. These modules were usually dedicated to different scientific and technical disciplines or functions. These included technological production with a shop, astrophysics, biological research, and medical. Launch of the first block, the service module *Kvant 2*, took place on November 26, 1989. The second block, *Kristall*, launched on May 31, 1990, became the third major addition to *Mir*. Nearly identical to *Kvant 2*, this building block was equipped with furnaces for materials experiments in space and a docking port for the originally planned *Buran* shuttle, later abandoned. The third module, *Spektr*, was launched on May 20, 1995; and the fourth module, *Priroda*, was launched on April 23, 1996.

After launch, *Mir*'s laboratory modules were docked initially to the axial position on the multiport docking fixture on the forward end of the station. They then were transferred by means of their own articulated robot arm to one of the four outward-facing ports on the docking fixture of the core module, forming a spokelike array around the core.

The station complex had about 23 kW power, generated by solar-cell arrays, and about 10,600 ft³ (300 m³) of pressurized (habitable) volume. Maximum crew size, for relatively short durations, was six persons, limited by the number of triple-crewed *Soyuz-TM* crew transport-escape vehicles that could be docked to the station. With the launch of *Mir*, the Soviet Union had had eight different stations in Earth orbit, including one failed attempt (*Salyut 2* in April 1973).

The first Mir crew, Leonid Kizim and Vladimir Solovyev, was launched on board Soyuz T15 on March 13, 1986. Docking and transfer to the station took place about 49 h later on March 15. More crews followed in succession, keeping the station continuously occupied until April 1989, when the crew of Soyuz TM-7 left Mir for a crewless period until the arrival of TM-8 in September 1989. By the time Mir was decommissioned and deorbited on March 23, 2001, having been in operation for 15 years, the station had been visited by 39 crews, totaling 88 cosmonauts, including seven times by the space shuttle Atlantis and once each by Endeavour and Discovery. During this time, Mir occupants set new marks three times for the longest human stay in space: 326 days by Yuri Romanenko in 1987, followed by 366 days by Vladimir Titov and Musa Manarov in 1988, and 437 days by Valery Polyakov in 1944-1995. In its later years, Mir required increasing maintenance and repairs by its crews, particularly after two serious emergencies in 1997: an on-board fire, caused by a backup oxygen generator device, and a collision of the automated Progress M-34 freighter with Mir's Spektr module during a docking exercise. When U.S. support of Mir ended at the conclusion of the joint Shuttle/Mir program (ISS Phase 1), further crewed operations of the space station were increasingly difficult and impractical for the Russian Space Agency. The last-best-one crew-cosmonauts Sergei Avdeev and Viktor Afanasyev, and French guest cosmonaut Jean-Pierre Haigneré-returned to Earth

on August 27, 1999, in Soyuz TM-29.

Most of *Mir*'s on-board systems were turned off on September 8. The station entered, for the first time in 10 years, a period of crewless dormancy, and preparations began for the possibility of a final decision by its Russian owners to terminate it with a controlled reentry into the Earth's atmosphere in early spring 2000. *Mir* gained another few months of life with the arrival of the final crew, Sergei Zalyotin and Alexander Kaleri in *Soyuz TM-30*, who prepared it for a flight extension. However, a deal sought by a commercial firm, MirCorp, fell through, and *Mir*'s end came on March 23, 2001, when it was deorbited and splashed into the South Pacific Ocean. *See* SPACE STATION.

Energia. With the successful first test flight, from Baikonur, of the powerful expendable heavy-lift launcher Energia on May 15, 1987, even if its payload did not reach orbit, the Soviet Union gained a tremendous new capability in launching large and heavy satellites, space station components, interplanetary spacecraft, and even the new Soviet space shuttle. The launcher has two stages, in parallel arrangement, and the payload is mounted not on top but on the side of the second core stage. Four liquid-fueled boosters (kerosine and liquid oxygen) make up the first stage. They are "strapped" to the cryogenic core stage, which is powered by four engines of 148 metric tons (1.45 meganewtons) thrust each (liquid hydrogen and liquid oxygen). With a total thrust at liftoff of about 3000 metric tons (30 MN) and a liftoff mass of 2650 tons (2400 metric tons), a height of only 197 ft (60 m), and a largest diameter of 65.6 ft (20 m), the Energia can carry 110 tons (100 metric tons) to Earth orbit or inject 20 tons (18 metric tons) to geostationary orbit, 35 tons (32 metric tons) to the Moon, and up to 31 tons (28 metric tons) to Venus and Mars.

With its second, and last, flight on November 15, 1988, the *Energia* launched the Soviet space shuttle *Buran*.

Buran. Described as the first of a fleet of Soviet shuttles, the space glider Buran (Blizzard) accomplished its first orbital test flight on November 15, 1988. The test mission was flown automated, that is, without crew, and lasted 3 h 25 min. The flight, while concluding with a return to the intended landing site, was only partially successful. It did not lead to further development efforts of this vehicle. The delta-winged Buran appeared to copy the major features of the United States shuttle, without, however, its own main propulsion system (which was provided by the expendable Energia core stage). The main mission of the Buran program reportedly was to support the Soviet space station Mir and its successors. After the collapse of the Soviet Union, the Energia/Buran programs were terminated.

European Program

European efforts in space have long had the basic goal of achieving European space autonomy: the capability of Europe to develop, launch, and operate its own satellites and space vehicles independent of the space systems of other nations, including the United States. This goal has been accomplished for only crewless launches of commercial and scientific satellites. In order to conduct human space activities, European astronauts have made use of the United States space shuttle and the Soviet/Russian *Soyuz* launcher.

European Space Agency (ESA). The multinational European space program is the responsibility of the European Space Agency, representing 13 member nations (Austria, Belgium, Denmark, France, Germany, Ireland, Italy, Netherlands, Norway, Spain, Sweden, Switzerland, and the United Kingdom, with Finland as associate member and Canada a cooperative nation). The ESA conducts operations at field installations in the Netherlands, Germany, and Italy; tracking stations in a number of countries; and its launch center at Kourou, French Guyana.

Ariane 5. The growing weight of geostationary satellites and the economic advantage of launching two satellites on one carrier led the European Space Agency to develop a more powerful version of the successful commercial expendable launcher Ariane. Ariane 5 is a $2^{1}/_{2}$ -stage launch vehicle, with a core stage (propelled by a cryogenic engine), two reusable solid boosters, similar to those of the United States space shuttle, and an upper stage. Developed by the French space agency [Centre Nationale d'Etudes Spatiales (CNES)] for the European Space Agency, Ariane 5 is able to launch payloads of up to 23 tons (21 metric tons) to low Earth orbit, and is also used for launching geostationary satellites.

The ECA (enhanced capability) version of the Ariane 5, designed to lift 10 tons (9 metric tons)

to geostationary transfer orbit, enough for two big communications satellites at once, uses a new cryogenic upper stage, an improved Vulcain 2 main stage engine, and solid boosters loaded with 10% more propellant. By December 31, 2006, the Ariane 5-G (generic) had flown 24 flights and the Ariane 5-ECA had flown 7 (after initial failures of both vehicles, in June 1996 and December 2002, respectively).

United States Programs

During the first decade after its inception in 1961, the United States crewed space program was conducted in three major phases—Mercury, Gemini, and Apollo. Each crewed flight led to increased knowledge of the systems and techniques needed to operate successfully in space, and each phase represented a significant advancement.

Development. Many differences in the three crewed United States spacecraft are readily apparent, such as size and weight (Fig. 6). The major differences, however, are in the complexity and refinement of subsystems. Apollo's requirement for hardware "maturity" was significantly higher than for the earlier programs. Each subsystem became progressively more complex, with many more demands made upon it and a correspondingly greater capability. For example, only Apollo had its own guidance and navigation system. Another good example of increased system complexity was electrical power. For the Mercury craft, it was supplied by six batteries; Gemini required seven batteries and two chemical fuel-cell power plants. For Apollo, electrical power was supplied by five batteries and three fuel cells.



Fig. 6. Comparison of United States spacecraft and launch vehicle configurations through 1972. (NASA)

Physically, the three systems do not appear greatly different, but in terms of operational demand they differed considerably. The *Mercury* electrical system had to supply power to sustain the 3000-lb (1360-kg) spacecraft and its single astronaut for $1^{1}/_{2}$ days. In *Gemini*, sufficient power had to be provided to operate a 7000-lb (3175-kg) craft containing two astronauts for as long as 2 weeks, while the *Apollo* system was designed to support a 100,000-lb (45-metric-ton) spacecraft carrying three persons for up to 2 weeks.

Perhaps the greatest difference between *Apollo* and the earlier spacecraft was in reliability. Everything had to work and keep working no matter what the circumstances. Unlike the previous crewed space programs, in which the crew could return to Earth almost within minutes if an emergency arose, the *Apollo* crew needed as much as 6 days to get back from a position outbound toward the Moon from the Earth's vicinity, requiring lunar circumnavigation to return safely to Earth. Such a situation occurred in the flight of *Apollo 13* in 1970 (Table 1).

Mercury. After Alan Shepard's first ride on *Mercury 3* atop a Redstone rocket, John Glenn in 1962 became the first American to orbit the Earth. His vehicle, a *Mercury*, was in its basic characteristics similar to the Soviet *Vostok*, but it weighed only about a third as much, as necessitated by the smaller missiles of the United States at that time (the Redstone and Atlas).

The one-person *Mercury* capsules (Fig. 3) used ballistic reentry and were designed to answer the basic questions about humans in space: how they were affected by weightlessness, how they withstood the gravitational forces of boost and entry, how well they could perform in space. The *Mercury* flights proved that humans not only could survive in space, but could also greatly increase humanity's knowledge of the universe. Between 1961 and 1963, a total of six astronauts flew on *Mercury* spacecraft, the last flight (Gordon Cooper's *Faith* 7) lasting a record 34 h. *See* WEIGHTLESSNESS.

Gemini. The second United States step into space was the Gemini Program. With the two-person Gemini capsule (Fig. 7), for the first time a crewed spacecraft had been given operational maneuverability in space. Its reentry module flew a lifting reentry trajectory for precise landing point control (Fig. 1). In addition, its design permitted extravehicular activity by one of the crew. The goal of the Gemini missions was to find out how humans could maneuver themselves and their craft, and to increase knowledge about such things as celestial mechanics, space navigation, and rendezvous and docking, as required later for the Apollo Program. All these extra capabilities required that the weight of the spacecraft be increased by 130% over the Mercury. This in turn called for a more powerful launch vehicle, the Titan 2, which had its origin as an Air Force ICBM.

Gemini has a record of 10 successful crewed flights between 1965 and 1966, with 20 astronauts and a total mission duration of 969 h 52 min in space.



Fig. 7. Gemini spacecraft. (NASA)

It set many records, including the longest duration (almost 14 days, *Gemini 7*), the first rendezvous by two maneuverable spacecraft (*Gemini 6* and 7), and the first docking (*Gemini 8*). Moreover, on June 3, 1965, *Gemini 4* crew member Edward White performed the first spacewalk by an American astronaut, drifting and maneuvering in space for 21 min at the end of a 25-ft (7.5-m) umbilical/tether. On later *Gemini* missions, EVAs were performed by E. A. Cernan, M. Collins, R. F. Gordon, and E. E. Aldrin, totaling 13 h 23 min, including open-hatch or stand-up EVAs (SEVAs).

Apollo. The third-generation spacecraft, *Apollo*, was 82 ft (25 m) tall and had five distinct parts: the command module (CM), the service module (SM), the lunar module (LM), the launch escape system (LES), and the spacecraft/lunar module adapter (SLA). The three modules made up the basic spacecraft; the LES and SLA were jettisoned early in the mission after they had outlived their function.

Command module. The command module served as the control center for the spacecraft and provided 210 ft³ (6 m³) of living and working quarters for the three-member crew for the entire flight, except for the period when two persons entered the lunar module for the descent to the Moon and return. The command module was the only part of the spacecraft that returned to Earth, flying a lifting trajectory with computer-steered maneuvers (Fig. 1).

Resembling a cone of almost 10-ft (3-m) height and 12.7-ft (3.9-m) width, the command module consisted of two shells separated by a layer of insulation: an inner crew compartment (pressure vessel) and an outer heat shield.

Service module. Attached to the command module at its heat shield end was the service module, a 22.8-ft-tall (7-m), 12.7-ft-wide (3.9-cm) cylinder housing the electrical power subsystem, most of the reaction control engines, part of the environmental control subsystem, and the service propulsion subsystem, including the main rocket engine for insertion into orbit around the Moon, for return from the Moon, and for course corrections.

Lunar module. The lunar module (**Fig. 8**) carried two astronauts from the orbiting command/service mod-



Fig. 8. Apollo lunar module. (NASA)

ule (CSM) down to the surface of the Moon, provided a base of operations there, and returned the two astronauts to a rendezvous with the command service module in orbit. It consisted of two components: the ascent stage (on top) and the descent stage. The latter had a descent engine and propellant tanks, landing gear assembly, a section for scientific equipment for use on the Moon, another section to transport a collapsed wheeled vehicle on the last three Apollo flights (the lunar rover), and extra oxygen, water, and helium tanks. The ascent stage contained the crew compartment with about 160 ft³ (4.6 m^3) of habitable volume, the ascent engine with propellant tanks, and all lunar module controls. Its subsystems were essentially the same kinds as those of the command module and service module. The lunar module measured about 23 ft (7 m) in height (legs extended), weighed about 32,500 lb (14.6 metric tons), and was constructed of aluminum with titanium fittings.

Flight history. Including Earth-orbital flights by *Apollo* 7 and 9, a total of 11 crewed missions were launched in the Apollo Program between 1968 and 1972 (Table 1). Of these, nine crews (27 astronauts) went to the Moon, with six landings (**Fig. 9**). Twelve astronauts walked on the Moon, accumulating almost 300 h of lunar stay time and 166 person-

hours of surface exploration. They traversed almost 60 mi (97 km) and brought about 850 lb (385 kg) of rock and soil samples back to Earth. Approximately 33,000 lunar photographs and 20,000 reels of magnetic tapes of geophysical data were collected, causing a major revolution in planetary science by providing the first deep understanding of a celestial body other than Earth. *See* MOON.

Lunar roving vehicle. On the last three lunar landings, Apollo 15-17, lunar exploration was supported by the lunar roving vehicle (LRV; Fig. 10). This vehicle weighed 460 lb (209 kg) on Earth (77 lb or 35 kg Moon weight) and 1600 lb (725 kg) fully loaded. Measuring 122 in. (3 m) in length and 6 ft (2 m) in width, it could carry two astronauts in their space suits and backpack life support units, scientific equipment, and lunar samples. Each wheel was driven by its own separate traction drive assembly consisting of a $\frac{1}{4}$ -horsepower (0.2-kW) dc brush-type electric motor, harmonic drive reduction gear, and brake assembly. Power was provided by two nonrechargeable 36-V silver-zinc batteries, each containing 23 cells, having a capacity of 115 ampere-hours and weighing 59 lb (27 kg). Both rear and front wheels were steered by a hand-operated control stick through two separate electric motors on the axles. Turning radius was 122 in. (3.1 m). The lunar roving



Fig. 9. Astronaut Edwin Aldrin, Jr., stands next to the passive seismometer experiment on the surface of the Moon. On July 20, 1969, Aldrin and Neil Armstrong became the first humans to land on the Moon. A television camera, American flag, laser reflector, and lunar module Eagle are behind Aldrin. (NASA)

vehicle had a range of 36 mi (58 km), traveled at a maximum speed of 8.5 mi/h (13 km/h), and negotiated slopes as steep as 25. During its transport to the Moon, it was stowed in folded and collapsed condition in the lunar module descent stage in a volume measuring $4.4 \times 4.4 \times 5.4$ ft ($1.3 \times 1.3 \times 1.6$ m). After touchdown, semiautomatic deployment was accomplished by the astronauts' pulling nylon tapes and releasing cables. By using the lunar roving vehicle, the astronauts were able to spend more time on the Moon's surface during exploration.

Extravehicular activity. Extravehicular activities became of paramount importance in Apollo. In addition to the 12 astronauts who walked on the Moon, four crew members performed in-flight EVAs, with a total of 3 h 46 min. Counting the nine space exits by *Gemini* astronauts, a total of 41 individual EVAs had been accumulated by the end of the Apollo Program. To these, *Skylab* astronauts later added 19 more (10 occasions, **Table 2**).

Launch vehicle. Prior to Apollo, the United States' launch vehicle inventory included the Atlas-Agena B with 5000 lb (2250 kg) Earth-orbital payload capability, and the Titan 2 with 7800 lb (3500 kg). Clearly, a more powerful booster was required to lift the *Apollo* spacecraft to Earth orbit and thence to the Moon.

At the Army Ballistic Missile Agency (ABMA) in 1958, a team of engineers under Wernher von Braun set out to prove that vastly more powerful space



Fig. 10. Lunar roving vehicle used in the Apollo 15 mission. (NASA)

Crewed periods	First	Second	Third	Total
Launch	5/25/73	7/28/73	11/16/73	
	9:00 a.m. EDT	7:10 a.m. EDT	9:01 a.m. EDT	
Splashdown	6/22/73	9/25/73	2/8/74	
	9:49 a.m. EDT	6:19 p.m. EDT	11:17 a.m. EDT	
Duration (day:h:min)	28:0:49	59:01:9	84:01:16	171:13:14
Revolutions	404	858	1214	2476
Distance, 10 ⁶ mi (10 ⁶ km)	11.5 (18.5)	24.5 (39.4)	34.5 (55.5)	70.5 (113.5)
SEVA*	0:37 (5/25/73)			,
EVA [†] 1 duration (h:min)	3:30 (6/7/73)	6:29 (8/6/73)	6:33 (11/22/73)	
EVA 2 duration	1:44 (6/19/73)	4:30 (8/24/73)	7:01 (12/25/73)	
EVA 3 duration		2:45 (9/22/73)	3:28 (12/29/73)	
EVA 4 duration			5:19 (2/3/74)	
Total EVA	5:51	13:44	22:21	41:56
Solar observatory photos	30,242	77,600	75,000	182,842
Earth resources photos	8,886	14,400	17,000	40,286

rockets could be built from existing hardware by clustering engines and tanks. The project evolved into the Saturn Program of NASA. There were three types of launch vehicles in the Saturn family: the Saturn 1 which had a perfect record of ten successful flights, the Saturn 1B with eight successful flights, and the Saturn 5, which became the Apollo lunar launch vehicle and had the capability to lift 250,000 lb (113 metric tons) into low Earth orbit and to send 100,000 lb (45 metric tons) to the Moon. The development of the Saturn 5 proceeded through three phases, with the two preceding launch vehicles as stepping stones. Saturn 1 used primarily modified existing equipment. Saturn 1B used a modified first stage of the Saturn 1 with uprated engines and a new second stage and instrument unit. Saturn 5 introduced new first and second stages, departing from the clustering approach, and used as its third stage and instrument unit the Saturn 1B's second stage and instrument unit (Table 3).

TABLE 3. Saturn launch vehicles*							
Launch vehicle and stage	Length, ft	Weight, 10 ³ lb	Payload to low orbit, lb	Diameter, ft	Engines, number and type	Propellants, weight and, type [†]	Stage thrust, [‡] 10 ³ lb
Saturn 1, block I (with payload)	162.5	1120 (gross)	No orbital capability				
1st stage, S-1	81.6	103 (dry)	_	21.4	8 H-1's	875,000 lb, RP-1/LOX	1504 s.l.
2d stage, dummy Saturn 1, block II (with payload)	44 187.5	142 (ballast) 1128 (gross)	24,000	18.3			
1st stage, S-1	80.3	105.7 (dry)		21.4	8 H-1's	875,000 lb, RP-1/LOX	1504 s.l.
2d stage, S-4	41.4	13.9 (dry)		18.3	6 RL10A-3's	100,000 lb, LH ₂ /LOX	90 v.
Instrument unit Saturn 1B	2.8 224	3.1 1313 (gross)	40,000	21.4			
1st stage, S-1B	80.3	84.4 (dry)		21.4	8 H-1's	885,000 lb, RP-1/LOX	1650 s.l.
2d stage, S-4B	58.4	25 (dry)		21.7	1 J-2	230,000 lb, LH ₂ /LOX	200 v.
Instrument unit Saturn 5	3 363	4.3 6200 (gross)	270,000	21.7			
1st stage, S-1C	138	303 (dry)		33	5 F-1's	4,578,000 lb, RP-1/LOX	7600 s.l.
2d stage, S-2	81.5	95 (dry)		33	5 J-2's	942,000 lb, LH ₂ /LOX	1160 v.
3d stage, S-4B	58.13	33.6 (dry)		21.7	1 J-2	232,000 lb, LH ₂ /LOX	230 v.
Instrument unit	3	4.3		21.7			

T tt = 0.3 m. 1 lb = 0.45 kg. 1 lbt = 4.4 newtons. $^{+}$ RP-1 = type of kerosine, LOX = liquid oxygen (-297°F, -183°C), LH₂ = liquid hydrogen (-422.9°F, -252.8°C).

[‡]s.l. = thrust at sea level, v. = thrust in vacuum.



Fig. 11. Space station *Skylab* in orbit. The sunshade rigged by crew repair is visible. (*NASA*)

In addition to its lunar mission in the Apollo program, the Saturn 5, in a two-stage version, served also—in its last and thirteenth flight—to launch the first United States space station, *Skylab*.

Skylab

The experimental space station Skylab (Fig. 11) was the largest object ever placed in space, and the first crewed project in the U.S. Space Program with the specific purpose of developing the utility of space flight in order to expand and enhance humanity's well-being on Earth. To that end, Skylab's equipment included Earth resources remote sensing instrument and the first crewed solar telescopes in space. A total of three crews of three astronauts each carried out experiments and observations on Skylab in a number of different areas, among them observations of the Earth, solar astronomy, stellar astronomy, space physics, geophysics, biomedical studies, zero-gravity, biological studies, zero-gravity technology studies, and spacecraft environment. Originally begun in the early 1960s, under the names Apollo Extension Support Study and Apollo Applications Program, Skylab emerged as a cluster of five major components: orbital workshop (OWS), airlock module (AM), multiple docking adapter (MDA), Apollo telescope mount (ATM), and command/service module (CSM). The CSM, adapted from the Apollo Program, served as crew carrier, launched by Saturn 1B's.

Skylab was launched crewless by Saturn 5 on May

14, 1973 (*Skylab 1*), and suffered a structural failure of its combined meteoroid/thermal shield during the first minute of flight which eliminated most of its electrical power supply. Subsequently visited by three astronaut crews for periods of 28, 59, and 84 days (Tables 1 and 2), the space station underwent repair in orbit, a first for the space program, which proved that humans are eminently capable of performing heavy emergency repair, service, and maintenance operations in space, thereby setting the stage for the future of orbital operations by space shuttle crews.

After the third crew had returned to Earth on February 8, 1974, the United States first space station remained dormant for 5 years while NASA worked on plans to save *Skylab* from orbital decay and reentry for future missions. The intention was given up in December 1978 when it became clear that the estimated decay rate of the orbit was approaching faster than anticipated and the likely date of space shuttle rescue mission was moving ahead. Under active control by NASA, *Skylab* finally returned to Earth on July 11, 1979, impacting in the Indian Ocean, with a few pieces landing in Australia.

Apollo-Soyuz Test Project

The last flight of an *Apollo* spacecraft occurred as the United States part in the world's first international crewed space flight, the Apollo-Soyuz Test Project on July 15–24, 1975. During that time, three American astronauts and two Soviet cosmonauts linked up in space for a total joint period of 47 h 17 min (**Fig. 12**).

The ASTP mission was planned to accomplish spacecraft rendezvous, docking, undocking, and crew transfer between spacecraft of disparate technical and management philosophies, the Soviet *Soyuz* and the United States *Apollo*, as well as interaction of spacecrafts crews. The successfully conducted mission tested compatible rendezvous and docking systems being developed for future United States and Soviet crewed spacecraft and space stations; in addition, the crews practiced docking and undocking, conducted intravehicular crew transfer, and performed 31 scientific experiments and engineering investigations.

Docking. The new compatible rendezvous and docking system provided the basis for a standardized international system for docking of crewed space-craft. Thus, it enhanced the safety of crewed flights in space and provided the opportunity for conducting subsequent joint missions and experiments.

To enable *Apollo* and *Soyuz* to dock, NASA developed and constructed a third element, the docking module (DM), that also served both crews as an airlock and transfer tunnel between the spacecraft (**Fig. 13**). It was 10 ft (3.2 m) long and 4.6 ft (1.4 m) wide, and weighed 4660 lb (2097 kg). It held two suited crew members and was attached to the forward end of *Apollo*. On the *Soyuz* side, it used the compatible docking system designed by NASA and Soviet engineers (**Fig. 14**). This docking system was of the androgynous type, which uses the principle of



Fig. 12. An artist's concept of the first international "handshake" in space, by the American and Soviet crews of the Apollo-Soyuz Test Project mission, July 17, 1975. (NASA)



Fig. 13. Apollo and Soyuz spacecraft in docked configuration during the Apollo-Soyuz Test Project mission with docking module serving as connecting link and airlock. 1 m = 3.3 ft. (NASA)



Fig. 14. Universal docking system, as used on Apollo-Soyuz Test Project (American version), suspended for testing. (NASA)

reverse symmetry to allow the mechanical linkage of otherwise dissimilar spaceships. Docking is usually performed by one (active) spacecraft on the other (passive) partner.

Since the *Soyuz* internal atmosphere consisted of nitrogen-oxygen at sea-level pressure as opposed to *Apollo's* pure oxygen environment at 5 lb/in.² (34.5 kPa), the docking module was required to serve as an air lock. To reduce lengthy waiting periods for denitrogenation, the *Soyuz* pressure was reduced to about 10 lb/in.² (69 kPa) before docking. During the mission itself, four transfers between spacecraft were accomplished by crew members of both nations.

Spacecraft modifications. To fit the mission needs, several modifications were made to both spacecraft. For Apollo, they included an increased number of propellant tanks for the reaction control system, added equipment to operate the docking module and the American-Soviet rendezvous and docking systems, and provisions for scientific and technical experiments. Among the major modifications of Soyuz were the new docking mechanism, additional communications equipment to accommodate the United States ultrahigh frequency of 296 MHz, a transponder (a combined receiver-transmitter that beams a signal when triggered by another radio signal) to allow Apollo to calculate distance during rendezvous, and alignment devices to aid Apollo in performing its active docking role.

Accomplishments. In addition to the rendezvous and docking test, the ASTP demonstrated that precision launch of complex vehicles from widely separated sites to meet the trajectory alignment requirements for a joint space mission was feasible, and that hardware, operating, and language differences can be surmounted for joint international space efforts, with enough planning and training in advance.

Space Transportation System

After the end of the joint American-Soviet space mission, the United States crewed space flight program had reached the end of a pioneering era of capability development. Attention began to focus on the routine application of newly acquired know-how systems and experience, specifically in the form of the emerging Space Transportation System (STS), with the ability to transport inexpensively a variety of useful payloads to orbit, as the mainstay and "work horse" of the United States space program in the 1980s and 1990s. Its design is aimed at reducing the cost and increasing the effectiveness of using space for commercial, scientific, and defense needs.

The two major components of the Space Transportation System are the space shuttle and the *Spacelab*. This system is complemented by a number of upper stages depending on the mission beyond Earth orbit, for example, the Payload Assist Module (PAM), the Transfer Orbit Stage (TOS), and the Inertial Upper Stage.

Space shuttle. The space shuttle (**Fig. 15**), as a reusable system, has the capability to conduct space missions in response to national and worldwide needs and the flexibility to respond to policy,



Fig. 15. Space shuttle on its crawler/transporter/launch platform. (NASA)



Fig. 16. Space shuttle vehicle. 1 m = 3.3 ft. (NASA)

discovery, innovation, and newly emerging needs. The shuttle provides extended capability in the areas of (1) easy access to and from low Earth orbit for nonastronaut passengers, (2) multipurpose reusable transportation, (3) standard equipment for a variety of spacecraft, (4) standard equipment for a variety of missions, (5) modularized subsystems for easy refurbishment, (6) refurbishable spacecraft/payloads in orbit, and (7) a technology base for more flexible future missions. The space shuttle flight system (Fig. 16) has three major elements: the orbiter, an external tank (ET) containing the liquid propellants to be used by the orbiter main engines for ascent, and two solid-propellant rocket boosters (SRBs) "strapped on" to the external tank. The orbiter and solid rocket booster casings are reusable; the external tank is expended on each launch. The total weight of the system at launch is 4,406,000 lb (1,998,500 kg).

Orbiter and external tank. The orbiter contains the crew and payloads. In size and weight (150,000 lb or 68,000 kg when empty of consumables) comparable to a DC-9 jet transport, it delivers payloads of 65,000 lb (29,500 kg) with lengths to 80 ft (18 m) and widths to 15 ft (4.5 m). The crew compartment accommodates seven crew members and passengers for some missions, but as many as ten persons in emergency operations.

The three main propulsion rocket engines in the aft fuselage of the orbiter, with a vacuum thrust of 470,000 lbf (2,090,000 N) each, use liquid hydrogen and liquid oxygen from the external tank, which is jettisoned shortly before orbit insertion after approximately 8 min of burn time. The orbiter also carries 46 auxiliary engines with a total vacuum thrust of 45,200 lbf (201,000 N). The external tank has a length of 153.7 ft (46.85 m) and a diameter of 331.0 in. (8.41 m). Its weight is 1,619,000 lb (734,400 kg) at launch and 69,000 lb (31,300 kg) when empty of fuel.

Solid rocket boosters. The two solid rocket boosters burn in parallel with the main engines of the orbiter to provide initial ascent thrust. Primary elements of each solid rocket booster are the motor, including casing, propellant, igniter and nozzle, forward and aft structures, separation and recovery systems, and thrust vector control subsystems. Each solid rocket booster has a length of 149.1 ft (45.45 m), a diameter of 146 in. (3.71 m), a weight of 1,265,000 lb (573,800 kg) at launch and 182,000 lb (82,600 kg) after the fuel has been expended, and a thrust of 2,700,000 lbf (12,000,000 N) at launch. The solid propellant grain is shaped to reduce thrust some time after liftoff to prevent overstressing the vehicle during the period of maximum dynamic pressure. The solid rocket boosters separate after about 2 min of flight from the external tank at an altitude of 30 mi (50 km) by pyrotechnic devices and eight small rockets; they descend on a system of parachutes and splash in the ocean about 170 mi (274 km) from the launch site-Kennedy Space Center in Florida or Vandenberg Air Force Base in California-to be retrieved for reuse.

Testing and flight. The first shuttle orbiter, named *Enterprise*, was rolled out at Palmdale, California, on September 17, 1976. Used for ground tests and approach and landing flights tests during 1977 and 1978, the *Enterprise* was superseded by the first space mission orbiter, *Columbia*, for the first orbital flight on April 12-14, 1981, piloted by astronauts John Young and Robert Crippen (**Fig. 17**). Three additional shuttles—*Challenger, Discovery*, and *Atlantis*—were added subsequently, so that by the end of 1985 the total shuttle fleet consisted of four vehicles. Planning for the construction of a possible fifth orbiter was held in abeyance pending a better understanding of future launch-frequency requirements.

By mid-January 1986, 24 shuttle missions to low Earth orbits has been conducted (up to and including *61-C*), all of them successfully, although a number of problems had to be overcome in the course of the program (Table 1). Highlights of these flights included the retrieval of two errant communications satellites (on *STS 51-A*), the in-space repair of two satellites (on *STS 41-C* and *STS 51-I*), the first untethered extravehicular activity of a human with an autonomous maneuvering system (on *STS 41-B*; **Fig. 18**), assembly of two large space structures (*STS 61-B*), and three flights of the space laboratory *Spacelab* (*STS 9, STS 51-B*, and *STS 61-A*).

With the launch of the twenty-fifth shuttle mission, *51-L*, on January 28, 1986, however, tragedy struck the American space program. At approximately 11:40 a.m. EST, 73 s after liftoff, the flight of *Challenger*, on its tenth mission, abruptly ended in an explosion triggered by a leak in the right solid rocket booster, killing the seven-member crew.

The loss of *Challenger* reduced NASA's shuttle fleet by one-fourth. An independent commission to investigate the causes of the *51-L* accident was established. Continuation of the United States crewed space program was suspended pending a thorough reassessment of flight safety issues and implementation of necessary improvements. The recovery period lasted 32 months (compared to 18



Fig. 17. Launch of the space shuttle *Columbia* on its first orbital flight, April 12, 1981. (*NASA*)

months after the Apollo 1 fire in 1967 which killed three astronauts). During this time, all shuttle systems, operations, and its entire management were subjected to meticulous reviews and modifications, overviewed by outside expert committees. The solid booster joints, whose inherent design weaknesses had caused the explosion of the Challenger, were redesigned with 73 changes and tested thoroughly. The orbiter itself, in the process, underwent 228 modifications, not counting 100 software changes. The main propulsion system (63 changes) and even the external tank were improved, along with the assembly and launch facilities, as well as all preflight procedures. In July 1987, NASA ordered a new orbiter as Challenger's replacement, called Endeavour.

Shuttle flight operations resumed on September 29, 1988, with the long-awaited launch of *Discovery* on the mission *STS 26* to place a *Tracking and Data Relay Satellite (TDRS)* in orbit. Subsequent flights followed with increasing confidence in the new systems and operations. At the end of 1989, after seven highly successful missions since the *Challenger* accident, NASA was ready to resume full-scale space shuttle operations, signaling a total recovery from the accident.

Highlights of post-Challenger missions included the launch of the interplanetary research probes Magellan to Venus (on STS 30) and Galileo to Jupiter [on STS 34), the retrieval of the 10.5-metric-ton (11.5-ton) research satellite Long Duration Exposure Facility (LDEF) [on STS 32], the placement in orbit of the Hubble Space Telescope (on STS 31), the launch of the European solar probe Ulysses (on STS 41), and the intricate repair and maintenance missions to the Hubble telescope (on STS 61, 82, 103, and 109). On STS 60, in February 1994 the first Russian cosmonaut, Sergei Krikalev, participated as a crew member, and in March 1995 Endeavour, on mission STS 67, established a long-duration record for the space shuttle program of 16 days 15 h 9 min. On STS 71, in June 1995, 20 years after the Apollo-Soyuz Test Project, Atlantis carried out the first of a series of dockings with the Russian Mir space station (Fig. 19), forming the largest human-made structure in space so far. Joint activities were carried out and crew members exchanged between the two spacecraft before separation. Eight more dockings between space shuttles and Mir were conducted in 1995-1998 as Phase 1 of the ISS program, with seven U.S. astronauts stationed for extended durations on the Russian facility. Since 2000, most of the shuttle missions have been in support of the assembly and operation of the International Space Station.

On February 1, 2003, after completing a successful research mission (*STS 107*), *Columbia* was lost with its crew during reentry. The Columbia Accident Investigation Board, chaired by Admiral (ret.) Harold W. Gehman, Jr., concluded that, unbeknown to crew



Fig. 18. Astronaut Bruce McCandless performing first free (untethered) extravehicular activity with an autonomous maneuvering system during shuttle flight *STS 41-B* on February 7, 1984.



Fig. 19. Space shuttle Atlantis docked to the Kristall module on the Russian space station Mir, photographed by cosmonant Nikelai Budarin on July 4, 1995. (NASA)

and ground, one of the left wing's leading-edge reinforced carbon-carbon (RCC) elements had been punctured during ascent to orbit by a "suitcase-sized" chunk of foam insulation, blown off the external tank by the supersonic airstream, hitting the leading edge and rendering the wing unable to withstand reentry longer than about 8 min after entry interface. Shuttle operations were halted as NASA and its contractors labored on intensive return-to-flight (RTF) efforts.

The space shuttle returned to the skies on July 26, 2005, with the liftoff of *Discovery* on mission *STS 114* to the *ISS*. The flight was successful but engineers were concerned about the shedding of some insulating foam material off the external tank. Prior to its final approach to the station, *Discovery* performed a maneuver to allow digital imagery of its

thermal protection system from the *ISS*, and during docking crewmembers performed a first-of-its-kind spacewalking heat shield repair. With more work required on redesign of the foam-based external tank insulation, new sensors for detailed damage inspection, and a boom to allow astronauts to inspect the vehicle externally during flight, the second RTF test flight was delayed for nearly a year. However, with test flights in July and September 2006 and an additional flight in December, the shuttle fleet resumed its operations.

As of December 31, 2006, the four U. S. shuttles, in over 24 years of operation, have flown a total of 117 missions, carrying 597 men and 101 women into space (including the lost crews of seven on *STS 51-L/Challenger* and *STS 107/Columbia*). Participating

in these flights were 81 crew members from 16 foreign nations. *See* SPACE SHUTTLE.

Spacelab. The Spacelab was a major adjunct of the shuttle. Developed and funded by member nations of the European Space Agency, the large pressurized Spacelab module with an external equipment pallet was designed to be the most important payload carrier during the space shuttle era. With its large transport capacity (in weight, volume, and power supply), the Spacelab was intended to support a wide spectrum of missions in science, applications, and technology by providing versatile and economical laboratory and observation facilities in space for many users. Another objective was to reduce significantly the time from experiment concept to experiment result, compared with previous space practice, and also to reduce the cost of space experimentation. It allowed the direct participation of qualified men and women scientists and engineers to operate their own equipment in orbit. Primary uses of the Spacelab were identified in the areas of process engineering and materials research, biomedical research, Earth observations and meteorology, communication and navigation, space flight technology, astrophysics, astronomy, solar physics, energy research, atmospheric and space physics, and basic chemical and physical sciences.

The Spacelab design consisted of three elements: (1) a pressurized module which could be assembled from modular sections in two sizes-a short version for missions requiring only minimal supplemental subsystems in a "shirt-sleeve" environment, and a longer version for more extensive experiments; (2) a pallet or instrument mounting platform to support sensors and antennas requiring vacuum and uninterrupted space exposure; and (3) support equipment such as air locks, viewports, storage vaults, and pointing systems. Depending on assembled size, the Spacelab was designed to have a total pressurized volume of 880-3500 ft3 (25-100 m3). In some missions, a pallet-only mode could be flown. Spacelab's average power use was designed to be in the 3-6-kW range. Orbiting the Earth at 115-445 mi (185-555 km) altitude and 28-104° inclination on missions lasting 7-30 days, the Spacelab was designed to remain firmly attached inside the open shuttle cargo bay. A connecting tunnel was designed to provide access to and from the shuttle orbiter crew compartment. The laboratory was designed to provide working space for up to four payload specialists, not counting the commander, pilot, and mission specialist (flight engineer) or the space shuttle itself.

The first flight of the European-built *Spacelab*, *SL-1*, took place November 28–December 8, 1983, on board the *Columbia* as the ninth shuttle mission. Its six-member crew included the first foreign astronaut in the United States space program, the German mission specialist Ulf Merbold. *SL-1* and subsequent *Spacelab* flights, notably the first foreign chartered mission, the German *D-1*, but also the Japanese *SL-1* on *STS 47* and the second German *Spacelab* mission, *D-2* on *STS 55*, proved the concept of the crewed space laboratory and confirmed the unique worth

of humans in space to operate, repair, maintain, and troubleshoot scientific equipment and conduct complex research programs. *Spacelab* flew 16 missions before it was decommissioned in 1998. *See* SPACE BIOLOGY; SPACE PROCESSING.

International Space Station

Interest in the development of a permanent crewed platform in Earth orbit dates back to the very beginnings of human space flight. While the practical realization of the concept was accomplished and proven by the Soviet Salyut/Mir and the American Skylab programs, the real breakthrough happened on January 29, 1998, when representatives of 16 nations (United States, Russia, Canada, Japan, the 11 nations of the European Space Agency (ESA), and Brazil) signed a partnership agreement for the joint development and operation of an International Space Station (ISS). With so many nations sharing in this unique megaprogram, the ISS has become a focal point of international understanding, cooperation, and collaboration, and thereby a strong catalyst for peace enhancement and conflict reduction.

The goal of the ISS program is to establish a permanent platform for humans to live and work in space in support of science and technology research, business, education, and exploration. Its objectives are to provide a world-class laboratory complex uniquely located in a microgravity and vacuum environment where long-term scientific research can be carried out to help fight diseases on Earth, unmask fundamental processes leading to new manufacturing processes and products to benefit life on Earth, and observe and understand the Earth's environment and the universe. The station is expected to create new opportunities for the private sector for business development and entrepreneurial ventures. As a "virtual classroom" in space, the ISS will inspire and support educators and students worldwide. As a testbed for studies of long-term effects of space on humans and for the development of new space-flight technologies, it will also become an essential gateway to future human space exploration.

When completed, the *International Space Station* will have a mass of about 1,040,000 lb (470 metric tons). It will be 356 ft (109 m) across and 290 ft (88 m) long, with almost an acre (0.4 hectare) of solar panels to provide up to 110 kilowatts of power to six state-of-the-art laboratories.

1998. The on-orbit assembly of the *ISS* began with the launch of the U.S.-owned, Russian-built Functional Cargo Block (FGB) *Zarya* ("Dawn") on November 20, 1998, from the Baikonur Cosmodrome in Kazakhstan on a Proton rocket. The second element, the U.S.-built Node (connecting module) Unity followed on the space shuttle *Endeavour*, launched on December 4, 1998, from the Kennedy Space Center, Florida. The shuttle crew attached Unity and *Zarya* during a 12-day mission, beginning the station's orbital construction which will extend over 12 years until its completion in 2010.

1999. Accomplishments during 1999 included the first crewed logistics-supply flight of a space shuttle to the early assembly in orbit in May and June; testing

of hardware and software for station elements to be launched in 2000 and 2001; successful implementation of a newly developed Multi-Element Integrated Test (MEIT), which functionally connects the first several U.S. elements of the station on the ground in a series of integrated tests; the completion of the Mission Control Center and the Payload Operations Integration Center; and delivery of the U.S. laboratory module Destiny and the airlock module to the launch site in Florida in preparation for MEIT, along with other components. Interface testing between the U.S. Mission Control Center in Houston and the Russian Mission Control Center in Moscow was successfully completed.

2000. Among the key events in 2000 was the July 12 launch of the service module *Zvezda* (Star), the first Russian-funded and owned element of the station, which linked up with the station on July 25. The docking, executed very gently at 0.6 ft/s (0.2 m/s) by *Zarya* as the active partner, marked the arrival of the first human living quarters and an initial command post for the station. *Zvezda* is also equipped with the propulsion system that will maintain the attitude and orbit of the space station. The station now spanned 119 ft (36 m) in length, and its mass had grown to almost 60 tons (54 metric tons). On August 8 the automated *Progress MI-3* docked with the station, carrying fuel for *Zvezda*'s propellant tanks, as well as equipment and supplies.

On October 31, a Soyuz-U carrier lifted off from Baikonur and placed in orbit *Soyuz TM-31*, carrying the first resident crew for the space station. The Expedition One crew consisted of American Commander William Shepherd and veteran Russian cosmonauts Yuri Gidzenko and Sergei Krikalev. On November 2, the *Soyuz* craft docked with the station and the crew entered, to begin what is intended to be a new era of permanent human habitation in space. The Expedition One crew remained on the station for 136 days, until March 18, 2001.

An *Endeavour* assembly mission on November 30-December 11 installed two solar array wings, jointly spanning 240 ft (73 m), at the space station (**Fig. 20**). The array is now providing the station with up to 62 kW of power.

2001. *ISS* assembly continued in 2001, notably with the installation of the United States Laboratory module *Destiny* on February 10. The 28×14 ft (8.5 × 4.3 m) laboratory was transported on *Atlantis*. The Canada-built Space Station Remote Manipulator System (SSRMS) and the U.S. airlock module *Quest* also arrived on space shuttles, and the Russian Docking Compartment (DC-1) *Pirs* was carried on a Russian Soyuz-U rocket, *Quest* provided station residents with their own "door" to space for extravehicular activities (EVAs), and DC-1 gave the station a second airlock for spacewalks in Russian Orlan-M spacesuits.



Fig. 20. International Space Station, photographed from the space shuttle Discovery on December 19, 2006. (NASA)

In the first crew rotation, the Expedition Two crew, consisting of Russian commander Yuri Ussachev and American flight engineers James Voss and Susank Helms, replaced Expedition One in March. In August, Expedition Two was replaced by the Expedition Three crew, consisting of American commander Frank Culbertson and Russian cosmonauts Vladimir Dezhurov and Mikhail Tyurin. The crews in both rotations were transported on the space shuttle *Discovery*. Successive expedition crews have served on the *ISS* during stays of four to six months.

2002. In April 2002, the first of eleven truss elements, S0 (Starboard Zero), was attached on top of *Destiny*, becoming the centerpiece of the 109-mlong (356-ft) truss for carrying the solar cell arrays of the station. In June, a new station crew (Expedition 5) delivered the Mobile Base System (MBS), to provide mobility for the SSRMS. The second truss segment, S1, arrived in October and was attached to S0 on the starboard side. Its counterpart on port, P1, followed in November and was also successfully mounted.

2003–2005. Early in 2003, further progress in ISS assembly was brought to a halt by the standdown of the space shuttles after the Columbia loss. Crew rotation flights could be accomplished only by Russian Soyuz vehicles. As an immediate consequence of the unavoidable reduction in resupply missions to the station, which now could be supported only by Russian crewless automated Progress cargo ships, station crew size was reduced from three to a twoperson "caretaker" crew per expedition, except for brief stays by visiting researchers or commercial "tourists" arriving and departing on the third seat of Soyuz spacecraft. Operations procedures had to be revised accordingly, and such vital areas as onboard systems maintenance and spares provision had to be replanned carefully to continue crewed occupancy and a viable science research program on board despite the sudden constriction in logistics.

2006. With the return to flight of the shuttle, *ISS* assembly resumed. In July the shuttle carried a third astronaut (Thomas Reiter) to the station to join Expedition 13. In September, the P3/P4 truss structure was installed with a second set of solar arrays (of four), and the two solar array wings and a huge radiator panel were deployed from the P4 truss. In December, the P5 truss segment, serving as a spacer at the end of the P4 truss, was installed, and the electrical power and thermal control systems were reconfigured to their final configurations and activated.

As currently planned, the station will require a total of 80 assembly flights. Interspersed among these missions will be logistics missions by the shuttle, Russian *Soyuz* flights to launch crews, and multiple *Progress* tanker flights for refueling and supplying the growing structure in orbit.

Vision for Space Exploration

On January 14, 2004, U.S. President Bush announced the Vision for Space Exploration, a long-term plan

for returning astronauts to the Moon to prepare for voyages to Mars and other planets in the solar system. The plan called for completion of the *International Space Station* and retirement of the space shuttle by 2010, and the development of new spacecraft and launch systems.

Project Constellation. To support this initiative, NASA is developing a family of spacecraft, launchers, and associated hardware, collectively known as Project Constellation.

The Orion spacecraft will be shaped like an Apollo capsule but will be significantly larger. It will have a crew module shaped similarly to the Apollo command module, capable of holding four to six crew members, and a cylindrical service module, containing the spacecraft's propulsion systems and supplies. It will be reusable for up to 10 flights, allowing the construction of a fleet of such spacecraft. Plans call for it to be launched into low-earth orbit, either for solo flights or for missions to the *International Space Station*, using the proposed Ares I rocket, which would consist of a single solid rocket booster derived from those used in the space shuttle system, and a liquid-fueled second stage.

Lunar flights would involve the use of a larger launch vehicle, the proposed Ares V, which would use five RS-68 engines (more powerful and less expensive than the space shuttle main engines) and a pair of five-segment solid rocket boosters. (Alternative plans have also been considered which would replace the Ares I and V rockets with a single launcher, the Ares IV.) The Ares V would carry a combination of the Earth Departure Stage (EDS), an enlarged version of the S-IVB upper stage of the Saturn V rocket, and the Lunar Surface Access Module (LSAM). The Ares V would fly for about 8 minutes before separating from the EDS/LSAM combination, which the EDS would then place into low-Earth orbit. The crewed Orion spacecraft would be launched separately on the Ares I at least two to four weeks later, and dock with the EDS/LSAM combination. After the systems were configured for lunar flight, the EDS would inject the combination into a translunar orbit and then be jettisoned.

Like the Apollo lunar module, the LSAM would consist of an ascent stage, which would house the four-person crew, and a descent stage, which would have the landing legs, most of the crew's oxygen and water, and scientific equipment. Unlike the Apollo lunar module, the LSAM's engines would place the Orion/LSAM combination in lunar orbit. The entire crew would transfer to the LSAM, which would undock from Orion and descend to the lunar surface. After completing operations there, the crew would enter the LSAM's ascent stage and lift off, powered by a single engine and using the descent stage as a launch pad. After docking with and transferring to the Orion spacecraft in lunar orbit, the crew would jettison the LSAM and fire the Orion's single engine for the return journey to Earth.

Lunar base. In December 2006, NASA announced plans to eventually construct a solar-powered lunar base located near one of the poles of the Moon,

rather than landing at several sites, as was done in the Apollo missions. An incremental buildup is envisioned, with astronauts making several sevenday visits to the Moon until their power supplies, rovers, and living quarters are operational. These visits would be followed by 180-day missions to prepare for journeys to Mars. Robotic precursor missions would provide landing site reconnaissance, natural resource assays, and risk reduction for the human landers. A decision on a site will probably not be made until after collecting data from several such missions. The Lunar Reconnaissance Orbiter (LRO) is scheduled to launch in 2008 to scout the lunar surface and search for resources. The Lunar Crater Observation and Sensing Satellite is to launch with the LRO and then crash into the lunar surface to search for water ice.

Jesco von Puttkamer; Jonathan Weil Bibliography. W. Ley, Rockets, Missiles, Space Travel, 1969; NASA, Report on the 90-Day Study on Human Exploration of the Moon and Mars, 1989; Office of Exploration, NASA Headquarters, Beyond Earth's Boundaries: Human Exploration of the Solar System in the 21st Century, 1989; S. Ride, Leadership and America's Future in Space, Report to the Administrator, NASA Headquarters, 1987; Round Trip to Orbit: Human Spaceflight Alternatives, Special Report, OTA-ISC-419, U.S. Congress, 1989; W. Sweetman (ed.), Jane's Space Directory 2006-2007, 22d ed., 2006; Synthesis Group, America at the Threshold: America's Space Exploration Initiative, 1991; T. D. Thompson (ed.), TRW Space Log, TRW Space and Technology Group, annually; W. von Braun et al., Space Travel: A History, rev. ed., 1985.

Space navigation and guidance

The determination of the position and velocity of a space probe with respect to a target body or a known reference body such as the Earth (navigation) and, based upon this determination, the application of propulsive maneuvers to alter the subsequent path of the probe (guidance).

Navigation. If the position and velocity of a space probe, plus some secondary quantities such as the positions and masses of any perturbing celestial bodies, can be determined at a given instant, then the future path of that vehicle can be predicted for all time by the application of the laws of celestial mechanics. A simple way of understanding this is that, an instant later, the new position of the space probe will be the original position incremented by the velocity times the time duration. Similarly, the new velocity will be the original velocity incremented by the acceleration acting on the probe times the time duration. Additionally, the acceleration at the new time will be slightly different since gravitational accelerations acting on the probe depend on the position. Other accelerations may also affect the probe, such as thrust due to engine firings, solar radiation pressure, drag, and so forth. Thus, at the future time, the probe will arrive having a new position and a new velocity; these coordinates can be used as the basis for predicting an additional short interval into the future. By taking sufficiently short steps, a very accurate path of the future trajectory can be calculated. *See* CELESTIAL MECHANICS.

Space navigation can be viewed as determining the current position and velocity of the probe and then using that determination as the basis of predicting future motion. This determination is made by taking a series of measurements relating to the probe's location and motion and combining these measurements in such a manner as to make the most accurate estimate of the probe's current position and velocity, taking into account possible small errors or inaccuracies in the measurements themselves. One of the most powerful (and accurate) measurements which can be made is the relative velocity between an Earth tracking station and the space probe itself. This is accomplished by broadcasting to the probe an electromagnetic signal which consists of a single tone having a stable frequency. The probe will receive this signal shifted in frequency in exact proportion to the relative station-probe velocity and then immediately will rebroadcast it back to the Earth. This frequency shift is known as the Doppler effect. It is also possible to measure the stationprobe distance (or range) using electromagnetic signals. Modern systems can make these measurements to an accuracy of about 0.02 mm/s (0.001 in./s) for the velocity and 1 m (3 ft) for the range at distances extending to the outer limits of the solar system. An additional, very powerful measurement is known as delta-differential-one-way-range (denoted Δ -DOR or delta-DOR). It is performed by simultaneously observing both a space probe and a nearby extragalactic radio source (quasar) from two tracking stations. These four simultaneous measurements are differenced to form an extremely accurate estimate of the angular distance between the probe and quasar as seen in the sky. The determination of the position and velocity of the probe relative to the Earth and hence the Sun can be made by observing these quantities over a period of time. See DOPPLER EFFECT.

The probe's orbit relative to the Earth, thus determined, is most useful when the target body's position relative to the Earth is also accurately known. This is true for lunar missions and missions to the planets. For missions to the satellites of the outer planets (Jupiter and beyond) or Pluto, for a mission to a body such as a comet or asteroid, the position of the target may be sufficiently uncertain as to make a strictly Earth-relative navigation scheme inadequate. Here it is necessary to make measurements that directly involve the target. Typical of these is to obtain optical measurements of the locations of the target relative to a star as seen from the probe itself using an onboard camera. Again, a series of measurements of the angular distance between the target and a known star as the probe approaches can be used to determine the probe's position and velocity, this time relative to the target.

Guidance. In its simplest form, guidance consists of comparing the predicted future motion of the probe against that which is desired, and if these are sufficiently different, executing a propulsive maneuver to modify that future position. Such maneuvers are called trajectory correction maneuvers (TCMs). Typically, the probe will contain a number of small rocket motors which can be fired in any desired direction by first rotating the spacecraft away from its cruise orientation and then holding the new attitude fixed while the rocket motor is firing. The cruise attitude is achieved by orienting the spacecraft relative to celestial references. Typically, the roll axis of the probe is oriented toward the Sun. A roll of the probe until a predetermined star enters the field of view of a fixed star sensor is then sufficient to fix the probe's complete attitude in space. The reorienting of the probe for the desired rocket firing is accomplished by gyroscopes within the probe which measure and command the desired turns from the celestial reference orientation. An alternative, for both probes cruising in a fixed orientation as well as probes that are designed to spin about a primary axis, is to fire multiple thrusters oriented in different directions either in sequence or simultaneously such that the combination of the several maneuvers adds up to the desired modification to the trajectory. See STAR TRACKER.

Modern missions generally employ sequences of these basic navigation and guidance functions. For example, for a typical trajectory to Mars, the navigation plan will call for a maneuver (TCM-1) shortly after launch (usually about 10 days) to correct for most of the errors due to the launch vehicle. A month or two later, TCM-2 is performed to correct for most of the errors from TCM-1. As the probe approaches Mars, one to three or even four more TCM's will be performed depending on a variety of requirements related to the specific mission. For each TCM, Earth-based tracking (Doppler, ranging, and usually Δ -DOR) will be used to estimate the trajectory prior to performing the maneuver. Missions to bodies requiring optical navigation will also include this data for the final approach maneuvers.

An entirely different method of making very large changes to the orbit of a space probe utilizes the gravity of intermediate bodies rather than the engines on the probe. These multiplanet missions require not only complex trajectory design techniques but also very accurate navigation to the intermediate targets. Such missions may gather scientific information at both the intermediate bodies and the ultimate target or may simply use the intermediate bodies as an extremely efficient means of getting to the final target. At its simplest, the method of "gravity-assist" uses a close flyby of a massive body to exchange energy and momentum between the space probe and the body. Although large changes to the orbit of the probe are made, no discernible effect is observed on the massive body due to the enormous difference in mass between the probe and flyby body. Illustration a shows that in the frame of reference of the flyby body (Jupiter in this case) the speed of the probe both approaching and departing Jupiter is the same although the direction changes. In illus. b it is seen that since Jupiter is moving in the Sun's frame of reference both the speed and direction of the probe are changed. In the 1970s the first missions to use gravity-assist were Mariner Venus/Mercury and both the Pioneer and Voyager missions to the outer solar system. Since that time Galileo (launched in 1989) and Cassini-Huygens (launched in 1997) have used gravity-assist both for the interplanetary transfer and in orbit at Jupiter and Saturn, respectively. Missions have been launched to the extremes of the Solar System (MESSENGER in 2004 to Mercury orbit using Earth, Venus, and Mercury flybys, and New Horizons in 2006 to Pluto using a Jupiter flyby), and the Stardust mission utilized an Earth flyby on its way to comet Wild 2 and returned a sample of comet dust to the Earth. Other than further missions to Mars, almost all of the future exploration of the solar system may make use of gravity-assist. All of these orbit designs



Jupiter flyby of a space probe in (a) Jupiter's frame of reference and (b) the Sun's frame of reference.

include a series of trajectory correction maneuvers on approach to each flyby body as well as one or more "clean-up" trajectory correction maneuvers shortly after the flyby.

The guidance maneuvers discussed so far are designed to make small corrections to the planned orbit. For missions designed to orbit another body, a large terminal guidance maneuver, also called an orbit insertion maneuver (for example, MOI for Mars orbit insertion, or JOI for Jupiter orbit insertion), is applied at the encounter with the target body. These maneuvers are quite large and may use the large majority of the propellant carried by the probe. For example, the Galileo JOI was about 0.65 km/s (0.4 mi/s), while MOIs for Mars Odyssey (2001), Mars Express (2003), and Mars Reconnaissance Orbiter (2006) were about 1.4 km/s (0.9 mi/s), 0.8 km/s (0.5 mi/s), and 1.0 km/s (0.6 mi/s), respectively. For orbiter missions, navigation and guidance takes place much as for interplanetary cruise with both orbit determination and correction maneuvers being performed on a regular basis. See GUIDANCE SYSTEMS; SPACE PROBE.

Nomenclature. Often "navigation" is used to mean all of orbit design, orbit determination, and maneuver design and execution. The terminology "guidance and control (G&C)" or "guidance, navigation, and control (GN&C)" is then used to denote the operation of the space probe itself—knowledge and control of its pointing and orientation in space plus the timing and firing of the onboard engines. Dennis V. Byrnes; David W. Curkendall

Bibliography. R. H. Battin, *An Introduction to the Mathematics and Methods of Astrodynamics*, 1987, paper 1999; M. Kayton (ed.), *Navigation: Land, Sea, Air and Space*, 1990; C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*, rev. ed., 1995; W. G. Melbourne, Navigation between planets, *Sci. Amer.*, 234(6):58-74, June 1976; W. E. Wiesel, *Spaceflight Dynamics*, 2d ed., 1996.

Space power systems

On-board assemblages of equipment to generate, store, and distribute electrical energy on satellites and spacecraft. A reliable source of electrical power is required for supplying energy to the spacecraft and its payloads during launch and through several years of operational lifetime in a space environment. Present-generation spacecraft fly power systems from tens of watts to several kilowatts. Each of the three fuel cells on the space shuttle delivers 12 kW continuous power and 16 kW peak power. The *International Space Station*'s solar arrays will generate 110 kW total power, with approximately 46 kW available for research activities. *See* SATELLITE (SPACE CRAFT); SPACE FLIGHT; SPACE PROBE; SPACE SHUTTLE; SPACE STATION.

With few exceptions, the power systems for United States satellites have used photovoltaic generation for power, batteries for energy storage, and a host of electrical equipment for appropriate regulation, conversion, and distribution. Fuel cells have been limited primarily to use in the crewed space program, supplying power for *Gemini, Apollo*, and the space shuttle. Radioisotope thermoelectric generators (RTGs) have powered, or augmented solar power on, many planetary missions and probes, powered lunar instruments left on the Moon by the *Apollo* missions, and augmented the solar-array battery power system on at least one Earth-orbiting spacecraft.

Many factors influence the final configuration of a power system. Basic to the initial consideration are the nature of the mission (Earth-orbiting or planetary) and mission lifetime. Other relevant factors include (1) spacecraft and payload requirements with consideration to average and peak power loads; (2) effect of environment such as orbit period, length of time the spacecraft is in sunlight and shadow, radiation particle damage, and space charging; and (3) constraints imposed by the spacecraft such as weight, volume, spacecraft shape and appendages, satellite attitude control, electromagnetic radiation limitations, characteristics of payloads, and thermal dissipation. For spacecraft that are shuttle-launched, additional considerations include compatibility with shuttle payload safety requirements and, in some cases, the retrievability by the shuttle for return to Earth. The weight of the power system ranges from 15 to 25% of the spacecraft weight.

Solar cells. The photovoltaic effect was discovered in 1877, but was not exploited until the silicon cell was invented in 1954. The first cells had conversion efficiencies of approximately 6%, and were little more than laboratory curiosities until the space program began in 1958.

The first solar cells flown, pn cells (a thin p layer diffused into an n-type substrate), were selected because of the lower sensitivity of the material. However, these cells degraded in a radiation environment much more rapidly than the np cells (a thin n-layer diffused into a p-type substrate). Since the early 1960s, the np cell has been used almost exclusively for United States spacecraft.

Development efforts for solar cells have emphasized improved conversion (sunlight to electrical) efficiency, increased resistance to charged-particle radiation, reduced manufacturing costs, and reduced weight. Conversion efficiency obtained in the conventional solar cells was approximately 11% until 1972, when the violet cell was perfected with an efficiency of over 13%.

The following additions have pushed silicon solarcell efficiency up to about 14%: a p^+ layer (called a black-surface field) that reduces hole-electron recombinations at the cell's back surface; a textured front surface and multilayer antireflection coatings that decrease reflected light; and a back-surface mirror that reduces operating temperature. For silicon cells, maximum achievable efficiency appears to be 16–18%. Consequently most development efforts are focused on improving the design of the array or perfecting the gallium aluminum arsenide/gallium arsenide (GaAlAs/GaAs) cell technology. The potential of GaAlAs/GaAs solar cells to obtain efficiencies of 20-25% has prompted major developments in this technology. A liquid-phase epitaxially grown GaAlAs/GaAs cell has achieved efficiencies as high as 18%. A metal/organic chemical vapor deposition process for cell fabrication and a multibandgap gallium aluminum arsenide/gallium indium arsenide (GaAlAs/GaInAs) cell that uses aluminum and indium material to enhance efficiency are other GaAlAs/GaAs developments. Developments anticipated for GaAlAs/GaAs cells indicate higher-performance cells with significant cost savings for space application. *See* SOLAR CELL.

Solar arrays. The process of assembling individual cells into series and parallel circuits to obtain the required working voltage and power is intricate and complex. There are two basic cell-mounting configurations. Body-mounted arrays, because of limited available space, occasionally use a flexible shingle arrangement. This arrangement usually provides more cells per unit area, thereby providing a higher output for a given area. However, flat mounting, with space provided between cells for electrical interconnects, is most frequently used (**Fig. 1**).

Many solar arrays now in use are in the 1-10-kW range, such as the array for the *Solar Maximum Mission* (Fig. 2). The four solar arrays for the *International Space Station* produce 31 kW each.



Fig. 1. Cross section of a solar array, illustrating the materials and components used in assembly with the flat-mounting technique.

The flexible roll-up solar-array (FRUSA) technology was first demonstrated in 1971 on an Air Force flight-test validation program. Solar arrays such as the FRUSA have a specific power of 20 W/lb (44 W/kg). Studies for a solar-array-powered propulsion stage have demonstrated the feasibility of a 12.5-kW array with a specific power of 30 W/lb (66 W/kg), and work has been undertaken to eventually achieve a specific power of 90 W/lb (200 W/kg).

The *Hubble Space Telescope* used two rolled-up arrays. They developed cooling-heating problems



Fig. 2. One of the two solar-array wings flown on the Solar Maximum Mission spacecraft. Overall dimensions are 7.2 by 11.5 ft (2.2 by 3.5 m).

which caused vibration during the transitions of the telescope from shadow to sunlight. The arrays were redesigned and replaced during a telescope reservicing mission. *See* HUBBLE SPACE TELE-SCOPE.

With the advent of GaAlAs/GaAs cells, solar arrays using solar concentrators have been studied. The useful range of conventional solar arrays for deepspace missions is limited to about 1.5 astronomical units (AU) [Mars orbit]. Studies have shown that, with solar concentration, the photovoltaic power system's useful range can be extended to at least 9 AU [Saturn orbit]. A concentrator system can be used with either silicon or GaAlAs/GaAs cells. However, the GaAlAs/GaAs solar arrays potentially have several advantages over silicon solar arrays, including higher beginning-of-life (BOL) efficiencies, less degradation in a radiation environment, higher operating temperatures, more rapidly increasing performance with concentrator ratios, and greater ability to recover from radiation damage at lower temperatures. With concentrator ratios of 8:25, an overall power density at beginning of life of 9 W/lb (20 W/kg) is projected. The potentially low cost of concentration may make this approach economically more attractive than RTGs.

One concentrator system is aboard an HS 702 geosynchronous spacecraft, designed by Hughes Space and Communications Company. The satellite, *Galaxy XI*, was launched on December 21, 1999. The solar arrays were deployed on January 21, 2000, and for the first time television cameras were used to watch the deployment. Most concentrator designs use a lens to focus more sunlight onto the cells. The Hughes design used reflector panels on each side of the solar panel. The cells were Ga-As and produced 12.4 kW at approximately 25% efficiency.

Another mission that used concentrators was *Deep Space 1*. It had 720 lenses to focus sunlight onto the solar arrays, producing 2500 W at Earth's orbit.

Other proposals have been made to improve the efficiencies and thus power outputs of solar arrays. One proposal is to stack cells on top of each other, each efficient at absorbing certain bandgap frequencies (colors) of light but allowing other frequencies to be transmitted to the next cell. A GaAs threejunction multibandgap cell has a calculated efficiency of over 34%. Another proposal would use the same principle, but instead of stacking, the cells would be placed side by side and a prism would spread the white light into the rainbow of colors that the cells would absorb. Five different cells would have approximately 32% efficiency in normal sunlight, and if concentrators were used before the prisms to concentrate the light by eight times, the efficiency would be 43%. See SOLAR ENERGY.

Batteries. Although both primary and secondary (rechargeable) batteries have been used in the space program, satellites now in use almost always employ rechargeable batteries. Five different rechargeable electrochemical systems are used in space power systems: nickel-cadmium (NiCd), silver-cadmium

(AgCd), silver-zinc (AgZn), and nickel-hydrogen (NiH), and lithium-ion (Li-ion).

The predominant factor in the use of nickelcadmium batteries is the exceptional long-cycle life capability (**Fig. 3**). An excess of 8 years (over 40,000 cycles) has been obtained on nickel-cadmium batteries operating in a near-Earth environment. At synchronous altitudes, several satellites have operated in excess of 10 years with nickel-cadmium batteries.

Sealed silver-cadmium batteries have seen limited applications in space power systems, even though the energy density is higher than that of a nickelcadmium battery. This limited use is due primarily to its very low cycle life (less than one-tenth that of the nickel-cadmium battery). Missions that use the silver-cadmium battery usually require a spacecraft with ultraclean magnetic properties. With an appropriately designed and wired silver-cadmium battery, the magnetic field is typically less than 1 gamma (1 nanotesla) at 1 ft (30 cm). Lifetimes in excess of 4 years have been obtained on missions with low battery duty cycles and a "cold storage" of the battery during nonduty cycles.

Because silver-zinc batteries offer excellent energy density but have a very limited life, particularly when used in a cycling mode, they have been used only in space power systems that require only a few days to a few months of operation.

Nickel-hydrogen batteries have replaced nickelcadmium batteries on some space missions. They have a higher energy density and energy per volume than nickel-cadmium. They have been used on the *Satcom V* communication satellite, the *Eartb Observation Satellite*, and the *Hubble Space Telescope*. In fact, the original batteries for the *Hubble Space Telescope* were nickel-cadmium, but they were changed to nickel-hydrogen because the telescope is colder than was expected, leading to more power being used by the heaters.

Lithium-ion batteries were formerly considered too dangerous for space missions since lithium can catch fire or explode. However, this changed with the launch of *Syracuse IIIA* from French Guiana in October 2005, the first spacecraft to carry a lithiumion battery. The advantage of lithium-ion batteries is that they can store the same amount of energy as a nickel-hydrogen battery in a smaller volume with 50% less mass. Also, after a month of storage, they still have 98% of their charge. This means that no recharge is needed on the launch pad.

There are several batteries under development to replace nickel-cadmium (since cadmium is toxic) and nickel-hydrogen (since such batteries are expensive to produce). Nickel-metal hydride (NiMH), silver-metal hydride (AgMH), and sodium-sulfur (NaS) batteries are the prime contenders. Nickelmetal hydride batteries are similar to nickel-hydrogen except that the hydrogen is stored in the metal hydride solid rather than as a gas. The pressure of nickel-metal hydride is about 50 lb/in.² (350 kilopascals), compared to 1000 lb/in.² (7 megapascals) for nickel-hydrogen. The nickel-hydrogen batteries



Fig. 3. A 20-Ah battery designed for space flight. Hermetically sealed nickel-cadmium cells are series-connected to provide spacecraft power at approximately 27 V.

are therefore in the form of a round pressure vessel, whereas the nickel-metal hydride batteries can be made rectangular. Nickel-metal hydride batteries have twice the energy density of nickel-cadmium and twice the energy per unit volume. They would also have twice as much energy per unit volume as nickelhydrogen. Silver-metal hydride batteries might have three times the energy density of nickel-cadmium. *See* ENERGY STORAGE.

Flywheels. Flywheels store mechanical energy using a spinning wheel. When connected to a motor/ generator, the solar arrays spin up the motor/flywheel. When the arrays are in shadow, the flywheel spins down and the generator produces electricity. Thus, the flywheel can be used to "store" electricity like a battery. Prior to 2001, NASA worked on development of a flywheel energy storage (FES) system for the International Space Station. The solar arrays provide power for the station, and part of that electricity is used to charge the batteries. When the station enters the shadow of the Earth, the batteries supply the power to the station through the Battery Charge/Discharge Unit (BCDU). The flywheel system was intended to replace the batteries and the BCDU. The flywheels were expected to store more energy than the batteries, allowing extra electrical capacity for the experiments. Batteries have a limited amount of charge/discharge cycles, whereas flywheels can be charged and discharged virtually forever. Therefore they last longer and were expected to save hundreds of millions of dollars over the life of the station. Testing at Johnson Space Center was scheduled to start in 2001 with a planned launch to the *International Space Station* in 2004. However, this program was canceled by the Bush Administration in 2001. *See* SPACE STATION.

Fuel cells. The technology resulting from the *Gemini* and *Apollo* fuel-cell programs served as the basis for the space shuttle fuel-cell development. The shuttle fuel cell has three sets of 32 series-connected cells that provide an operating voltage range of 27.5-32.5 V and power of 16 kW (with maximum 21 kW). Each cell is designed for up to 2400 h of operation and weighs 281 lb (127 kg). The shuttle power plant consists of three fuel cells that feature a startup time of 15 min (compared with 24 h for *Apollo*).

Fuel cells have four distinct advantages for crewed missions: high power, high specific power (56.9 W/lb or 125.8 W/kg), high efficiency (70%), and a benign by-product—water. In fact, the astronauts drink the water produced by the shuttle fuel cells. *See* FUEL CELL.

Tethers. Tethers are a space power system currently in development. A conducting tether many kilometers long is deployed from a spacecraft in orbit about a planet with a magnetic field, such as Earth. The solar wind issues electrons, which travel in spiral paths around the magnetic field lines. As the conducting tether crosses the field lines, it induces an electric field, the same principle that terrestrial generators use. Using an electron gun on the spacecraft to shoot electrons back into space completes the circuit and current flows. The electromotive force (emf) for low-Earth-orbit (LEO) satellites is a few tenths of a volt per meter. A 6-mi (10-km) tether would produce several kilovolts, leading to several kilowatts of power. The TSS-1 (Tethered Satellite System 1) shuttle payload tried to test this idea but had problems reeling out the tether. Tether voltages as high as 3500 V and current levels of 480 mA were obtained in a subsequent flight, generating a modest amount of power, before the tether severed at 12.2 mi (19.6 km). This was better than was predicted by the analytical models. *See* ELECTROMAG-NETIC INDUCTION.

Solar dynamic systems. Solar dynamic systems have been tested to achieve higher efficiencies and power while using less mass and volume than a solar array would use. The principle is that a parabolic or nearly parabolic reflector would concentrate solar energy on a receiver. The receiver has a thermal storage area (for when the station enters Earth's shadow) and a fluid. The solar energy heats the fluid, which turns a turbine-generator to supply electricity. Efficiencies of roughly 30–40% are possible.

A joint NASA and Russia program to put a solar dynamic system on the space station *Mir* was canceled in February 1996 after several successful tests. The flight hardware built for the *Mir* program is currently in storage at the NASA Glenn Research Center. Plans by NASA to use most of the existing hardware and redesign the system to be installed on the *International Space Station* were canceled.

Nuclear power systems. Nuclear power system concepts encompass radioisotopes and nuclear reactor heat sources coupled with thermal-to-electric converters. Nuclear power is now limited to a few experimental applications. Significant advances in materials must be made before a prototype converter can be demonstrated. The relatively heavy weight (2.5-3 W/lb or 5.5-6.6 W/kg), high cost, and lack of demonstrated reliability make nuclear power systems unattractive for Earth-orbiting space power systems below the kilowatt range.

Estimates indicate that the nuclear system could become competitive with solar power at 25 kW or greater. However, with the reentry of a nuclearreactor-powered Soviet satellite (*Cosmos 954*) in Canada during the summer of 1979, the use of nuclear reactors in space moved from the technical domain to the political arena. The need and desirability of nuclear power in space is a controversial issue with worldwide ramifications.

RTGs have been used for space-borne power primarily for special or unique applications. Present RTGs are descended from power sources developed by the Space Nuclear Auxiliary Power (SNAP) program starting in the 1950s. Although RTGs have been used occasionally in Earth-orbiting satellites, their principal space use has been to power space probes and lunar equipment. RTGs remain on the lunar surface from *Apollo 12, 14, 16*, and *17* missions. These units used lead telluride thermoelectric material and produced approximately 75 W of power. The *Galileo* probe to Jupiter, the *Ulysses* probe of the Sun, and the *Cassini* probe to Saturn use the most advanced RTG so far, the general-purpose heat source (GPHS) RTG. These use silicon-germanium thermoelectric material and produce 300 W of power. *See* NUCLEAR BATTERY; THERMOELECTRICITY.

For power levels above 1 kW, a dynamic isotope power system (DIPS) is desirable. The radioisotope is used to power a Rankine, Brayton, or Stirling device. This system has roughly three times the efficiency of a thermoelectric generator. NASA is currently (2006) testing Stirling generators. *See* BRAYTON CYCLE; RANKINE CYCLE; STIRLING ENGINE. Jeffrey C. Mitchell

Bibliography. J. A. Angelo, Jr., and D. Buden, *Space Nuclear Power*, Orbit Book Company, Malabar, FL, 1985; S. W. Angrist, *Direct Energy Conversion*, 3d ed., Allyn and Bacon, Boston, 1976; M. D. Griffin and J. R. French, *Space Vehicle Design*, 2d ed., American Institute of Aeronautics and Astronautics, Washington, DC, 2004.

Space probe

An automated crewless vehicle, the payload of a rocket-launching system, designed for flight missions to other planets, to the Moon, and into interplanetary space, as distinguished from Earth-orbiting satellites (see **table**).

The space probe is used primarily for scientific purposes, which are stated as the mission objectives. Missions have been designed to explore many diverse targets, such as the rocky bodies of the inner solar system, including Mercury, Venus, Mars, and the asteroids; the giant gaseous planets in the outer solar system, Jupiter, Saturn, Uranus, and Neptune; and icy bodies originating in the outer solar system, such as comets, Kuiper Belt objects, and the dwarf planet Pluto. Other missions are designed to study solar physics, the properties of interplanetary space, or targets far beyond the solar system. Most spacecraft launched to a planet or other body also study the environment of charged particles and electromagnetic fields in interplanetary space during their cruise phase en route to the destination.

Missions may also be categorized by complexity. The simplest are flyby spacecraft, which study their target body during a relatively brief encounter period from a distance of hundreds to thousands of miles as they continue past. Next are orbiters, which circle a planet or other body for extended study; some may carry atmospheric descent probes. Even more complex are lander missions, which touch down on a planet or other body for the collection of on-site data; some may bear exploration rovers designed to range beyond the immediate landing site. Finally, the most complex space probes envisaged are samplereturn missions, which would collect specimen material from a target body and return it to Earth for detailed study.

Spacecraft Subsystems

In the broadest terms, a space probe may be considered a vehicle that transports a payload of sensing

Norse	Loupeb data	Commente
Name	Launch date	Comments
Luna 1	Jan. 2, 1959	Lunar flyby; passed within 3278 mi (5257 km) of the Moon
Pioneer 4	Mar. 3, 1959	Cosmic rays; passed 37,300 mi (60,000 km) from Moon
Luna 2	Sept. 2, 1959	Impacted Moon Photographed far side of Moon
Pioneer 5	Mar 11 1960	Study of space between Earth and Venus: magnetic fields and cosmic rays
Mariner 2	Aug. 26, 1962	First planetary flyby: Venus probe
Ranger 7	July 28, 1964	Lunar impact and approach photography
Mariner 4	Nov. 28, 1964	Mars flyby; photography, magnetic fields, cosmic rays
Ranger 8	Feb. 17, 1965	Lunar impact and approach photographs
Ranger 9	Mar. 21, 1965	Lunar impact in Alphonsus; approach photography
Zond 3	July 18, 1965	Flyby of Moon; photographs, other data
Pioneer 6	Dec. 16, 1965	Solar orbiter
Luna 10	Mar 31, 1900	First solt landing on Moon: collected data, but no camera instrument
Surveyor 1	May 30, 1966	Soft landing on Moon: environmental data and photography
Lunar Orbiter 1	Aug. 10, 1966	Lunar photographs
Pioneer 7	Aug. 17, 1966	Solar orbiter
Luna 11	Aug. 24, 1966	Lunar orbiter; TV camera failed
Luna 12	Oct. 22, 1966	Lunar orbiter; successful return of photographs and other data
Lunar Orbiter 2	Nov. 6, 1966	Lunar orbital photography
Luna 13	Dec, 21, 1966	Lunar lander; surface photography and soil information
Venera A	Feb. 5, 1907	Analysis of Venus atmosphere: lander failed during descent
Mariner 5	June 14, 1967	Venus flyby: atmospheric and magnetospheric data
Surveyor 3	Apr. 17, 1967	Lunar surface photography and surface properties
Lunar Orbiter 4	May 4, 1967	Lunar orbital photography
Lunar Orbiter 5	Aug. 1, 1967	Lunar orbital photography
Surveyor 5	Sept. 8, 1967	Lunar surface photography and surface properties, including elemental analysis
		of surface
Surveyor 6	Nov. 7, 1967	Same as Surveyor 5; landing in Sinus Medii
Pioneer 8 Surveyor 7	Dec. 13, 1967	Solar orbiter
Luna 14	$\Delta nr 7 1968$	Lunar orbiter
Pioneer 9	Nov. 8, 1968	Solar orbiter
Zond 6	Nov. 10, 1968	Lunar flyby; crashed on return to Earth
Venera 5	Jan. 5, 1969	Same as Venera 4
Venera 6	Jan. 10, 1969	Same as Venera 4
Mariner 6	Feb. 25, 1969	Photography and analysis of surface and atmosphere of Mars
Mariner /	Mar. 27, 1969	Same as <i>Mariner 6</i>
Zond 7	July 14, 1969 Aug. 8, 1969	Lunar orbiter (crashed during attempted lunar landing) Reentered Aug. 14, 1969; third crewless circumlunar flight; recovered in the Soviet Union
Venera 7 Luna 16	Aug. 17, 1970 Sept. 12, 1970	Lander capsule transmitted 23 min from surface of Venus, Dec. 15, 1970 Reentered Sept. 24, 1970; crewless Moon lander touched down on Sea of Fertility Sept. 20, 1970; returned lunar soil samples
Zond 8	Oct. 20, 1970	Circled Moon; recovered Oct. 27, 1970
Luna 17	Nov. 10, 1970	Landed on Moon Nov. 17, 1970; crewless Moon rover
Mars 2	May 19, 1971	First Mars landing
Mars 3	May 28, 1971	Mars probe
Mariner 9	May 30, 1971	Mars probe
Luna 10	Sept. 2, 1971	Lunar photography mission
Luna 20	Feb 14 1972	Lunar sample return
Pioneer 10	Mar. 2. 1972	Jupiter encounter: transiovian interplanetary probe
Venera 8	Mar. 27, 1972	Venus landing July 22, 1972
Luna 21	Jan. 8, 1972	Lunar lander with rover
Pioneer 11	Apr. 5, 1973	Jupiter encounter and trasjovian interplanetary probe; also Saturn encounter
Mars 4	July 21, 1973	Mars orbiter
Mars 5	July 25, 1973	Mars londer
Mars 7	Aug. 5, 1973	Mars lander
Mariner 10	Nov. 3, 1973	Venus and Mercury encounter
Luna 22	May 29, 1974	Lunar orbiter
Helios 1	Dec. 10, 1974	Inner solar system, solar wind exploration
Venera 9	June 8, 1975	Venus probe
Venera 10	June 14, 1975	Venus probe
Viking 1	Aug. 20, 1975	Mars lander and orbiter
Viking 2	Sept. 9, 1975	Mars lander and orbiter
Luna 24	Jun 9 1976	Interpreterary, similar objectives to those of Hellos T
Voyager 2	Aug. 20, 1977	Jupiter, Saturn, Uranus, and Neptune encounters; also satellites and ring systems
Voyager 1	Sept. 5, 1977	Same objectives as <i>Voyager 2</i> with some orbital differences giving differing encounter trajectories
Pioneer Venus Orbiter Pioneer Venus	May 20, 1978 Aug. 8, 1978	Returning atmospheric, surface, and particle and field information Penetration of Venus atmosphere by four probes; returned atmospheric data
Multi-Probe Bus Venera 11	Sept. 8, 1978	Venus lander; returned information on surface properties; detection of lightning and thunderlike sounds
Venera 12	Sept. 14, 1978	Similar mission to Venera 11 (conti.)

Important space probes (cont.)					
Name	Launch date	Comments			
Venera 13	Oct. 30, 1981	Venus lander			
Venera 14	Nov 4 1981	Venus lander			
Venera 15	June 2, 1983	Venus lander: surface topography			
Venera 16	lune 7, 1083	Similar mission to Venera 15			
International Compton	Julie 7, 1300	Originally International Sun Forth Evalurar 2 (ISEE 2) Forth actallita, redirected			
Explorer (ICE)	—	using a lunar swingby on Dec. 22, 1983, to encounter with Comet Giacobini-Zinner; plasma and magnetic field measurements			
Vega 1	Dec. 15, 1984	Venus probe-Halley intercept			
Vega 2	Dec. 21, 1984	Venus probe-Halley intercept			
Sakigake	Jan. 8, 1985	Halley intercept; precursor to Suisei, upgraded to full mission			
Giotto	July 2, 1985	Halley intercept			
Suisei	Aug. 19, 1985	Halley intercept; plasma and magnetic field measurements			
Phobos 1	July 7, 1988	Mars/Phobos probe, lost by command error			
Phobos 2	July 12, 1988	Mars/Phobos probe; some data, but communications lost			
Magellan	May 4, 1989	Venus radar mapper			
Galileo	Oct. 18, 1989	Jupiter orbiter and atmospheric probe			
Hiten/Hagoromo	Jan 24, 1990	Moon orbiter and relay probe: orbiter transmitter malfunctioned			
l Ilvsses	Oct 6 1990	Solar polar orbiter			
Mars Observer	Sept 25 1992	Contact lost 3 days before Mars arrival			
Clementine	Jan 25 100/	Orbited Moon: thruster malfunction prevented asteroid flyby			
Solar and Holicophoria	Dec 2 1005	Orbite L1 libration point to study the Sup			
Observatory (SOHO)	Dec. 2, 1995 Feb. 17, 1996	Asteroid orbiter			
Rendezvous	100.11,1000				
Mars Clobal Surveyor	Nov 7 1006	Mare orbitor			
Mara 06	Nov. 16, 1006	Mars orbiter and landers: launch vehicle failed			
Mara Dathfindar	Dec 4 1006	Mars lander and rever			
Advanced Composition	Aug. 25, 1997	Orbits L1 libration point to study charged particles			
Explorer	0-+ 15 1007	Ostum subitsu/Titsu desearch and s			
Cassini	Oct. 15, 1997	Saturn orbiter/ litan descent probe			
Lunar Prospector	Jan. 6, 1998	Lunar orbiter			
Nozomi Deep Space 1	July 4, 1998 Oct. 24, 1998	Mars orbiter; failed to reach Mars orbit Test of ion engine and 11 other advanced technologies; asteroid and comet			
Marra Oliverata Ortaitan	D 11 1000				
Mars Climate Orbiter	Dec. 11, 1998	Lost during Mars arrival			
Mars Polar Lander	Jan. 3, 1999	Lost during Mars arrival			
Stardust	Feb. 7, 1999	Comet flyby, dust sample return			
Mars Odyssey Wilkinson Micowave	April 7, 2001 June 30, 2001	Mars orbiter Solar orbiter studying cosmic background radiation			
Anisotropy Probe	A	Option wheels are also wetween the and increased we are the adia at in 1 the b			
Genesis	Aug. 8, 2001	Solar wind sample return; hard impact upon landing in Utan			
Comet Nucleus Tour	July 3, 2002	Intended for two comet flybys; failed shortly after launch			
Hayabusa	May 9, 2003	Comet sample return mission; sample collection uncertain			
Mars Express	June 2, 2003	Mars orbiter			
Beagle 2 Spirit (Mars Exploration	June 2, 2003 June 10, 2003	Lander carried on Mars Express; failed during descent Mars rover			
Rover A) Opportunity (Mars	July 7, 2003	Mars rover			
Exploration Rover B) Spitzer Space	Aug. 25, 2003	Infrared telescope in Earth-trailing heliocentric orbit			
Telescope					
	Sept. 27, 2003				
HUSEIIA	Narch 2, 2004	Comet orbiter-lander			
wessenger	Aug. 3, 2004	iviercury orbiter			
Deep Impact	Jan. 12, 2005	Released impactor to excavate crater in comet nucleus			
Mars Reconnaissance Orbiter	Aug. 12, 2005	Mars orbiter			
Venus Express	Nov. 9, 2005	Venus orbiter			
New Horizons	Jan. 19, 2006	Pluto flyby			
Solar Terrestrial Relations Observatory	Oct. 26, 2006	Twin spacecraft in solar orbit			

instruments to the vicinity of a target body. Thus, the spacecraft must include a number of subsystems to provide power, to communicate with Earth, to maintain and modify attitude and perform maneuvers, to maintain acceptable on-board temperature, and to manage the spacecraft overall. *See* SPACE TECHNOLOGY.

Scientific instruments. The scientific payload may be divided into remote-sensing instruments such as cameras, and direct-sensing instruments such as magnetometers or dust detectors. They may be classified as passive instruments, which detect radiance given off by a target body, or active, which emit energy such as radar pulses to characterize a target body.

Examples of passive remote-sensing instruments are cameras and other imagers, polarimeters, photometers, spectrometers, and radiometers. Passive direct-sensing instruments include those that study magnetic fields and charged particles, such as highand low-energy particle detectors, plasma instruments, plasma wave detectors, dust detectors, and magnetometers, as well as instruments on landers such as those of the *Viking* Mars mission. Active remote-sensing instruments include syntheticaperture radar and altimeters.

Early missions up to and including Voyager used vidicon tubes similar to those in terrestrial television cameras of that period as the detector of the onboard camera. Some, including Lunar Orbiters and Luna missions, actually used film which was developed and scanned on board. The image information was then digitized before being transmitted to Earth. Subsequent missions have used solid-state chargecoupled devices (CCDs), similar to those in contemporary television cameras. A spacecraft's camera system frequently is outfitted with two sets of optics to provide wide- and narrow-field views; the narrowfield camera is the equivalent of a small on-board telescope. In addition, the camera system includes a filter wheel with various colored filters which may be placed in the optical path. Color images are made by rapidly taking a series of three exposures through different filters; these are transmitted to Earth and combined into a single frame during image processing. Most spacecraft cameras, such as those on Voyager, Galileo, or Cassini, capture light in discrete snapshot-type exposures. By contrast, some orbiters, such as Mars Global Surveyor and Mars Reconnaissance Orbiter, carry pushbroom-type cameras; these collect data continuously, building up a long ribbonlike image as the spacecraft passes over the planet's surface. See ASTRONOMICAL IMAGING; CHARGE-COUPLED DEVICES; COLOR FILTER; REMOTE SENSING; TELEVISION CAMERA TUBE.

Power subsystem. Electrical power is required for all spacecraft functions. The total required power ranges from about 300 to 2500 W for current missions, depending on the complexity of the spacecraft. The power subsystem must generate, store, and distribute electrical power.

All space probes launched so far have generated power either via solar panels or via radioisotope power systems. Solar panels are generally employed on missions to the solar system's inner planets, where sunlight remains relatively strong. The strength of sunlight diminishes as the square of the distance from the Sun. Missions beyond the orbit of Mars have therefore relied on radioisotope systems. In addition, radioisotope systems have been used on the *Viking* Mars landers because of the unreliability of sunlight during dust storms and other periods.

Solar panels typically employ photovoltaic devices constructed of materials such as crystalline silicon or gallium arsenide. A silicon cell 2.4 in. (6 cm) in diameter, for example, produces a current of about 1 A at 0.25 V under direct sunlight at a distance of 1 astronomical unit (1 AU, the average distance from Earth to the Sun, is about 9.3×10^7 mi or 1.5×10^8 km). Solar panels are mounted on arrays which may be articulated to optimize pointing at the Sun; alternatively, they may be off-pointed during periods to generate less electrical energy in the interest of on-board power management. The performance of solar panels degrades gradually at a rate of perhaps 1–2% per year. On the solar-powered *Magellan* mission, power loss was unusually great. After a 15month interplanetary cruise and 4 years in orbit at Venus, careful power management was required in the mission's final days before *Magellan* was intentionally directed into the planet's atmosphere to burn up in October 1994. *See* ASTRONOMICAL UNIT; SOLAR CELL.

The other principal power source, the radioisotope power system, produces electrical power through the radioactive decay of plutonium-238. The radioisotope device is not a nuclear reactor, however; as the plutonium naturally decays, it gives off heat which is converted to electricity. In early devices this conversion was accomplished by thermocouples made of silicon-germanium junctions, whereas one newer design uses thermocouples composed primarily of lead telluride, and another envisages the use of a dynamic Stirling converter. Waste heat not converted to electricity is radiated into space via metal fins on the generator casing. The output of radioisotope power systems gradually decays at about 1-2% per year, slightly more rapidly than actively used solar panels. See NUCLEAR BATTERY.

Because they typically go through alternating sunlight and shadow, spacecraft powered by solar panels must carry batteries to store electricity for periods when the panels are in shade. Nickel hydrogen, nickel metal hydride, and lithium batteries are often used. Their performance degrades after many cycles of charging and partial discharging, but the batteries may be reconditioned by discharging them completely, much like rechargeable batteries in terrestrial applications. Because the electrical power output of radioisotope power systems does not vary significantly from hour to hour, spacecraft equipped with them do not require batteries, but may require short-term energy-storage devices, such as capacitors, to provide for high-pulse current needs.

The other function of the power subsystem is to distribute electrical power. A typical spacecraft supplies its various on-board subsystems with a specific constant bus voltage, although some instruments may require special voltages. During the course of the mission, various instruments are switched on and off as they are used to make observations. This switching is performed with solid-state or mechanical relays. A shunt regulator is generally used to maintain a constant voltage on the main bus; this converts electrical energy into heat, which is radiated into space. In some cases it may be necessary to switch instruments on and off in a particular order to manage the voltage on the main bus. *See* RELAY; SPACE POWER SYSTEMS.

Telecommunications subsystem. In order to accomplish its mission, the spacecraft must maintain communications with Earth. This includes receiving commands sent from ground controllers, as well as transmitting scientific data and routine engineering "housekeeping" data. All of these transmissions are made in various segments of the microwave spectrum. Early missions typically used lower frequencies in the L or S band (1–4 GHz), while subsequent missions have used the X band (7–9 GHz). The

design of the telecommunications subsystem takes into account the volume of data to be transmitted and the distance from Earth at which the spacecraft will operate, dictating such considerations as the size of antennas and the power of the on-board transmitter. *See* MICROWAVE; TELEMETERING.

Advanced planetary probes have carried a dishshaped high-gain antenna which is the chief antenna used to both transmit and receive. These antennas typically consist of a large parabolic reflector, with a subreflector mounted at the main reflector's focus in a Cassegrain-type configuration. Most large high-gain antennas are fixed to the spacecraft's body; the entire spacecraft must maintain Earth pointing to avoid disrupting communication. Smaller parabolic antennas used in relatively close missions such as those to Mars, however, may be on a steerable mount. Most large parabolic antennas are rigid; the high-gain antenna for Cassini, for example, is fabricated from a lightweight composite material. Galileo used a 16-ft (4.8-m) high-gain antenna designed to unfold like an umbrella after launch; although similar antennas functioned well on Earth-orbiting satellites, the antenna on Galileo did not open fully when ground controllers attempted to deploy it in April 1991, rendering it unusable. The beam width of parabolic high-gain antennas is typically very narrow, requiring an Earth pointing to within a fraction of a degree. The advantage of the high-gain antenna is that it allows a much greater transmission rate than do other antenna types. See ANTENNA (ELECTRO-MAGNETISM).

In the interest of redundancy and in the event that Earth pointing is lost, spacecraft virtually always carry other on-board antennas. These may be low-gain antennas, which typically offer nearly omnidirectional coverage except for blind spots shadowed by the spacecraft body. *Galileo* was outfitted with two low-gain antennas which combined to afford completely omnidirectional capability. Other spacecraft have used medium-gain antennas, which provide a beam width of perhaps $20-30^{\circ}$, with a gain midway between that of high- and low-gain antennas. *Magellan*, for example, was equipped with a conical feed horn antenna which it used when its high-gain antenna was off Earth point.

The spacecraft's on-board transmitter typically provides a signal in the tens of watts. The final stage of the transmitter is a power amplifier which may be a traveling-wave tube or a solid-state amplifier. The carrier signal is modulated with a subcarrier which contains the measured scientific or engineering data. Ground controllers frequently make use of the technique of Doppler tracking, wherein slight changes in the frequency of the probe's signal are measured to reveal subtle acceleration changes by the spacecraft. In some cases, spacecraft may be commanded to transmit a pure carrier signal for purposes such as Doppler tracking; this was used with great success in the Magellan mission to map Venus's gravity field by measuring slight acceleration changes in the spacecraft as it passed over various portions of the planet. See DOPPLER EFFECT; MICROWAVE SOLID- STATE DEVICES; POWER AMPLIFIER; RADIO TRANSMITTER; TRAVELING-WAVE TUBE.

The spacecraft's receiver is designed to be sensitive to a narrow range of frequencies within a very small percentage of a specified frequency. Many spacecraft receivers incorporate phase-lock-loop circuitry, which locks on to a signal detected within the specified bandwidth and follows any slight changes in frequency, particularly the Doppler shift produced by the motion of the probe relative to Earth. Although this design has worked well for most missions, Voyager 2 presented special challenges because one of its redundant receivers failed midway through the mission, and the second receiver's frequency varied depending on the temperature of the on-board electronics. Ground controllers thus had to calculate the expected frequency to which the onboard receiver would be sensitive and had to tailor transmissions of spacecraft commands accordingly. See PHASE-LOCKED LOOPS; RADIO RECEIVER.

In most modern cases, the transmitter and receiver may be combined into a single unit called a transponder. Some components of the telecommunications subsystem may be used for purposes other than communications; there may in fact be a radio-science team of investigators who use the on-board radio as a scientific instrument. As noted above, Magellan measured Venus's gravity by means of Doppler tracking of the radio carrier signal. On many missions such as Voyager, the radio signal is tracked as the spacecraft passes behind a planet, revealing information on the chemical makeup of the planet's atmosphere. Magellan's main scientific instrument, its imaging radar, used the spacecraft's high-gain antenna to send and receive radar pulses to and from the planet's surface. See RADAR; RADAR ASTRONOMY; SPACE COMMUNICA-TIONS.

Attitude-control subsystem. It would be impossible to navigate the spacecraft successfully or point its scientific instruments or antennas without closely controlling its orientation in space, or attitude. Many specialized techniques have been developed to accomplish this.

Some spacecraft, particularly earlier ones, have been spin-stabilized; during or shortly after launch, the spacecraft is set spinning at a rate on the order of a few revolutions per minute. Much like a rotating toy top, the spacecraft's orientation is stabilized by the gyroscopic action of its spinning mass. Thruster jets may be fired to change the spacecraft's attitude. Spin stabilization is preferred for scientific instruments that measure electromagnetic fields and charged particles, because they are continuously swept in a circle in such a way as to cancel out interference from the spacecraft itself. The stabilization method is not as good for scientific instruments that require pointing, such as optical cameras. The Pioneer 10 and 11 spacecraft were spin-stabilized, capturing photolike images with sweep polarimeters. Ulysses exclusively carries fields and particles instruments, and thus is also spin-stabilized.

Most planetary spacecraft, however, are three-axisstabilized, meaning that their attitude is fixed in

relation to space. This method makes the operation of pointed instruments much simpler, although it is not as preferable for fields and particles instruments. The spacecraft's attitude is maintained and changed via on-board thruster jets or reaction wheels, or a combination of both. Reaction wheels are electrically powered devices similar to gyroscopes which rotate the spacecraft about a given axis by spinning up or slowing down a massive wheel. Although most planetary spacecraft are three-axis-stabilized, Galileo was the only spacecraft to use a complex dual-spin design. Most of the Galileo spacecraft spins at three revolutions per minute, while a despun segment carrying the camera and other pointed instruments is held in a three-axis-stabilized attitude. The two segments of the spacecraft are joined by a complex spin bearing that must transfer hundreds of electrical signals. See GYROSCOPE.

Other parts of the attitude-control subsystem include Sun sensors and star trackers which are used to establish points of reference in space. The subsystem's functions are usually overseen by a dedicated computer system distinct from the spacecraft's main command and data computer. Because spacecraft must operate at great distances from Earth where radioed commands would take many minutes or even hours to arrive, the attitude-control subsystem is usually one of the most autonomous parts of the spacecraft, maintaining the spacecraft's orientation with only infrequent intervention from the ground. *See* STAR TRACKER.

Propulsion subsystem. Most spacecraft are outfitted with a series of thruster jets, each of which produces approximately 0.2-2 pounds-force (1-10 newtons) of thrust. The thrusters are placed so that at least two lie in opposite directions on each of the spacecraft's three orthogonal axes. Additional thrusters are often added to lend redundancy. Thrusters are usually fueled with a monopropellant, hydrazine, which decomposes explosively when it contacts an electrically heated metallic catalyst within the thruster. The system also includes helium used to pressurize the propellant.

In addition to maintaining the spacecraft's attitude, on-board thrusters are used for trajectorycorrection maneuvers. These maneuvers are usually carried out at several points during interplanetary cruise to progressively fine-tune the spacecraft's aim point at its destination. Maneuvers are conducted by firing one or more thrusters for periods from several seconds to several minutes, accelerating or decelerating the spacecraft typically on the order of several meters per second.

Spacecraft designed to orbit a planet or similar target body must carry a larger propulsion element capable of decelerating the spacecraft into orbit upon arrival. *Magellan* used a solid-fuel motor to insert it into orbit at Venus, whereas spacecraft such as *Galileo* and *Cassini* use bipropellant hypergolic substances, monomethylhydrazine and nitrogen tetroxide, which expand rapidly in a flameless combustion when mixed. The main engine on *Galileo* and *Cassini* used to place the spacecraft in planetary orbit provided 90 and 110 lbf (400 and 490 N) of thrust, respectively. Mars orbiters have taken advantage of different propulsion approaches. *Mars Global Surveyor* and *Mars Odyssey* each used a large main engine [135 and 154 lbf (600 and 685 N), respectively] to enter orbit at Mars. *Mars Reconnaissance Orbiter*, by contrast, braked into orbit by firing a group of six thrusters supplying 38 lbf (170 N) of thurst each. *See* SPACECRAFT PROPULSION.

Thermal control subsystem. The environment of space is very harsh, ranging from temperatures near absolute zero in the outer solar system to the extremely hot locales of inner planets such as Venus and Mercury. In order to minimize the impact of temperature variations on the electronics onboard, spacecraft nearly always incorporate some form of thermal control. Mechanical louvers, controlled by bimetallic strips similar to those in terrestrial thermostats, are often used to selectively radiate heat from the interior of the spacecraft into space. Other thermal strategies include painting exterior surfaces; Magellan used white thermal blankets to reflect heat and keep interior units cooler. Magellan also carried optical solar reflectors composed of quartz mirror tiles on a number of surfaces, including the back sides of its solar arrays. Internal components are frequently painted black to encourage heat exchange between these components.

In some cases, spacecraft may also carry one or more active forms of heating to maintain temperature at required minimums. Often this is an electric unit that radiates heat from a resistive element similar to electric heaters on Earth; it may be controlled autonomously by on-board logic or commanded from the ground. Some spacecraft carry several radioisotope heating units (RHUs), each containing a small amount of plutonium which heats nearby spacecraft components. Because of their power needs, active cooling systems are not generally feasible on planetary spacecraft.

Command and data subsystem. This designation is given to the main computer that oversees management of spacecraft functions and handling of collected data. Blocks of commands transmitted from Earth are stored in memory in the command and data subsystem and are executed at prescribed times. This subsystem also contains the spacecraft clock in order to accurately pace its activities, as well as all the activities of the spacecraft.

In its data-handling function, the subsystem receives data destined for the ground on the spacecraft's data bus from other on-board subsystems, processes them, formats them for telemetry, and delivers them to the telecommunications subsystem for transmission to Earth. Included are data from scientific instruments as well as housekeeping engineering data such as voltage and temperature readings. Data are usually stored on tape recorders on board while awaiting transmission to Earth. Since the 1990s, solid-state data recorders have taken the place of traditional tape-driven devices for this purpose. Most spacecraft encode data to be sent of Earth in order to correct transmission errors; Viterbi, Golay, and Reed-Solomon encoding schemes have been used, and Turbo coding is now the most common. Some spacecraft also perform data compression to reduce the number of bits to be transmitted. This technique was used extensively on *Galileo* to compensate partially for the fact that the spacecraft's high-gain antenna did not deploy completely. *See* DATA COMPRESSION; INFORMATION THEORY.

Another critical function of the command and data subsystem is fault protection for the spacecraft in the event of a mishap. One of the chief strategies designed to respond to a contingency is safing, which shuts down or reconfigures subsystems to prevent damage either from within or from the external environment. After carrying out these activities, the fault-protection software usually commands the spacecraft to begin an automated search to regain Sun reference and reestablish communications. Spacecraft safing events that have taken place in actual flight missions have usually been caused by single-event upsets, probably resulting from a cosmic ray interfering with on-board electronics or from transient electrical phenomena from on-board sources. Spacecraft are usually recovered after safing events, although one of the Phobos probes was believed to be lost because of a faulty software command. Another example of on-board fault protection is the command-loss timer, which counts how much time has passed since a command has been received from Earth. If this count passes a designated value, it is assumed that a failure has occurred in the telecommunications or command subsystems. The spacecraft then switches to backup systems and takes other actions in an attempt to reestablish reliable contact.

Structure subsystem. The spacecraft's physical structure is considered a subsystem itself for the purposes of planning and design. Usually the heart of this structure is a spacecraft bus, often consisting of a number of bays, which houses the spacecraft's main subsystems. Besides providing physical support, the bus generally offers an electrical grounding reference, radio-frequency interference shielding, and protection from radiation and meteoroids. The bus also provides attachment points for appendages such as booms and scanning platforms bearing scientific instruments. Magnetometers, for example, are usually mounted on a boom which may extend 30 ft (10 m) or more away from the spacecraft bus. For launch, this boom collapses and is stowed in a compact canister. The spacecraft bus is almost always covered with one or more layers of thermal blankets to help stabilize on-board temperature and to shield the spacecraft from micrometeoroid impacts. Special shielding may be added to protect the spacecraft from radiation in environments such as that at Jupiter, where ionized particles may degrade on-board subsystems, or from temperature extremes at planets such as Venus or Mercury. Spacecraft designed as landers usually also carry aeroshells and parachutes to aid during descent. See ATMOSPHERIC ENTRY; SPACECRAFT STRUCTURE.

Miscellaneous subsystems. Among other miscellaneous spacecraft subsystems, the pyrotechnics subsystem includes electrically initiated devices, called squibs, that are used to deploy components such as booms, explode bolts that separate or jettison hardware such as descent probes or launch adapters, ignite solid rocket motors, and operate some types of valves.

Redundancy. As noted above, redundancy traditionally has been a prominent feature of planetary spacecraft, which typically carry two transmitters, two receivers, two command computers, and so on. In the event of a mishap, the spacecraft can switch to a backup system and carry on; this has happened many times. In an effort to reduce spacecraft mass and complexity, some designers have considered dropping redundancy and allowing spacecraft to operate on a single string of components. Although this may save mass, it also usually means a spacecraft cannot function if any of its principal subsystems fails.

Mission Stages

The process of conceiving, designing, building, launching, and operating a spacecraft mission takes many years. *Voyager*, for example, was conceived during the late 1960s and given formal project approval—called a "new start"—in 1972 by the National Aeronautics and Space Administration (NASA). Five years later, the twin *Voyager 1* and 2 were launched, and executed their planetary flybys between 1979 and 1989. In the 1990s and the first decade of the twenty-first century, they continued to return data as they left the solar system. Efforts have been made to simplify missions and to shorten the lead time required to design and launch spacecraft, but prelaunch activities still involve a minimum of 2–3 years even for modest missions.

In the United States, the programmatic steps leading to preparation of a mission are commonly designated by a series of terms. A conceptual study is conducted, which outlines the general scheme of a mission concept. If approved, this proceeds to a phase A study, also called a preliminary analysis, which creates a tentative design and project plan outlining the mission. This specifics what to build, when to launch, what course the spacecraft is to take, what is to be done during cruise, when the spacecraft will arrive at its destination, and what operations will be carried out. The project then moves into phase B, also called the definition phase, in which more formal specifications and requirements are determined. If the project wins formal approval, it progresses to phase C/D, or design and development, where the spacecraft is designed, assembled, tested, and launched. The process concludes with the mission's operations phase, phase E, including its interplanetary cruise and eventual flyby, orbit insertion, or landing. If the spacecraft is healthy at the conclusion of its specified mission period, it may be approved to continue into an extended mission.

Throughout the design process, mission planners must be sensitive to various constraints. These include the project's overall acceptable budget,

which will influence many decisions on spacecraft and mission complexity. Another constraint is the availability of launch vehicles, which in turn will dictate the maximum spacecraft mass and feasible trajectories. The launch period during which planetary geometry is appropriate to carry out the mission must also be taken into account. The launch period for a transfer orbit to Mars, for example, occurs once every 26 months. Contention for available resources must also be considered, such as tracking time on the giant dish antennas of the U.S. Deep Space Network used to communicate with deepspace probes. *See* SPACECRAFT GROUND INSTRUMEN-TATION.

Launch. Virtually all planetary launches in the United States take place at the Cape Canaveral Air Force Station in Florida, which is situated relatively close to the Equator with an eastward launch range extending over the Atlantic Ocean. Launches conducted by Russia take place at the Baikonur Cosmodrome in Kazahkstan. In the 1980s, the United States and the Soviet Union were joined by Japan and the European Space Agency in launching solar system probes. Japan's launch facility is at the Uchinoura Space Center, near the southwest tip of the island of Kyushu. European launches take place at the Center Spatial Guyanais at Kourou, French Guiana, on the northeast coast of South America. *See* LAUNCH COMPLEX.

United States space probes traditionally have been launched on such vehicles as the Atlas, Delta, and Titan rockets. In 1989 and 1990, the *Magellan*, *Galileo*, and *Ulysses* spacecraft were launched from the payload bay of the United States' space shuttle, using specially adapted solid-fuel upper-stage engines to inject them on their interplanetary trajectories. The Russian launch fleet includes the liquidfuel Proton. Japan has launched planetary missions on its solid-fueled Mu-5, while Europe has developed the Ariane 5. *See* ROCKET PROPULSION; SPACE SHUTTLE.

After a spacecraft arrives as the launch site, it is generally tested to ensure that all systems are working normally after being transported from the facility where it was assembled. The probe is then mated to its upper-stage engine, and the combined stack is placed atop the launch vehicle. The probe's propulsion tanks are filled with fuel and pressurant. After launch, the main rocket usually places the probe and upper stage in a relatively low orbit around Earth. The upper stage then fires to send the probe on its interplanetary trajectory. At that point, pyrotechnic squibs may be fired to deploy on-board appendages, and the entire spacecraft goes through a checkout phase.

Cruise. Many activities take place while the spacecraft is en route to its destination planet or other target. Some scientific instruments, such as those studying fields and particles, may be fully powered to collect data during the interplanetary transit. Other instruments used only at the target body may be switched off. It is usually necessary to conduct one or more trajectory correction maneuvers in which on-board thrusters are fired to fine-tune the spacecraft's flight path.

Although it is common to think of a spacecraft trajectory as a simple path from Earth to the target body, the fact that Earth and all other bodies are moving around the Sun makes trajectory design more complex. To travel from Earth outward to Mars, for example, using the least energy possible, mission planners usually take advantage of a Hohmann transfer orbit. This involves adjusting the existing orbit around the Sun so that it is more elliptical. Under the new orbit, the spacecraft's perihelion, or closest approach to the Sun, will coincide with the orbit of Earth, while its aphelion, or farthest point from the Sun, will coincide with the orbit of Mars. In order to arrive at a time when Mars is at the same point as the spacecraft, it must be launched during an appropriate time window.

Beginning in the 1970s, mission planners took advantage of a technique called gravity assist trajectories to send spacecraft on to more than a single planetary target. In this technique, the spacecraft passes within a few hundred miles of a given planet; this accelerates (or decelerates) the spacecraft and changes its trajectory. (Because no energy is ever created or lost, the encounter also results in the planet being slowed, or speeded, by an immeasurably small amount in its orbit around the Sun.) Gravity assist was first used by Mariner 10 in 1973, when the spacecraft used the gravity of Venus to send it on to three flybys of Mercury in 1974 and 1975. The technique was then used very extensively by Voyager to send spacecraft past Jupiter, Saturn, Uranus, and Neptune. In the 1990s, gravity assist became useful to send heavier spacecraft to destinations that would not be possible via direct trajectories by using available launch vehicles. Galileo, for example, flew by Venus once and Earth twice on its way to Jupiter. See SPACE NAVIGATION AND GUIDANCE.

Arrival. Activities at the target body vary, depending on whether the spacecraft executes a flyby, goes into orbit, or lands on the body's surface. Flyby activities are generally divided into a far-encounter period, which may extend for several weeks, and a close encounter, which includes the most hectic days directly surrounding the spacecraft's closest approach to its target body.

Spacecraft designed to go into orbit around the target body must fire a large engine to brake their velocity sufficiently so that the target body's gravity can capture them. In a typical mission scenario, the spacecraft may enter an initial capture orbit which may be highly elliptical; over a period of several weeks, thruster firings will be used to circularize the orbit so that the spacecraft can carry out mapping operations. In addition, it has become routine for Mars orbiters to circularize their orbits after arriving at the planet by using the drag of Mars's atmosphere to reduce the orbit's apoapsis, or high point, over time. This technique is called aerobraking. Most spacecraft are placed into a polar or fairly highly inclined orbit at a planet to maximize the amount of surface that can be studied. Magellan was placed in
a near-polar orbit at Venus that caused the planet to rotate once beneath it every 243 days. In the *Mars Reconnaissance Orbiter* mission, the spacecraft was placed in a Sun-synchronous orbit that took it over regions uniformly at about 3 p.m. Mars local time. In this orbit type, the spacecraft's orbit plane precesses with nearly the same period as the planet's solar orbit period. This is useful for many scientific instruments because it makes the incidence angle of sunlight on surface features uniform from orbit to orbit.

Landing a spacecraft on a planet requires even greater deceleration than in the case of orbit insertion. Spacecraft approaching planets with an atmosphere, such as Mars, usually take advantage of the technique of aerobraking, or using friction with the atmosphere, to slow the spacecraft. An aeroshell is usually required to protect the spacecraft from the effects of heating. After jettisoning the aeroshell, the spacecraft may use a propulsion system to achieve a soft landing, perhaps augmented by a parachute system. Although the Soviet Union's Luna spacecraft returned soil samples from the Moon during the 1970s, sample return missions have not yet been mounted to other planets, although samples of cometary dust and solar and interstellar particles have been brought to Earth. Japan's Hayabusa spacecraft apparently failed to collect a sample when it landed on the asteroid Itokawa in 2005, but probably obtained some dust from the asteroid that will be carried to Earth in 2010.

Significant Missions

Soon after the Soviet Union launched the Earthorbiting *Sputnik* in 1957 and the United States responded with *Explorer 1* in early 1958, both countries turned to developing probes that would escape



Fig. 1. Model of *Luna 1* (Mechta). Long metal strips are antennas, the tapered post is the magnetometer boom, and raised windowlike apertures are ion detectors.



Fig. 2. Full-scale model of the *Pioneer 3* and *4* space probes, showing the antenna, the paint stripes for thermal control, and the despin device about the major diameter. (*JPL/NASA*)

Earth orbit and travel to the Moon and beyond.

Luna. The first successful mission to escape Earth orbit was the Soviet Union's Luna 1 (Fig. 1), which executed a 3728-mi (5275-km) lunar flyby after launch on January 2, 1959. Later that year, the Soviet Union launched Luna 2, which made the first lunar impact, and Luna 3, which studied the far side of the Moon. Luna 4, launched in 1963, missed the Moon by 5282 mi (8500 km). The Soviets had a string of failures in 1965 in attempting a lunar soft landing; Luna 5 failed when the spacecraft made a hard impact on the Moon, Luna 6 missed the Moon by 100,000 mi (160,000 km), and Luna 7 and 8 failed because retrorockets fired too early or late. Luna 9 made the first lunar soft landing on February 2, 1966, and returned photographs from the Moon's surface. Luna 10, 11, 12, 14, 15, and 19 were lunar orbiters, while Luna 13 was a soft lander. In 1970 the Soviet Union launched Luna 16, which carried out that country's first successful lunar sample return. Later the same year, Luna 17 made a soft landing on the Moon with a robotic rover vehicle. Luna 18 was intended as a sample-return mission but failed during lunar descent. Luna 20 in 1972 executed a successful sample return. The series continued with Luna 21, a lander and rover, in 1973, and Luna 22, an orbiter, in 1974; and concluded with Luna 23 and 24, sample return missions conducted in 1974 and 1976 (Luna 24 was successful, while Luna 23 was not). Much information about the missions was

withheld by the Soviet government for many years, but some details have become available since the dissolution of the Soviet Union in the early 1990s. *See* MOON.

Pioneer. This name was given to a family of largely dissimilar United States spacecraft used mainly for interplanetary physics, although some later missions in the series achieved success as planetary probes. *Pioneer 1, 2,* and *3* were failed lunar probes in 1958, although *Pioneer 1* and *3* returned data on the environment of space before reentering Earth's atmosphere. *Pioneer 4* (**Fig. 2**) passed within 37,300 mi (60,000 km) of the Moon in 1959, followed by *Pio*-

neer 5 in 1960, which relayed solar system data to a distance of 22.5×10^6 mi (36.2 $\times 10^6$ km). Following a hiatus of 5 years, *Pioneer 6* studied the space environment in solar orbit, as did *Pioneer 7* in 1966, *Pioneer 8* in 1967, and *Pioneer 9* in 1968.

The series took a very different direction with *Pioneer 10* and *11*, the first space probes sent into the outer solar system in 1972 and 1973, respectively. *Pioneer 10* executed a flyby of Jupiter in December 1973, while *Pioneer 11* flew by Jupiter in December 1974 and Saturn in 1979. In 1978 the series took yet another departure with *Pioneer Venus* (**Fig. 3**), which used two launches to send an orbiter and



Fig. 3. Expanded views of the structure of (a) the large probe and (b) one of the smaller probes launched into the Venus atmosphere from the Pioneer Venus bus. (After L. J. Nolte and S. C. Sommer, Probing a Planetary Atmosphere: Pioneer-Venus Spacecraft Description, 1975)



Fig. 4. Partial schematic of *Surveyor* to show part of the propulsion system associated with the vernier engines which provided midcourse trim and attitude control during firing of the large retrorocket, and final velocity removal near the lunar surface. (*JPL/NASA*)

multiple atmospheric probes to Venus. *See* JUPITER; SATURN.

Ranger. In the early 1960s, the United States carried out this series of missions to relay photographs of the Moon as the spacecraft descended to a crash landing. *Ranger 1* and 2 ended in launch vehicle failures; *Ranger 3* missed the Moon by 22,862 mi (36,793 km); *Ranger 4* impacted the Moon but experiments proved inoperative when a timer failed; *Ranger 5* missed the Moon by 450 mi (725 km); and *Ranger 6* impacted the Moon but its television camera failed. *Ranger 7* was the first success, returning 4308 photos during its lunar descent in July 1964. In 1965 *Ranger 8* and 9 sent back 7137 and 5814 photos, respectively.

Surveyor. These United States missions (**Fig. 4**) soft-landed robot spacecraft on the Moon as precursors to the crewed Apollo missions. *Surveyor 1* landed on the Moon on June 2, 1966, and returned 11,150 photos as it operated for 6 weeks. *Surveyor 2* suffered from a vernier engine failure and crashed into the lunar surface. *Surveyor 3* carried a soil sampler for its soft landing in 1967. *Surveyor 4* touched down on the Moon, but its signal was lost $2^{1}/_{2}$ min later. *Surveyor 5* soft-landed in 1967, and relayed 19,000 photos and conducted soil analyses. *Surveyor 6* executed the first takeoff from the Moon during its

1967 mission. *Surveyor* 7, which soft-landed in 1968, was the last in the series.

Lunar Orbiter. These spacecraft launched during the same era as the *Surveyors* were the first attempt to make detailed maps of the Moon. *Lunar Orbiter 1* was launched August 10, 1966, and relayed photos for the next 20 days. *Lunar Orbiter 2* sent 205 photographic frames later that year; *Lunar Orbiter 3* sent 182 photos in 1967, followed by *Lunar Orbiter 4* and 5.

Mariner. This designation was given to the United States' early planetary probes, which chiefly studied Venus and Mars, although the final in the series also visited Mercury. Mariner 1 was lost in 1962 after a typographical error in a software program caused the launch vehicle to veer off course. Mariner 2 became the first spacecraft to encounter another planet when it flew by Venus at a distance of 21,594 mi (34,752 km) later that year. It continued to relay data to a distance of 53.9×10^6 mi (86.7 × 10⁶ km). Mariner 3 and 4 (Fig. 5a) targeted Mars; the first experienced a shroud failure, while the second became the first spacecraft to encounter Mars in 1964. Mariner 5 (Fig. 5b) executed a 2480-mi (3991-km) Venus flyby in 1967. Mariner 6 and 7 (Fig. 5c) again targeted Mars, sending back 75 and 126 television pictures, respectively, during flybys at a distance of about 2150 mi (3450 km) in 1969. Mariner 9 was a Mars orbiter, sending back detailed pictures after its launch in 1971. Mariner 10 was the first spacecraft to use a gravity-assist trajectory, encountering Venus before going on to three flybys of Mercury in 1974 and 1975. The United States then planned a mission with the working title Mariner Jupiter-Saturn, but the name was changed to Voyager before launch. See MARS; MERCURY (PLANET); VENUS.

Venera, Mars, and Zond. These missions were the Soviet Union's equivalent of the Mariners. The Soviets had the greatest success with Venera, which lofted a number of spacecraft to Venus over more than two decades. Although the first three Venera failed in 1961 and 1965, Venera 4 transmitted during its 94-min descent, though it failed at a considerable altitude above to Venus's extremely hot surface. The series continued with Venera 5 and 6 in 1969. Venera 7 made the first soft-landing of Venus in 1970 and transmitted for 23 min, which was remarkable considering the planet's 900°F (500°C) surface temperature. Venera 8 made another successful landing in 1972. In 1975 the Soviet Union launched Venera 9 and 10, which combined Venus orbiters and landers; Venera 9 returned the first pictures from the surface of another planet. In 1978 and 1981 the landers Venera 11, 12, 13, and 14 followed. Venera 15 and 16, launched in 1983, were orbiters that used syntheticaperture radar to pierce Venus's opaque atmosphere and map the planet's surface.

The Soviets' Mars program achieved more mixed results. After the failure of *Mars 1* in 1962, the Soviets attempted to make soft landings in 1971 with *Mars 2* and *3*. The first failed when it evidently impacted the planet's surface with excessive force; the

second touched down but failed 110 s later, having transmitted only a partial photograph that revealed no detail. In 1973 the Soviets made another attempt with a four-spacecraft armada, the orbiters *Mars 4* and 5 and the landers *Mars 6* and 7. *Mars 4* and 7 missed the planet entirely, while *Mars 6* stopped transmitting during its descent. *Mars 5* entered orbit, but communications proved to be erratic.

Zond was the designation given to several Soviet missions with various targets from 1964 through 1970. Zond 1 and 2 were failed Venus and Mars probes, respectively, while Zond 3 returned photos from a lunar flyby in 1965. Zond 5 flew around the Moon and returned to Earth, splashing down in the Indian Ocean in 1968. In addition to other objectives, this and later Zond missions carried tortoises and other organisms for biological research. Zond 6 flew around the Moon before returning to crash in the Soviet Union. Zond 7 also was quite successful. Zond 8 performed another lunar return in 1970.

Viking. One of the most ambitious planetary explorations ever mounted was the United States' *Viking* program. Two launches of combined orbiter-landers, *Viking 1* and 2, took place in August and September 1975, with arrival at Mars in June and August 1976. *Viking*'s chief purpose was the search for biological activity on the Martian surface. Both the orbiters and landers also carried out extensive imaging; the orbiters reconnoitered landing sites in preparation for descent of the landers and also carried out a global mapping program. The landers (**Fig. 6**) sent back the first extensive detailed pictures from the surface of another planet.

The spacecraft failed to detect any sign of life on the planet. Many scientists, however, later concluded that *Viking*'s search methodology was very narrow, and began to devise other ways to search for past or present life on Mars. The *Viking* orbiters revealed a landscape that appeared to have been heavily refigured by liquid water, but it appeared for many years that none exists currently on the frigid planet. In 2006, images taken by *Mars Global Surveyor* provided evidence suggesting that on current-day Mars water has broken out of cliffs or crater walls and run downhill before evaporating. Mar's polar caps have been found to be a mixture of carbon dioxide ice and water ice.

Helios. This is a German-built and United Stateslaunched solar system probe which because of its close approach to the Sun (0.29 AU) is called a solar probe. *Helios 1* was launched in 1974, followed by *Helios 2* in 1976. The experiments, directed toward phenomena taking place in the solar wind, were chiefly magnetometers and particle detectors. *See* SOLAR WIND.

Voyager. In the late 1960s, mission designers began to contemplate a coming opportunity that would not be repeated for another 177 years. In the 1970s it would be possible to launch a spacecraft that could sequentially visit all four giant gaseous outer planets—Jupiter, Saturn, Uranus, and Neptune—by using the technique of gravity assist to send the probe on to its next destination.



Fig. 5. Evolution of the *Mariner* spacecraft. (a) *Mariner 4*. (b) *Mariner 5*. Active thermal louvers, shown on the sides of the main body, are driven by temperature-sensitive motors which control the degree of opening, and thus the heat-emission rate from the spacecraft. Many changes from *Mariner 4* are not seen in this gross view from the outside. (c) *Mariner 6* and 7. Although only about 300 lb (135 kg) heavier than *Mariner 4* and 5, this type of spacecraft is much more complex because of advances in electronics. (NASA)



Fig. 6. Viking lander, with scientific experiments. (Viking Press Kit, NASA News Release 75-183, 1975)



Fig. 7. Voyager spacecraft showing primary subsystems. 1 m= 3.3 ft. (NASA)

Efforts culminated with the launch in 1977 of Voyager 1 and 2 (Fig. 7), which began a decade-long "grand tour" of the outer solar system. Although Voyager 2 was launched first, Voyager 1 took a shorter trajectory to Jupiter and arrived in 1979; it went on to encounter Saturn in 1980. Voyager 1's trajectory at Saturn was optimized to study the planet's major satellite, Titan, which was found to contain complex organic chemistry in its opaque atmosphere. This flight path precluded the spacecraft from later planetary encounters. Voyager 2 visited Jupiter in 1979 and Saturn in 1981, then continued on to flybys of Uranus in 1986 and Neptune in 1989. During the 1990s and the first decade of the twenty-first century, both spacecraft continued to relay back fields and particles data as they exited the solar system. See NEPTUNE; URANUS.

Comet missions. The arrival of Comet Halley to the inner solar system in the mid-1980s prompted a great deal of international interest in sending spacecraft to meet it. The Soviet Union and Japan each dispatched two spacecraft, and the European Space Agency a single craft. The efforts were notable for a high degree of international cooperation, including the use of data obtained by the NASA Deep Space Network from the early-arriving Soviet craft to target more precisely the flyby of the European craft. *See* COMET; HALLEY'S COMET.

Plans originally called for a United States spacecraft to join the Halley armada, but that mission was the victim of a budgetary cancellation. United States mission planners developed an ingenious scheme to redirect the *International Sun-Earth Explorer 3 (ISEE 3)* satellite to conduct a flyby of the tail of Comet Giacobini-Zinner on September 11, 1985, about 6 months before the Comet Halley flybys.

International Comet Explorer (ICE). This was the new name given the ISEE 3 (Fig. 8) when it was redirected to make its comet visit. The craft was originally placed in a halo orbit, a conditionally stable circular orbit with its center at the Lagrange point upstream (toward the Sun) from the Earth-Moon system. In 1982 and 1983 a series of maneuvers in the Earth-Moon system culminated in a final close flyby of the Moon that sent the craft on a new trajectory to fly through the tail of Comet Giacobini-Zinner, some 6000 mi (10,000 km) from the comet's nucleus. The craft's payload was a half-dozen instruments probing plasma electrons, plasma ions, energetic protons, magnetic fields, plasma waves, and radio waves. See CELESTIAL MECHANICS.

Suisei and Sakigake. These two spacecraft represented the first solar system exploration mission launched by the Japanese Institute of Space and Astronautical Science (ISAS). *Sakigake* was launched in January 1985 and flew within a few million miles of Comet Halley on March 11, 1986. The spin-stabilized craft carried a solar-wind ion detector, a magnetometer, and a plasma-wave probe.

The *Suisei* craft (**Fig. 9**), also called *Planet-A*, was launched in August 1985 and approached within about 90,000 mi (150,000 km) of Comet Halley on March 9, 1986. Like its sibling craft, *Suisei* was spin-



Fig. 8. International Sun-Earth Explorer 3 (ISEE 3), the last satellite of a coordinated series of three designed to study the Earth's magnetosphere. This satellite was redirected toward Comet Giacobini-Zinner by a lunar encounter, and was renamed International Cometary Explorer (ICE). (ISEE Science Working Team)

stabilized, but it carried a different science payload: an ultraviolet camera and a solar-wind experiment.

VEGA. This enterprising mission by the Soviet Union sent a pair of spacecraft on flybys of Venus, where each released a lander and an atmospheric balloon, before continuing on to encounter Comet Halley (**Fig. 10**).

VEGA 1 and 2 were launched in December 1984 and flew by Venus in June 1985. VEGA 1's Halley encounter took place March 6, 1986, while VEGA 2's followed 3 days later. Each craft carried optical cameras, spectrometers, dust detectors, and fields and particles instruments. Unlike the Japanese and European craft, the VEGAs were three-axis-stabilized.

The VEGAs achieved a closer flyby distance from Halley than did the Japanese probes, encountering the comet at distances of 5523 and 4990 mi (8889 and 8030 km). They were able to provide highly useful targeting data for the European probe, which was to make by far the closest pass by the comet.

Giotto. This spacecraft (**Fig. 11**) was the European Space Agency's representative in the Comet Halley fleet. At its closest approach, *Giotto* came within about 360 mi (600 km) of the comet.

The spin-stabilized craft's payload included a camera, spectrometers, a photopolarimeter, dust experiments, and fields and particles instruments. Launch took place in July 1985, with the Halley flyby on March 14, 1986.

Because the environment near the comet's nucleus was known to be very dusty, *Giotto* was designed to endure a high number of particle impacts. Although only 2–3% of the spacecraft's solar-panel capacity was lost during the encounter, the mirror on the camera began to suffer degradation when



Fig. 9. Model of the Japanese comet probe *Suisei* (*Planet-A*). (a) Assembled spacecraft with the solar cells assembly, high-gain antenna, and ports for instruments. (b) Inside view of the spacecraft with instruments arranged cylindrically about the spin axis. The extensive foil is for thermal protection and heat control. (*Institute of Space and Astronautical Science*)

Giotto was inbound at a distance of about 3000 mi (5000 km), and the last usable image was taken at a distance of about 1220 mi (1963 km). In addition, particle impacts caused the craft to wobble with a nutation of 0.8° , resulting in a loss of radio signal 14 s before closest approach. The wobble was overcome, however, about 32 min later, and *Giotto* remained a highly functional probe years after the encounter.

Phobos. An ambitious project for the exploration of Mars and its natural satellite, Phobos was undertaken by the Soviet Union in the late 1980s. Unfortunately, the two spacecraft were not to fulfill completely their planned activities.

Phobos 1 and 2 were launched 5 days apart in July 1988. The two nearly identical craft were to arrive at Mars in January 1989 and begin orbiting the planet. A series of maneuvers would have taken the craft near Phobos, where they would approach for a surface-contour-following flyby about 150 ft (50 m) from the rocky object's surface. The craft would also have deployed a 110-lb (50-kg) long-duration lander, expected to operate for about a year, and a 112-lb (51-kg) hopper, which would have moved about the surface of Phobos in 60-ft (20-m) leaps. Altogether, the craft included a total of 37 scientific experiments.

A series of commands sent to *Phobos 1* in August 1988 apparently contained a coding error that instructed the spacecraft to turn off its attitude-control system. The craft then began to tumble slowly; as its solar panels were no longer pointed toward the Sun, it lost power and all contact was permanently lost.

Problems arose with *Phobos 2* by the end of 1988 when its television system malfunctioned en route to Mars. Those difficulties were resolved by ground command. More serious was the loss of the on-board high-power radio transmitter, which would have permitted a high rate of data transmission. It was still possible, however, to use the slower low-power transmitter to send data to Earth.

Phobos 2 survived the transit of Mars, entered orbit in late January 1989, and began its scientific surveys of the red planet. By late March, the craft was positioned near the satellite Phobos, capturing high-detail images from a distance of several hundred miles. On March 27, 1989, however, a failure in the attitude control system caused the craft to begin tumbling erratically in an orbit just outside that of Phobos. Ground controllers spent a week trying to resume communications before giving up the craft as lost.

Magellan. In the late 1970s and the 1980s, improvements were made in synthetic-aperture radar, a technology that uses radar pulses to capture images where optical systems are not practical. This technology is ideal for the study of the planet Venus, which is perpetually shrouded by a thick layer of opaque clouds. *See* SYNTHETIC APERTURE RADAR (SAR).

The first craft to perform radar imaging at Venus, *Pioneer Venus* of the United States was followed by the Soviet *Venera 15* and *16* probes, which were able to improve resolution by about a factor of 10. They were followed by the United States' *Magellan*



mission, which in turn achieved a factor-of-10 improvement over the resolution delivered by the Soviet missions.

Launched in May 1989, *Magellan* (Fig. 12) was also the first of several solar system probes to be lofted by the United States space shuttle. The probe was carried into Earth orbit in the cargo bay of the shuttle and released. A solid-fuel upper-stage engine attached to the probe then injected it into its interplanetary trajectory.

Magellan entered Venus orbit in August 1990, initiating a highly successful 4-year study of the planet. Each 8 months, Venus rotated once underneath the spacecraft as it passed overhead in a polar orbit. Magellan spent several of these 8-month orbital cycles producing radar maps of 98% of the Venusian surface with a resolution of 1000 ft (300 m) before compiling a high-resolution, comprehensive gravity field map for 95% of the planet. In May 1994, ground controllers experimented with aerobraking, making use of the planet's atmosphere to lower the spacecraft's orbit. With solar panels and other components gradually failing, mission controllers decided to plunge the spacecraft into Venus's atmosphere to study the craft's behavior and atmospheric characteristics. Contact was lost with Magellan on October 12, 1994, as the craft presumably descended and broke apart.

Galileo. After the initial reconnaissance of the solar system performed by the *Mariners* and *Voyagers* from the 1960s through the 1980s, mission planners



Fig. 11. Cutaway view of the *Giotto* comet-encounter spacecraft showing the primary subsystems. (*European Space Agency*)



Fig. 12. Magellan spacecraft. (a) Undergoing testing. (b) Deployed, showing primary subsystems. (JPL/NASA)

began to consider the next step in planetary exploration. Their plans called for a series of orbiters that would return to each of the planets visited by flyby spacecraft to conduct detailed studies. The first such approved mission was *Galileo*, which would orbit the solar system's largest planet, Jupiter.

Galileo (Fig. 13) had a complex dual-spin design in which most of the spacecraft was spin-stabilized to optimize fields and particles experiments, while the platform that carried pointed instruments such as cameras was capable of being three-axis-stabilized. The orbiter also carried a descent probe designed to relay data as it dived into the giant planet's atmosphere. En route to Jupiter, *Galileo* performed a flyby of Venus and two flybys of Earth. It also performed the first-ever flyby of an asteroid when it encountered the rocky body Gaspra in October 1991. In August 1993, *Galileo* flew by a second asteroid, Ida. When images were sent to Earth from the craft's on-board tape recorder 6 months later, the first-ever natural satellite of an asteroid was discovered when the tiny moon Dactyl was found to be orbiting Ida. *See* ASTER-OID.

The failure of *Galileo*'s umbrella-like high-gain antenna to open fully prompted an extensive redesign of planned observations at Jupiter. After its arrival in December 1995, the mission's highest priority was





Fig. 13. Galileo spacecraft. (a) Being tested. (b) Deployed, showing primary subsystems. The scan platform contains the photopolarimeter radiometer, near-infrared mapping spectrometer, solid-state imaging camera, and ultraviolet spectrometer. (JPL/NASA)

to relay data from the atmospheric probe during its descent. *Galileo* then began sending selected images and other scientific data that had been compressed in on-board computers to make the best use of the craft's available low-gain antennas.

Galileo made about 30 orbits of Jupiter in the first 5 years after its arrival. After completing an initial $2^{1}/_{2}$ -year prime mission, *Galileo* was approved for an extended mission focusing on the Jovian satellites Europa and Io. The mission was extended a second time to allow the spacecraft to focus on observations of Jupiter's largest satellite, Ganymede, and on collaborative studies with Cassini as that spacecraft passed Jupiter in late 2000 en route to Saturn. Galileo discovered strong evidence that Europa has a melted saltwater ocean under an ice layer on its surface. The spacecraft also found indications that Ganymede and another Jovian satellite, Callisto, have layers of liquid saltwater as well. Other major science results from the mission include details of varied and extensive volcanic processes on the moon Io, measurements of conditions within Jupiter's atmosphere, and the discovery of a magnetic field generated by Ganymede.

On September 21, 2003, *Galileo* was intentionally plunged into Jupiter's atmosphere to avoid the risk of impacting and possibly contaminating Europa.

Hiten and Hagoromo (Muses-A). After their success with its two Comet Halley craft, mission planners in Japan turned to other projects they could execute as the nation's space program developed. Their next

effort was Muses-A, a mission to study the Moon.

An expendable rocket launch in January 1990 carried aloft two spacecraft, *Hiten* and *Hagoromo*. The latter, a basketball-sized package carrying an instrument payload, was to go into orbit around the Moon. The larger *Hiten* was to occupy a highly elliptical orbit around Earth, ranging in distance between 19,000 and 470,000 mi (30,000 and 760,000 km). This orbit would take *Hiten* out toward the Moon, where it would receive and relay to Earth signals from *Hagoromo*. As *Hagoromo* approached the Moon, however, its radio transmitter failed, ending all contact with the craft.

Ulysses. By the 1980s, the Sun had been studied for many years by ground-based equipment. Missions such as the *Orbiting Solar Observatory* satellites and the crewed *Skylab* in the 1970s extended these observations to Earth orbit. All observing equipment, however, was limited to the two-dimensional surface of the ecliptic, the plane in which most planets orbit the Sun.

Ulysses was designed to extend that point of view by sending a spacecraft out of the ecliptic and into the unexplored regions above and below the Sun's poles. After collaborative work between NASA and the European Space Agency and the budgetary cancellation of a United States craft that was to have taken part, a project emerged wherein the European agency built and provided the single *Ulysses* spacecraft (**Fig. 1**4) and contributed a computer system



Fig. 14. Ulysses spacecraft, showing primary subsystems. The propellant storage tank is part of the reaction control equipment (for attitude and so forth), and the thruster cluster is essentially a set of small engine nozzles for this equipment. The preamplifier is part of the unified radio and plasma wave experiment. (*European Space Agency*)

to control the craft from the ground. NASA provided some instruments, the radioisotope power source, launch by a United States space shuttle, tracking via its global Deep Space Network, and ground facilities at the Jet Propulsion Laboratory in Pasadena, California.

Launched in October 1990, *Ulysses* flew first to Jupiter, where the giant planet's gravity sent the craft out of the ecliptic in February 1992. The encounter of the Sun's south pole took place in 1994, with an encounter of the north pole in 1995. As the mission progressed, one of the key scientific findings revealed a Sun with a homogeneous magnetic field lacking latitudinal variations, very different from the magnetic climate observed by ground-based instruments. *Ulysses*'s observations continued after its flight over the Sun's poles, and it made another passage over the Sun's polar regions under conditions of high solar activity in 2000 and 2001. In 2006 it began its third passage over the Sun's poles.

Mars Observer. This mission was to be the United States' first return to the red planet since the *Viking* effort of the 1970s. A polar orbiter, *Mars Observer* carried a high-resolution camera and other instruments to study Mars's geology and atmosphere. The spacecraft was adapted from an assembly-line commercial Earth-orbiting satellite in an effort to streamline design and save costs.

Launched on a Titan 3 rocket in September 1992, the mission proceeded according to plan until 3 days before Mars arrival in August 1993, when all contact was lost with the craft. The loss took place during a period in which *Mars Observer*'s propulsion system was being pressured in preparation for the firing of its main engine to place it on orbit; the craft's radio transmitter was intentionally turned off at the time to protect it from any forces during the pressurization. A review panel concluded that a failure probably occurred in the propulsion system that either ruptured a fuel line or caused an explosion.

Clementine. Unlike most United States missions, carried out by NASA, this novel effort was the work of the Department of Defense. *Clementine*'s flight plan called for it to orbit Earth for several days before departing for the Moon, where it would spend more than 2 months in orbit capturing high-resolution images of the lunar surface. The spacecraft was then to leave once more on a trajectory that would take it within 60 mi (100 km) of the asteroid 1620 Geographos.

After launch in January 1994, *Clementine* (Fig. 15) arrived in lunar orbit in February and carried out systematic mapping; on May 5, it left lunar orbit. Two days later, a malfunction of the on-board computer caused several of the craft's thrusters to be activated during a 20-min period when *Clementine* was out of touch with ground controllers. This depleted all fuel from the craft's tanks, preventing the asteroid flyby.

Although it returned a wealth of lunar imagery, the chief objective of *Clementine* was not scientific, but testing 23 advanced technologies of interest to the military. Moreover, an aim was to demonstrate a streamlined project management approach that would make it possible to build and launch a small yet highly capable spacecraft cheaply in less than 2 years.

Cassini. Considered the last of the its era of large United States planetary exploration missions, *Cassini* (**Fig. 16**) was launched in 1997 to study the ringed planet Saturn and its major natural satellite, Titan. The *Cassini* orbiter carried an instrumented probe provided by the European Space Agency, *Huygens*, designed to descend to the surface of Titan. Shrouded by an opaque atmosphere containing a rich diversity of organic compounds, Titan possibly contains liquid oceans of methane. Like *Galileo*, the craft has made use of multiplanet gravity assists, flying by Venus, Earth, and Jupiter on the way to its destination.

The *Cassini* and *Galileo* spacecraft made coordinated observations of Jupiter on *Cassini*'s flyby of the giant planet on December 30, 2000. *Cassini* data were used to track daily changes of some of the planet's most visible storms and to study patterns in natural radio emissions near the edge of Jupiter's magnetic environment.

After 7 years in transit, *Cassini* entered orbit at Saturn on July 1, 2004. Six months later, on January 14, 2005, the *Huygens* probe took photos as it descended to the surface of Titan, where it touched down and continued to send data for about 90 min. Preliminary findings confirmed that the landing site is near the shoreline of a liquid ocean.

The *Cassini* orbiter, meanwhile, continued on numerous close approaches to Saturn's moons and studies of its atmosphere, rings, and magnetic field. One notable early science finding was the discovery that Saturn's moon Enceladus may have liquid water reservoirs that erupt in geysers.



Fig. 15. *Clementine* spacecraft being prepared for testing at Naval Research Laboratory, Washington, DC.



Fig. 16. Cassini spacecraft, showing primary subsystems. (JPL/NASA)

Discovery Programs. Faced with constrained budgets but increasingly complex and costly planetary spacecraft, NASA in the mid-1990s undertook a conscious program of downsizing missions. At the center of this effort was the Discovery Program, designed to sponsor low-cost solar system missions with highly focused science objectives. The program periodically makes competitive selections from proposals submitted by teams led by scientific principal investigators. Early selected missions included *Near Earth Asteroid Rendezvous, Mars Pathfinder, Lunar Prospector, Stardust, Comet Nucleus Tour, Messenger*, and *Deep Impact*.

Near Earth Asteroid Rendezvous (NEAR). The *NEAR* probe was launched in 1996 to rendezvous with and orbit the asteroid Eros, one of the small rocky bodies that pass relatively near Earth. After passing asteroid 253 Mathilde in 1997, the probe was scheduled to enter orbit around Eros in late 1998. Due to a problem with flight software, however, the orbital insertion engine firing did not take place correctly, and *NEAR* sailed past Eros. However, the craft was able to try again about a year later after circling once around the Sun. *NEAR* correctly entered orbit around Eros on February 14, 2000, and began a series of maneuvers to take it close to the surface of the small rocky

body. The spacecraft, renamed *NEAR-Shoemaker* in honor of the late astronomer Eugene Shoemaker, gently landed on asteroid Eros on February 12, 2001, after orbiting the rocky body for a year. It captured 69 images during the final 3 mi (5 km) of its descent, revealing features as small as 0.4 in. (1 cm) across. The final signals from the spacecraft before it was shut down were received February 28, 2001.

Lunar Prospector. One of the smallest of the early Discovery missions, *Lunar Prospector* was launched in 1998 and reached the Moon after 5 days in flight. Orbiting 60 mi (100 km) above the lunar surface, the probe carried five instruments designed to help understand the origin and evolution of the Moon and to determine whether water ice may be present in shaded areas near the Moon's poles. After a year and a half in orbit, *Lunar Prospector* was intentionally crash-landed in a perpetually shadowed crater near the Moon's south pole; scientists hoped that the impact would throw off water vapor detectable from ground- and space-based observatories. No visible debris plume was detected, however.

Stardust. Launched in 1999, this probe (**Fig. 17**) was designed to fly through the cloud of dust that surrounds the nucleus of a comet and, for the first time, collect cometary material for return to Earth.



Fig. 17. Stardust spacecraft, showing primary subsystems. (JPL/NASA)

The spacecraft had a pair of tennis-racket-shaped collectors using a substance called aerogel to snag minuscule dust particles. During the course of the mission, the spacecraft collected dust streaming through the solar system from interstellar sources as well as dust thrown off by Comet Wild-2 during *Stardust*'s flyby on January 2, 2004. On January 15, 2005, the spacecraft returned to Earth, where a sample capsule descended by parachute to the salt flats of the Utah desert.

Genesis. Somewhat similar in concept to Stardust, Genesis was a mission designed to collect particles and return them to Earth. Launched August 8, 2001, Genesis traveled sunward to the Lagrange 1 point, where it spent more than 2 years collecting samples of solar wind particles. On September 8, 2004, the spacecraft brought its sample return capsule to Earth and put it on a trajectory to descend over the salt flats of Utah. The capsule's drogue parachute and parafoil did not deploy, however, and the capsule impacted the ground at a speed of 193 mi/h (311 km/h). Despite this setback, technicians found that much of the payload was intact, allowing the science team to fulfill its highest objectives, including the measurement of the isotopic ratios of oxygen and nitrogen in the solar wind.

Comet Nucleus Tour. The *Comet Nucleus Tour (Contour)* mission was designed to visit two diverse comets. The spacecraft was launched July 3, 2002, but 6 weeks later contact was lost after a maneuver intended to send the probe out of Earth orbit.

Messenger. Messenger—short for Mercury Surface, Space Environment, Geochemistry, and Ranging was conceived as the first probe to orbit Mercury, the closest planet to the Sun. Mercury was previously studied at close range in the 1970s during three flybys by the *Mariner 10* spacecraft. *Messenger* was launched August 3, 2004, and was to fly by Earth once, Venus twice, and Mercury three times before entering orbit at Mercury in 2011.

Deep Impact. This novel mission concept called for a spacecraft to fly to a comet and dispatch a copper

projectile that would excavate a substantial crater in the comet's nucleus, giving scientists a unique view of the nucleus' makeup. Launched January 12, 2005, *Deep Impact* encountered Comet Tempel 1 on July 4, 2005, delivering the impactor was planned. The collision—documented both by the penetrator before its demise and by the flyby spacecraft—threw off a large bright cloud of debris. Investigators were surprised to find that the comet nucleus was buried in a deep layer of powdery dust.

New Millennium Program. At about the same time that NASA established the Discovery Program to sponsor smaller planetary missions, it also created a program called New Millennium designed to launch missions flight-testing advanced technologies. The first New Millennium mission was *Deep Space 1*, a probe launched in 1998 to test an ion engine and 11 other technologies. After an initial problem in starting the propulsion system, the ion engine worked well, and the technology validation effort was viewed as highly successful, although a pointing problem prevented the probe's camera from obtaining useful photos during an asteroid flyby in 1999. *See* ION PROPULSION.

After the failure of the star tracker on *Deep Space 1*, the spacecraft's science camera was reprogrammed to maintain the craft's three-dimensional orientation in flight. After completing its primary mission of flight-testing the new space technologies, *Deep Space 1* executed a flyby of Comet Borrelly in September 2001.

Mars initiatives. Following the failure of *Mars Observer*, NASA quickly approved an orbiter mission, *Mars Global Surveyor*, to duplicate most of the scientific objectives of the lost spacecraft. In addition, the space agency established a program to send pairs of orbiters and landers to Mars during each biannual launch period, culminating in sample return missions in the early twenty-first century.

Mars Global Surveyor. Launched in 1996, *Mars Global Surveyor* spent 2 years circularizing its orbit using the technique of aerobraking, wherein the spacecraft

is slowed by frictional drag as it skims through the planet's thin upper atmosphere at each orbital periapsis. First tested at the end of Magellan's mission at Venus in 1994, aerobraking allows planetary orbiters to use less propellant to carry out orbital insertion at the time of arrival. Mars Global Surveyor embarked on its main mapping mission in 1999, collecting highresolution images and elevation data. Among early findings, scientists determined that the Martian magnetic field is not globally generated in the planet's core but is localized in particular areas of the crust. But the most startling science results came shortly before contact was lost with Mars Global Surveyor in 2006 following nearly a decade in orbit around the Red Planet. By comparing photos taken over a series of several years, scientists found evidence of bright new deposits in two gullies on Mars suggesting that liquid water flowed there briefly sometime during those years.

Mars Pathfinder. This mission was initially conceived as an engineering demonstration of a way to deliver a payload to Mars using the innovative technique of inflatable airbags for landing, and was subsequently equipped with an instrumented robotic rover to carry out a science mission. Launched in 1996, *Mars Pathfinder* (Fig. 18) landed on July 4, 1997, in Ares Vallis, an ancient floodplain that scientists selected because it was believed to contain a wide variety of rocks. The last successful data transmission from *Pathfinder* was completed on September 27, 1997. Both the lander and rover outlived their design lifetimes, the lander by nearly three times and the rover by twelve times. *Pathfinder* returned 2.3×10^9 bits of information, including more than 16,500 images from the lander and 550 images from the rover, as well as more than 15 chemical analyses of rocks and soil and extensive data on winds and other weather factors. Findings from the investigations carried out by scientific instruments on both the lander and the rover suggest that Mars was at one time in its past warm and wet, with water existing in its liquid state and with a thicker atmosphere.

Mars 96. During the same biannual launch period when *Mars Pathfinder* and *Mars Global Surveyor* were launched, Russia launched a complex spacecraft called *Mars 96*, which combined an orbiter and several landers. Originally named *Mars 94*, the mission was renamed due to a launch delay. The *Mars 96* probe failed to leave Earth orbit when booster rockets misfired, and reentered Earth's atmosphere.

Mars 98. Mars 98 consisted of the Mars Climate Orbiter and the Mars Polar Lander. The orbiter carried the balance of science instruments from the payload of the lost Mars Observer that were not reflown on Mars Global Surveyor. The lander, meanwhile,



Fig. 18. Mars Pathfinder, as deployed on Mars, showing primary subsystems. (JPL/NASA)

was designed to be the first spacecraft to settle down near the planet's south pole to study the history of water on Mars. In addition, the lander carried two *Deep Space 2* microprobes, developed under NASA's New Millennium Program, designed to crash into the planet's surface independently and test for the presence of water ice as a technology demonstration.

Launched in December 1998, Mars Climate Orbiter was lost when it entered the Martian atmosphere too low and presumably burned up; a subsequent investigation established that the failure was caused because navigation-related data were not properly converted from U.S. Customary to metric units. Mars Polar Lander was launched in January 1999 and was also lost upon arrival. Although the cause of this failure was less certain, investigators concluded that flight software did not properly take into account the behavior of sensors on the spacecraft's landing legs, and as a result thrusters were probably cut off too early during the craft's final descent. No signal was received from the Deep Space 2 microprobes, which were to be jettisoned just before atmospheric entry.

Mars Odyssey. A follow-up to *Mars Global Surveyor*, the *Mars* Odyssey spacecraft was an orbiter launched April 7, 2001, that reached the planet 6 months later. Aerobraking ended in January 2002, and the science mission commenced a month later. *Mars Odyssey* carries three science instruments, including one package designed to study the radiation environment in space and at Mars for the purpose of protecting future astronauts; it also serves as a communication relay for the Mars Exploration Rovers.

Mars Express/Beagle 2. For the 2003 Mars launch opportunity, the European Space Agency conceived an orbiter, *Mars Express*, equipped with a science payload including an imaging radar that could penetrate the planet's surface. In addition, *Mars Express* carried *Beagle 2*, a United Kingdom-built lander designed to search for signs of Martian life, past or present. Launched June 2, 2003, on a Soyuz-Fregat rocket from Kazakhstan, the orbiter reached the Red Planet on December 25, 2003. *Beagle 2* entered Mars's atmosphere the same day, but contact was never established with it following its descent. *Mars Express*, meanwhile, embarked on its orbital study, producing early results such as evidence of buried impact craters from the imaging radar.

Mars Exploration Rovers. During the same period in which Europe launched *Mars Express* and *Beagle 2*, the United States lofted two identical robotic rovers designed to explore Mars, *Spirit* and *Opportunity*. The travel capabilities of the rovers were much greater than those of *Sojourner*, the United States' previous Mars rover that landed in 1997. *Spirit* was launched June 10, 2003, and landed January 4, 2004, while *Opportunity* was launched July 7, 2003, and landed January 25, 2004. They touched down on opposite sides of the planet—*Spirit* in Gusev Crater and *Opportunity* in Merdiani Planum, both sites being viewed as promising for the study of rocks that could

reveal past water activity on Mars. Despite a software problem that prevented *Spirit* from performing science for 10 days shortly after its landing, both rovers went on to travel around studying rocks for several years, far beyond their originally planned 90-day missions.

Mars Reconnaissance Orbiter. On August 12, 2005, the United States launched *Mars Reconnaissance Orbiter*, which reached the planet on March 10, 2006, and, following aerobraking, began its science mission in November. One of its instruments is a high-resolution camera that is the most powerful ever carried on a planetary mission. In addition, the orbiter is equipped with a subsurface radar designed to operate in conjunction with the radar on *Mars Express.* Soon after arrival, the science team released photos taken by the orbiter of several landers on the planet's surface, demonstrating the power of the high-resolution camera.

Other missions. Although the United States' Discovery and Mars programs account for many of the missions sent beyond Earth orbit in the early twenty-first century, several other projects have been launched under other sponsorships.

Wilkinson Microwave Anisotropy Probe. Intended to survey the sky at infrared wavelengths to measure heat left over from the big bang, the *Wilkinson Microwave Anisotropy Probe* was launched June 30, 2001. It orbits the Lagrange 2 point some 9×10^5 mi (1.5×10^6 km) from Earth. The mission was able to measure many cosmological parameters with higher accuracy than was previously possible; it puts the age of the universe, for example, at 13.7×10^9 years, plus or minus 200×10^6 years. *See* WILKINSON MICROWAVE ANISOTROPY PROBE.

Hayabusa. Known before launch as *Muses-C*, *Hayabusa* is a Japanese mission to collect samples from an asteroid and return them to Earth. The spacecraft was launched May 9, 2003, and approached the small near-Earth asteroid 25143 Itokawa in September 2005. It did not enter orbit around the asteroid, but took up a heliocentric orbit nearby. After making observations of the object, the spacecraft prepared for a series of touchdowns to collect samples. It was uncertain whether the sample collection was carried out as planned, but the spacecraft was believed to have probably obtained at least some dust in its sample collector. The spacecraft is scheduled to deliver a reentry capsule in June 2010 to Earth, where it will land via parachute in Australia.

Spitzer Space Telescope. Similar in concept to a previous Earth-orbiting infrared observatory, the *Spitzer Space Telescope* was launched August 25, 2003, and occupies an Earth-trailing heliocentric orbit, drifting away from Earth at about 0.1 AU per year. The telescope features a 33.5-in. (85-cm) mirror and three instruments that allow it to conduct a variety of photometric, spectroscopic, and spectrophotometric studies. *See* SPITZER SPACE TELESCOPE.

Smart 1. Smart 1 (the name derives from Small Missions for Advanced Research in Technology) was a European spacecraft that orbited the Moon.

The mission's primary objective was testing of a solar-powered ion thruster. Launched September 27, 2003, from French Guiana, *Smart 1* spent 13 months in an orbit gradually spiraling outward from Earth, shaped by the ion propulsion. In November 2004 the ion thruster was activated to decelerate it into orbit around the Moon—a maneuver that was expected to take $4^{1}/_{2}$ days but which required approximately 3 months. It was intentionally crash-landed onto the Moon on September 3, 2006.

Rosetta. The European *Rosetta* probe was launched March 2, 2004, on a mission to rendezvous with Comet 67P/Churyumov-Gerasimenko more than 10 years later, in May 2014. En route, the spacecraft is scheduled to fly by Earth three times and Mars once, in addition to a flyby of the asteroid 2867 Steins in 2008. *Rosetta* carries a lander named *Philae*, which will use harpoons and drills to secure itself to the surface of the comet nucleus.

Venus Express. Another European mission, *Venus Express* was launched November 9, 2005, and entered orbit around Venus on April 11, 2006. The purpose of the mission is to make detailed studies of the planet's atmosphere, clouds, plasma environment, and surface characteristics.

New Horizons. During the 1990s, NASA began conducting studies of how to carry out a mission to Pluto, which was then regarded as the solar system's only planet that had not been visited by a robotic spacecraft. The mission concept that was eventually approved and moved forward was *New Horizons*. Launched January 19, 2006, *New Horizons* was to fly by Jupiter in 2007 and fly by Pluto in July 2015, continuing on and out of the solar system. One or two encounters with Kuiper Belt objects may be possible in the event of an extended mission. A few months after the launch, the International Astronomical Union reclassified Pluto as a dwarf planet. *See* KUIPER BELT; PLUTO.

Solar Terrestrial Observatory. The Solar Terrestrial Observatory (Stereo), is a mission designed to study the Sun. Two identical spacecraft were launched October 26, 2006, into highly elliptical geocentric orbits with an apogee at about the distance of the Moon. After a few orbits, they were to pass close to the Moon, where a gravity-assist effect would send them into heliocentric orbits inside and outside the Earth's orbit. The dual spacecraft allow stereoscopic imaging of the Sun and solar phenomena such as coronal mass ejections.

Trends. Missions of the early twenty-first century demonstrate ongoing refinement in a number of areas. One area that has seen considerable development is that of increasing on-board autonomy. For example, the Mars Exploration Rovers are able to execute simple decision-making about their environments, and how to carry out science observations that makes operations proceed more quickly than if every move was commanded from Earth. In addition, newer spacecraft have great capabilities in their anomaly response. The fault protection system of an older spacecraft might direct it simply to enter a pro-

tective or "safing" mode in response to unexpected conditions. Newer spacecraft are programmed to carry out additional diagnostics and take corrective actions.

Another new initiative is in the area of in situ science instruments. The development of greatly miniaturized microprocessors, memory, and sensors has enabled the creation of more sophisticated science instruments. Many of these are designed to search for evidence of exobiological processes on Mars and other bodies. Franklin 0'Donnell

Bibliography. J. K. Beatty, C. C. Petersen, and A. Chaikin, The New Solar System, 4th ed., Cambridge University Press, 1999; W. E. Burrows, Exploring Space: Voyages in the Solar System and Beyond, 1990; D. Doody and G. Stephan, Basics of Space Flight, JLP D-9774, 1993; W. T. Huntress, Jr., Grand challenges for space exploration, Space Times, American Astronautical Society, May-June 1999; Jet Propulsion Laboratory, Cassini Mission Plan, JPL D-5564, 1994; D. A. Mendis and R. T. Tsurutani, The spacecraft encounters of Comet Halley, Eos, 67(20):478-481, 1986; B. C. Murray, Journey into Space: The First Three Decades of Space Exploration, 1989; F. O'Donnell, Mariner 2 and beyond: Planetary exploration's first 25 years, IEEE Aerosp. Electr. Sys., 2(8):2-6, August 1987; M. D. Rayman, The successful conclusion of the Deep Space 1 mission, Space Technol., 23(2-3):185, 2003; A. Rosenthal, Satellite Handbook: A Record of NASA Space Missions 1958-1980, Goddard Space Flight Center, 1981; Space Studies Board, New Frontiers in the Solar System: An Integrated Exploration Strategy, National Research Council, 2003; E. C. Stone, Mars and the Search for Life Elsewhere: Innovations in the Third Era of Space Exploration, American Institute of Aeronautics and Astronautics, 1999.

Space processing

Experiments conducted in space in order to take advantage of the reduced gravity in studies of the growth, behavior, and properties of materials. Spacelab, developed by the European Space Agency (ESA), was a laboratory module that flew in the space shuttle payload bay. First launched in 1983 on the Columbia, the module became the workhorse for United States and international science missions emphasizing low gravity. Twelve such missions were flown, and an additional seven missions utilized Spacelab components (such as a pallet to support experiments not requiring direct crew interaction). The experiments in fluids, combustion, materials science, and biotechnology conducted on these missions, together with their related ground-based research, resulted in more than 2200 scientific publications, about 1200 of which appeared in peerreviewed journals.

The Spacehab laboratory module, which also flies in the shuttle payload bay, was privately developed by Spacehab, Inc., beginning in 1993. A number of commercial experiments have been conducted on the 21 Spacehab missions (as of January 2007), several of which carried experiments to the Russian *Mir* space station.

Biotechnology. Since biotechnology experiments generally have modest space and power requirements, they were able to be accommodated by middeck lockers in other shuttle missions as well as on the *Spacelab* microgravity emphasis missions, which provided them with many more flight opportunities. Even though the microgravity environment on some of these missions was not ideal, biomolecular crystal growth experiments for structural analysis as well as cell and tissue culturing experiments produced many interesting results, some of which could have commercial applications.

Protein crystal growth. It has now been established beyond question that at least some proteins and other macromolecular systems form higher-quality crystals in a microgravity environment. A survey of 221 macromolecular growth systems flown on 19 shuttle missions found that 57.1% produced crystals large enough for diffraction analysis, and 27.2% (or almost half of those that grew large enough for diffraction analysis) showed improvement in the diffraction data compared to the best crystals ever grown on the ground by any means. One criterion used for selecting growth systems for flight was the ability to grow diffraction-sized crystals in the flight apparatus in a time commensurate to the time available on orbit. Since crystals growing under diffusionlimited transport in space would be expected to grow more slowly than those under convective enhanced transport on the ground, there is little question that many more of the space-grown crystals would have grown to a size large enough for x-ray analysis if more time had been available. This supposition has been confirmed by crystal growth experiments that were carried to the Mir station. Even though conditions on Mir were not ideal for crystal growth, several of these growth systems have produced centimeter-sized crystals that are large enough for neutron diffraction analysis. (Neutron diffraction can directly reveal the positions of hydrogen atoms, whereas the positions of these atoms must be inferred from the positions of heavier atoms seen by x-ray diffraction.) Crystal size limitations have restricted the number of protein structures that have been determined by neutron diffraction analysis to fewer than a dozen. See CRYSTAL GROWTH; CRYSTAL STRUCTURE; NEUTRON DIFFRACTION; X-RAY CRYSTAL-LOGRAPHY; X-RAY DIFFRACTION.

In some cases the improvement in diffraction resolution seen in the space-grown crystals has been key to solving the structure of a particular protein for the first time, and in other cases the space-grown crystals contributed significantly to the refinement of the structure of certain proteins. Several drugs based on structures in which space-grown crystals played a part are in clinical trials.

Cell culturing experiments. A number of experiments have demonstrated that cell metabolism and biosynthesis of cell products are altered in microgravity.

For example, the analysis of the essential oils produced by a rose flower on orbit found products that are not normally found in normal gravity. Several experiments have also shown that gene regulation is significantly altered on exposure to microgravity. Included are genes that regulate cell-cycle, apoptosis, heat shock response, metabolism, structural proteins, cancer cell growth, intracellular signaling, mitochondrial proteins, and a number of transcription factors and DNA binding proteins that are either up-regulated or down-regulated. In particular, it was found that a specific cell-cell signaling matrix material, important to the layering of cells during heart development, was up-regulated in microgravity. This information could be important to the development of tissue-engineered heart patches to replace damaged cardiac tissue. See BIOMEDICAL ENGINEERING; CELL (BIOLOGY); TISSUE CULTURE.

Normal cells do not tend to form threedimensional structures in cultures in vitro. However, a new rotating bioreactor developed by NASA mimics one aspect of microgravity by keeping cells in suspension as they continuously fall through their culture medium. This device has been used on the ground by a variety of investigators at the NASA/NIH Center for Three-Dimensional Tissue Culture and at several universities to study a broad spectrum of biomedical research issues. However, it has also been found that the shear forces on the cells in the rotating reactor often inhibit their attachment to the threedimensional assembly. This was justification for flying the rotating bioreactor on several shuttle flights to the Mir space station. In microgravity, the rotation rates required to maintain the cells in suspension are orders of magnitude smaller than on Earth. Techniques developed to provide nutrients to the culture and remove waste products were successfully demonstrated, and several cultures were carried out which had proven difficult in normal gravity.

Combustion experiments. There are two compelling reasons for the study of combustion in microgravity. One is the issue of fire safety in the design and operation procedures of orbiting laboratories; the other is to take advantage of the weightless state to study certain combustion phenomena in more detail and to test various models in which convection has been ignored in order to be mathematically tractable. Examples of the latter are a number of droplet combustion experiments in which the burning of free-floating or tethered droplets was studied. The absence of gravity allowed the droplet size to be increased to as much as 5 mm (0.2 in.) so that more detailed observation could be made. The objective is to test theories of droplet combustion and soot formation that are of importance to improving the efficiency of internal combustion engines, gas turbine engines, and home and industrial oil-burning heating systems. See COMBUSTION; GAS TURBINE; IN-TERNAL COMBUSTION ENGINE.

In an interesting variation of this experiment, a container was filled with various combustible mixtures near their lean limit of combustion (the lowest concentration of fuel-to-air mixture that will ignite and/or support combustion). A flame ball was created by an electrical spark. A stationary spherical flame front develops as fuel and oxygen diffuse into, and heat and combustion products diffuse out of, the flame ball. This is the simplest possible geometry in which to study the chemical reactions and the heat and mass transport of lean combustion processes. It is expected that these experiments will provide new insight into combustion processes in the lean burning limit, which are important in improving the efficiency of engines and heating systems.

Materials science. Processing of metallic alloys and composites has been carried out in space in order to study their microstructure, thermal properties, and crystal growth.

Evolution of microstructure. The strength and other properties of an alloy depend on its microstructure, which is characterized by the size, orientation, and composition of the grains that make up the solid. One of the main tasks of a materials scientist is to design solidification processes to produce the microstructure that will give the material the desired properties. With current computational capabilities, it is possible to design a complex mold so that the heat and mass flow will produce the desired microstructure throughout the final casting. However, in order to do this the basic laws governing the development of the microstructure must be known, along with the thermophysical properties of the components. Establishing the physical basis for the various laws that describe the solidification process has, over the last half-century, transformed metallurgy from an industrial art based on empiricism to a more exact science. Because of the complicating effects of convection, many of the laws in use today are based on theories that (for simplicity's sake) assume no convective flows. The effects of convection are then added to the calculation to make the result somewhat more realistic. However, most of these theories have never been tested in the absence of convection, so various subtleties may have been overlooked. The ability to experiment in microgravity provides an opportunity to test some of these theories to make sure they are basically correct. Later Spacelab experiments investigated theories that predict the lamella spacing in eutectic structures, the growth rate of dendrites in castings, the coarsening and particle agglomeration in composites formed by liquid-phase sintering, the particle-solidification interactions and ripening in alloys with a dispersed second phase, and the phase-separation mechanisms that act in monotectic alloys (systems which exhibit liquid phase immiscibility). See ALLOY; COMPOSITE MATERIAL; CONVEC-TION (HEAT); MATERIALS SCIENCE AND ENGINEERING; METAL CASTING.

Measurement of thermal properties. A number of selfdiffusion and binary diffusion experiments in liquid metal systems have been carried out on the Spacelab missions. It was found that diffusion coefficients measured in space were generally 30–50% lower than the accepted values based on ground-based experiments carried out in capillaries to minimize convection. This would suggest that most, if not all, diffusion measurements in normal gravity have been contaminated to some degree by convection. Furthermore, the increased precision of the space data allowed an accurate determination of the temperature dependence of diffusion in liquid metal systems. Surprisingly, instead of following the Arrhenius behavior typical of diffusion in solids, it was found that the diffusion data appear to fit a power law in which temperature is raised to exponents ranging from 1 to 2, depending on the system. This unexpected power law behavior of diffusion may shed some light on the structure of the liquid state of metals, which is still poorly understood. *See* DIFFUSION.

Conducting materials can be levitated by electromagnetic induction in normal gravity, but the current required for levitation also heats the sample. In microgravity, the levitation and heating may be decoupled, which allows samples to be deeply undercooled in the liquid state with no physical contact that would induce the nucleation of the solid phase. Thus it becomes possible to measure thermophysical properties of liquids well below their normal freezing temperatures, which are important in developing new metallic glasses and other novel nonequilibrium solids. Several very clever techniques have been developed for measuring the temperature, surface tension, viscosity, and electrical resistivity of levitated samples with no physical contact using the TEMPUS facility developed by the Institute for Space Simulation in Cologne, Germany. These properties of several metallic glass-forming systems were measured in the undercooled state with this apparatus. One particularly interesting result was evidence for magnetic ordering of ferromagnetic Co80Pd20 in the undercooled liquid state near the Curie temperature of the solid, which demonstrates that ferromagnetism can indeed exist in the liquid state. See ELECTROMAG-NETIC INDUCTION; FERROMAGNETISM.

Crystal growth experiments. On four Spacelab missions was flown a sophisticated solidification experiment apparatus developed by the French Space Agency (CNES) and the French Atomic Energy Commission (CEA) called Materials for the Study of Interesting Phenomena of Solidification on Earth and in Orbit (MEPHISTO). It was used by French and American investigators to study the stability and breakdown of the solidification interface of metal alloy systems. The Seebeck voltage generated by the melt-solid interfacial was measured, which served as an indicator of incipient interfacial breakdown. The results were used to test and refine theories of interfacial stability of both faceting and nonfaceting alloy systems. Such theories are important in developing crystal growth processes for alloy-type semiconductor and photonic systems in which interfacial breakdown must be avoided. See ALLOY STRUCTURES; SEEBECK EFFECT; SEMICONDUCTOR; THERMOELECTRICITY.

A number of attempts were made to grow a variety of compound and alloy-type semiconducting systems from the melt with the aim of improving the compositional uniformity through diffusioncontrolled transport. This goal has proven to be elusive, even in microgravity, which underscores the extreme sensitivity of such systems to buoyancydriven flows even at acceleration levels on the order of a micro-g. It was found, however, that dislocations and twinning were significantly reduced in many of the space experiments, especially in regions where the forming solid was not in contact with the wall of the growth ampoule. This observation would implicate wall effects as a major source of imperfections in terrestrially grown crystals, and has encouraged some crystal growers to try to develop growth methods that avoid or least minimize wall contact. See CRYSTAL DEFECTS.

Several semiconductor systems were grown in space by the floating zone process to eliminate wall contact. The absence of hydrostatic pressure in space allows much larger floating zones to be deployed than are possible under normal gravity. Well-ordered crystals of gallium arsenide (GaAs) with diameters up to 20 mm were grown using a mirror furnance to focus light from an arc lamp onto the molten zone. The dislocation densities were very low; however, dopant striations attributed to nonsteady Marangoni convection (convective flows driven by surface tension gradients) were observed. Such striations were avoided in other growth systems that had a native oxide skin which served to suppress the Marangoni flows. See ZONE REFINING.

Robert J. Naumann

Bibliography. P. A. Curreri and D. E. McCauley (eds.), Second United States Microgravity Payload: One Year Report, NASA Tech. Mem., no. 4737, 1996; P. A. Curreri, D. E. McCauley, and C. Walker (compilers), Third United States Microgravity Payload: One Year Report, NASA Tech. Mem., no. 4737, 1996; P. Downey (ed.), Life and Microgravity Spacelab (LMS) Final Report, NASA Conf. Publ., NASA CP-1998-206960, 1998; NASA Microgravity Research Program 1998 Annual Report, NASA Tech. Mem., TM-1999-209757, 1999; N. Ramachandran et al. (eds.), Proceedings of the Joint Launch + One Year Science Review of USML-1 and USMP-1 with the Microgravity Measurement Group, NASA Conf. Publ., no. 3272, 1994; M. B. Robinson (compiler), Microgravity Science Laboratory (MSL-1) Final Report, NASA Conf. Publ., NASA/CP-1998-208868, 1998; P.R. Sahm et al. (eds.), Proceedings of the Symposium on Scientific Results of the German Spacelab Mission D-2, Norderney, March 14-16, 1994 (available from the European Space Agency); R. Snyder (compiler), Second International Microgravity Laboratory (IML-2) Final Report, NASA Ref. Publ., no. 1405, July 1997; Spacelab Science Results Study, Final Report, vol. 2: Microgravity Science, Contract no. NAS8-97095, Task no. H-30194D, August 18, 1999 (available from George C. Marshall Space Flight Center, Huntsville, AL 35812).

Space shuttle

A reusable crewed orbital transportation system. The space shuttle, along with crewless (robotic) expendable launch vehicles such as Atlas and Delta, make up the United States Space Transportation System (STS). The shuttle provides the unique capability for in-flight rendezvous and retrieval of faulty or obsolescent satellites, followed by satellite repair, update, and return to orbit or return to Earth for repair and relaunch. The space shuttle also plays an essential continuing role in the construction and provisioning of the International Space Station (ISS) by transporting major components, such as the giant solar-cell arrays and the Canadian computer-driven double-ended robotic arm, from Earth to the ISS and installing them using extended extravehicular activity (EVA, or space walks) by trained ISS resident and shuttle-visiting astronauts. See SATELLITE (SPACE-CRAFT); SPACE STATION.

Early in its history the space shuttle was expected to be a low-cost replacement for expendable launch vehicles. Following the Challenger shuttle accident in 1986, it became clear that crewless vehicles would have a continuing place in the United States launch vehicle fleet, that they have advantages over the shuttle such as lower cost and shorter lead time for tasks within their capability, and that the shuttle fleet would be fully occupied for much of the first decade of the twenty-first century doing complex jobs requiring human presence, in particular associated with the ISS, which no other launch system can do. For such reasons the STS was expanded to include the expendable launch vehicle families in use by the Department of Defense (DOD) and the National Aeronautics and Space Administration (NASA).

Prior to permanent occupation of the ISS, the shuttle carried attached payloads, such as the European Space Agency (ESA) Spacelab module, which made use of the low-gravity environment for mission durations of up to 10 days or so. Such experiments are now almost always flown on the ISS to take advantage of the longer microgravity duration implicit in the ISS "continuing human presence in space" philosophy.

The space shuttle orbiter accommodates a crew of four to seven for orbital mission durations up to about 10 days. The space shuttle flight system (Fig. 1) consists of the orbiter, which includes the three liquid-fueled shuttle main engines; an external fuel tank; and two solid-fuel strap-on booster rockets. The external tank is discarded during each launch. The solid-fuel rocket casings are recovered and reused. The orbiter lands horizontally as an unpowered ("dead stick") aircraft on a long runway at the NASA Kennedy Space Center in Florida or, when conditions for landing these are unacceptable, at the Edwards Air Force Base in California. See ROCKET **PROPULSION**

Flight sequence. A nominal flight sequence is shown in Fig. 2. The shuttle is launched with all three main engines and both strap-on solid-fuel booster rockets burning. The booster rockets are separated 2 min after liftoff at an altitude of 30 mi (48 km), 28 mi (44 km) downrange. Main engine cutoff occurs about 8 min after liftoff, nearly 70 mi (110 km) above the Atlantic Ocean, 890 mi

(1430 km) downrange from Kennedy. The external tank (ET) is separated from the orbiter shortly after the main engine cutoff, before orbital velocity (speed) is reached, so that the relatively light empty tank can burn up harmlessly during its reentry into the atmosphere above the ocean. Main propulsion for the rest of the mission is provided by the orbital maneuvering system (OMS) engines, whose first task is to complete insertion of the shuttle into its final, nearly circular path during its first orbit (flight once around Earth) after liftoff.

After the shuttle orbiter completes the orbital phase of its mission and is ready to return to Kennedy, its pilot rotates the spacecraft 180 (tail-first relative to the orbiter's direction of motion) and fires the orbital maneuvering system engines to decrease the space vehicle's speed. This maneuver reduces the orbiter's speed just enough to allow the orbiter to fall into the tenuous outer atmosphere, where atmospheric drag causes the orbit to decay (altitude and velocity decrease) along a spiral path. This reentry slowdown occurs in a precise, computercontrolled manner so that the landing may be made manually, halfway around the Earth, on the desired runway at Kennedy or Edwards. At orbital altitudes, small reaction-control engines (jets) maintain the desired orientation of the orbiter. As the spacecraftbecoming-airplane descends into the denser lower air, its aerodynamic control surfaces gradually take over their normal maintenance of heading, pitch, and roll, and the small jets are turned off. Landing is made at a speed of 220 mi/h (100 m/s). The brakes are helped to bring the vehicle to a stop by a parachute (drag chute) which is deployed from the tail after touchdown of the nose wheel and released well before the vehicle comes to a complete stop.

Vehicle data. The orbiter carries into space the crew and the payload for the space shuttle vehicle system. It is comparable in size to a medium commercial jet transport, with a length of 121 ft (37 m) and a wing span of 78 ft (24 m). Its dry (unfueled) mass is 161,400 lb (73,200 kg). The orbiter can deliver cylindrical payloads of up to 65,000 lb (29,500 kg) of mass, with lengths as great as 60 ft (18 m) and diameters of up to 15 ft (4.6 m).

The external tank is 157 ft (48 m) long by 28 ft (7.6 m) in diameter and contains a combined mass of 1,550,000 lb (703,000 kg) of available liquid oxygen (LOX) and liquid hydrogen propellants. The solid-fuel booster rockets are each 149 ft (45.4 m) long by 12 ft (3.7 m) in diameter and have a loaded mass of 1,292,600 lb (586,200 kg). Together, the three space shuttle main engines and the two solid-fuel rocket boosters generate a thrust of 6,400,000 lbf (28,500,000 newtons) at liftoff. The total mass of the shuttle system at liftoff is 4,491,000 lb (2,037,000 kg).

Orbiter vehicle systems. Major orbiter systems are the environmental control and life support, electric power, hydraulic, avionics and flight control, and the space shuttle main engines. Each system is designed with enough redundancy to permit either continua-



Fig. 1. Space shuttle lifting off from the Kennedy Space Center.

tion of mission operations or a safe return to Earth after any single-element failure. For example, three fuel-cell power plants generate the in-flight electric power. Each fuel cell feeds one of three independent electrical distribution buses. Similarly, three independent hydraulic systems, each powered by an independent auxiliary power unit, operate the aero-surfaces. *See* FUEL CELL; SPACE POWER SYSTEMS.

The avionics system uses five general-purpose computers, four of which operate in a redundant set while the fifth operates independently. These computers perform guidance, navigation, and control calculations, and operate the attitude (orientation) controls and other vehicle systems. They also monitor system performance and can automatically reconfigure redundant systems in the event of a failure during flight-critical mission phases. Each of the three space shuttle main engines has an independent computer for engine control during ascent. *See* DIG-ITAL COMPUTER; MULTIPROCESSING.

The thermal protection system, which is not redundant, presented a major technical challenge



Fig. 2. Typical space shuttle mission profile.

to the orbiter development schedule. This system, unlike those of previous single-use spacecraft, has a design requirement of reuse for 100 missions. Performance requirements also dictate that the system withstand temperatures as high as 3000° F (1650° C) while maintaining the vehicle's structure at no more than 350° F (177° C). *See* ATMOSPHERIC ENTRY.

The key to meeting this challenge was to develop a material possessing an extremely low specific heat capacity and thermal conductivity, together with adequate mechanical strength to withstand the launch and reentry vibration and acceleration. These thermal characteristics must remain stable up to temperatures of at least 3000°F (1650°C). The solution was found in silica ceramic tiles, some 24,000 of which cover most of the orbiter's surface. During the highest thermal load portion of the reentry trajectory (path), the orbiter's wings are level and the nose is elevated well above the flight path. This causes the temperature of the undersurface to be substantially higher than that of the upper surface. For this reason, the undersurface is covered with black borosilicate glass-coated high-temperature tiles. Most of the upper surface of the shuttle is covered with a lowertemperature silica blanket made of two outer layers of woven fabric and an insulating buntinglike center layer stitched together in a quiltlike pattern. Finally, the nose cap and wing leading edges are subject to the highest (stagnation) temperatures, which requires the use of molded reinforced carbon-carbon material.

Flight testing. Like any commercial or military transport aircraft, the space shuttle system had to pass a series of flight tests before it could be committed to operational service. These tests were conducted in two phases: approach and landing, and orbital flight.

For the approach and landing phase, which tested the capabilities of the shuttle as an unpowered aircraft, the first test vehicle, *Enterprise*, was airlaunched from a modified Boeing 747 airplane which was used later to airlift ("piggyback") operational shuttle orbiters between California and Florida. Five successful flights in 1977 verified the subsonic (lowspeed) aerodynamics and handling qualities of the orbiter.

In 1981 the first protoflight powered space shuttle *Columbia* was launched from Kennedy on the first of four orbital test missions, STS 1 through STS 4. All major test objectives were successfully accomplished. The most significant problem that surfaced involved damage to *Columbia*'s thermal tiles, but the orbiter's underlying structure suffered no

temperature problems and the repaired areas survived subsequent flights without further thermal difficulties.

Flight operations. STS 5 initiated the operational shuttle flight phase in 1982. Many tests of the orbiter remained to be accomplished and were scheduled in conjunction with operational flights. The ability to do deployment, rendezvous, proximity operation, and retrieval using the remote manipulator system (robotic arm), was demonstrated in 1983 using the West German Shuttle Pallet Satellite Spacecraft (SPAS). One year later the first shuttle landing took place on Kennedy's 3-mi (5-km) runway. Longer-term problems were found and fixed. For instance, an unexpected chemical reaction with the tile waterproofing agent caused degradation of the tile bonding adhesive, requiring replacement of the degraded material and development of a new waterproofing agent. In 1985 the planned four-shuttle fleet was completed with the delivery of the orbiter Atlantis

Payload accommodations. The orbiter payload bay is divided into four cylindrical sections, each of which can support one major payload. Standard services to each section include electric power, data recording, and two-way telemetry (data exchange). Optional services, such as active payload cooling, may also be purchased by the customer.

Small experiments for researchers are accommodated by a variety of carriers. These are normally manifested on a space-available basis and flown at a relatively modest cost. Self-contained experiments requiring only turn-on and turn-off by the crew are carried in canisters mounted along the forward portion of the payload bay sidewall. They are known informally as "get-away special" (GAS) canister (CAN) payloads. Small experiments can also be acommodated in the middeck section of the pressurized cabin. Larger experiments are mounted on a carrier in the payload bay, making orbiter power and telemetry services available.

Experiments requiring precision pointing can be mounted on a free-flying carrier that is deployed and retrieved with the robotic arm. The carrier platform provides highly accurate pointing and a microgravity environment.

Figure 3 shows a typical collection of payloads and experiments to illustrate the diversity of purposes which have been served on each shuttle flight. Specifically, it shows the STS 61-B cargo configuration, which was made up of three communications satellites: the Australian AUSSAT 2; the Mexican MORELOS-B; and the United States commercial SATCOM KU-2. The attached payloads are Experiment Assembly of Structure in Extravehicular Activity (EASE), Assembly Concept for Construction of Erectable Space Structure (ACCESS), IMAX Cargo Bay Camera (ICBC), and G-479 Telesat-Canada Getaway Special (GAS). In the crew compartment are the Continuous Flow Electrophoresis System (CFES), Diffusive Mixing of Organic Solutions (DMOS), Morelos Payload Specialist Experiments (MPSE), and Orbiter Experiments (OEX).



Fig. 3. STS 61-B cargo configuration. Acronyms are explained in the text.

It should be noted that, were this mission (which was launched in November 1985) planned after 1986, probably each of the communications satellites would be handled by an expendable launcher, and after 2000 many of the experiments would be done on the *ISS*.

Challenger disaster. Shortly before noon on January 28, 1986, *Challenger* lifted off from Kennedy Space Center on the twenty-fifth shuttle mission. At 72 s into the flight, with no apparent warning from real-time data, the external tank exploded and *Challenger* was destroyed.

All seven crew members perished, the first to die in NASA's human space-flight program spanning 25 years and 56 crewed launches. Further flight operations were suspended while a Presidential (Blueribbon) Commission reviewed the circumstances surrounding the accident, determined its probable cause, and developed recommendations for corrective actions.

Recovery. Nearly 5 months later the Commission submitted its report which contained recommendations to be implemented before returning the space shuttle to flight status.

The cause of the accident was determined to be inadequate design of the solid rocket motor field joint. "Field joint" refers to the assembly of the solid rocket motor at Kennedy rather than at the manufacturer's factory, because the massive motor must be shipped in cylindrical sections. Deflection of the joint with deformation of the seals at cold temperature allowed hot combustion products to bypass both O-rings, resulting in erosion and subsequent failure of both primary and secondary seals.



Fig. 4. Solid-fueled rocket motor field joint. (a) Original design. (b) New design.

The redesigned field joint (**Fig.** 4) has an overlapping tang that helps prevent separation during the startup phase of motor operation when the case is responding dynamically to the internal pressure buildup. A third O-ring seal, in the overlapping tang, was also included in the new design, along with the ability to verify the seal's integrity after installation. Other improvements were also made in the design of the solid rocket motor.

NASA completely reviewed all activities associated with launch aborts and established an additional transatlantic abort landing site at Ben Guerir, Morocco. An escape system was devised that allows the crew to parachute out of the orbiter once the vehicle has achieved controlled gliding flight. The system uses a pole extended out of the orbiter side hatch (**Fig. 5**).

NASA also modified its shuttle program management structure and established an Office of Safety, Reliability, Maintainability, and Quality Assurance, reporting directly to the NASA Administrator.

Return-to-flight actions. Before resuming space shuttle flights, in addition to actions mandated by the Presidential Commission, NASA initiated a Systems Design Review (SDR), a detailed review of all shuttle program hardware and software. Its objective was to establish a series of changes that would result in improved safety, with reliability and efficiency being of second and third priority, respectively. As a result, 254 modifications were made to the orbiter, 24 to each main engine, 30 to each solid rocket booster, and 13 to the external tank.

Because the solid rocket cases are recovered and refurbished, it was known that the primary (but not the secondary) O-ring had failed on several earlier flights. For human space flight systems, a basic requirement of redundancy for safety, which includes the field joint O-rings, is that should the designed redundancy be known or suspected of being compromised, the mission must be delayed until the redundancy is known to be restored. This requirement was waived by NASA for several shuttle missions, including *Challenger*'s final flight.

On September 29, 1988, 32 months after *Challenger*'s final flight, space shuttle *Discovery* successfully resumed shuttle operations. During the next 14 years, more than 80 missions were flown with no anomalies more serious than several foreshortened missions because of the failure of a redundant component or subsystem.

Columbia disaster. On February 3, 2003, while the shuttle *Columbia* was reentering the atmosphere, it disintegrated. All seven crew members perished, and pieces of the orbiter fell in a long swath over East Texas. As much of the debris as possible was collected and taken to Kennedy Space Center, where pieces were placed in their appropriate position on a hanger floor. The Columbia Accident Investigation



Fig. 5. Orbiter crew escape system. (a) Typical profile. 1 ft = 0.3 m. (b) U.S. Navy jumper demonstrating use of the crew escape system pole in a test aircraft.

Board (CAIB) was formed to determine the cause of the accident and to present recommendations. As during the *Challenger* investigation, all flights were suspended.

Recovery. Seven months after the disaster, the CAIB issued its report. The cause of this accident was a breach in the thermal protection sytem on the leading edge of the left wing, caused by a piece of insulating foam that separated during launch from the left bipod ramp section on the external tank. This breach in the reinforced carbon-carbon panel allowed superheated air to penetrate through the leading edge of the left wing, which melted the aluminum structure of the wing. Aerodynamic forces caused loss of control of the shuttle and its subsequent breakup.

The CAIB made 15 recommendations. The independent NASA Return to Flight Task Group agreed that NASA had met the intent of 12 of the recommendations. NASA could not meet the major recommendation, to eliminate foam from leaving the external tank, due to engineering problems. NASA did remove the foam on the bipod ramp and made improvements in foam application to lessen the foam debris. NASA also met other recommendations for better ground-based imagery of the vehicle, and better high-resolution imagery of the orbiter and the external tank.

STS 114 return to flight. When *STS 114* launched in July 2005, the new cameras on the external tank, installed for high resolution of the orbiter, recorded a large piece of foam coming off the tank. On orbit, the crew used the new Orbiter Boom Sensor System (OBSS) to inspect the reinforced carbon-carbon and the tiles for damage. Only two gap fillers were found, extending between the tiles. NASA decided that the fillers had to be removed or cut off. An astronaut on the station arm pulled them out easily. On another spacewalk, NASA also tested several methods to repair the tiles and the reinforced carbon-

carbon, should further damage be detected in the future.

Six in-flight anomalies were taken due to the unexpected foam loss. The analysis and wind tunnel testing required to understand and eliminate the loss led to further delays in the launch of the next mission. Worse, when hurricane Katrina struck the Gulf Coast, it damaged the Michoud Assembly Facility, which builds the external tanks, causing further delays. The next mission, *STS 121*, did not launch until July 4, 2006, almost exactly one year after *STS 114*. The most significant change in the tank was the elimination of the liquid hydrogen and liquid oxygen protuberance air load (PAL) ramps. The tank was shipped from Kennedy Space Center back to Michoud for the modifications.

Technical achievements. The first operational shuttle accomplished the first commercial satellite deployments in 1982. In 1984 the first in-orbit satellite repair was accomplished on the *Solar Maximum Mission*'s control system and a primary experiment sensor. Later that same year *Palapa* and *Westar*, two communications satellites in useless orbits, were recovered and returned to Earth on the same space shuttle mission. These repair and recovery missions demonstrated the usefulness of humans in these new space tasks.

In September 1988, a *Tracking and Data Relay Satellite* (*TDRS*) was launched by the first space shuttle to fly after the *Challenger* accident. The initial *TDRS* network was completed 6 months later by the second shuttle launch of a *TDRS*, which enabled the closing of many Spaceflight Tracking and Data Network (STDN) ground stations and a major improvement in the duration and bandwidth of communication with robotic and crewed spacecraft. *See* SPACE COMMUNICATIONS; SPACECRAFT GROUND IN-STRUMENTATION.

The year 1989 was a banner one for solar system exploration, thanks to space shuttle launches of

Magellan to Venus and Galileo to Jupiter. In 1990 the first mission, Ulysses, was undertaken to investigate the magnetic field configuration and variations around the Sun's poles, after the craft's shuttle launch. STS 32 launched the long-awaited Hubble Space Telescope (HST) in 1990; however, this telescope's most important work was not accomplished until later that decade after two shuttle EVA visitsthe first in 1993 to correct a bad case of astigmatism produced by the HST mirror manufacturer, and the second in 1997 to update the sensors and replace major spacecraft subsystem components that were failing. Additional servicing missions were carried out in 1999 and 2002. In 2006, after much debate, NASA decided to perform one more Hubble servicing mission, enhancing its capabilities and extending its life by replacing failed components and updating some sensors and data processors.

Science accomplishments. Successful launches of Magellan, Galileo, Ulysses, and HST (together with the two in-orbit servicing calls) constitute a major contribution to the space sciences. The deployment of the Long Duration Exposure Facility (LDEF) as a free-flying satellite by a shuttle in 1984 and its shuttle recovery in 1990 alerted materials researchers to the rapid degradation of many materials used in space, even though some data were lost due to the delay of the LDEF recovery because of the hiatus in shuttle operations in 1986-1988. Finally, the experience of the bioscientists and materials researchers developing research methods and equipment for experiments aboard ISS is helping to get this work off to a good start. See REMOTE MANIPULATORS; SPACE BIOL-OGY; SPACE PROCESSING.

Safety upgrades. Until 2001, official NASA statements projected the availability of a single-stage-toorbit space shuttle replacement by 2012. Cancellation or suspension in 2001 of the X-33 and other steps toward a shuttle replacement led to speculation that the space shuttle might be the major crewed link with *ISS* until 2020 or later. Consequently, the Space Flight Advisory Committee advised NASA to give higher priority and funding to space shuttle high-priority safety upgrades, which NASA had already identified and was investigating.

In 2001, NASA planned to develop four highpriority safety upgrades for the shuttle fleet. An electric auxiliary power unit would replace the hazardous hydrazine-fueled high-speed turbine-driven hydraulic systems. New engine sensors in the Main Engine Advanced Health Monitoring system could cut the risk of a catastrophic engine failure by about 25%. The Solid Rocket Booster Advanced Thrust Vector Control project would shift away from hydrazine in the turbines that drive the nozzle hydraulics. Cockpit upgrades would install advanced avionics in the shuttle fleet, including the flat-panel cockpit displays already available on *Atlantis* to ease the crew workload during ascent.

In 2004, President Bush announced the Vision for Space Exploration. Part of the Vision required the shuttle to be retired no later than 2010, only 5 years after the first return-to-flight mission. Consequently, some of the safety upgrades were canceled, and other upgrades were added. The electric auxiliary power units and Solid Rocket Booster Advanced Thrust Vector Control programs were canceled. The Advanced Health Monitoring System was flown in monitor-only mode on *STS 116* in December 2006. The results will be examined before it is placed in operational mode. All three orbiters received the Cockpit Avionics Upgrade.

A new upgrade was added, the Station/Shuttle Power Transfer System (SSPTS). Since the *ISS* is powered from an inexhaustible supply of sunlight and the shuttle is limited in power by the amount of hydrogen and oxygen carried onboard, a system was designed to power the shuttle from the *ISS*. The SSPTS converts the 125-volt DC *ISS* power to 28-volt DC shuttle power. This will allow the shuttle to stay at the *ISS* for longer periods of time. The shuttle *Edeavour* has been modified with the SSPTS and is scheduled to fly in June 2007 to test it. Installation of the SSPTS is also planned on the shuttle *Discovery*, but not on *Atlantis*, due to the retirement of the shuttles. John F. Clark; Jeffrey C. Mitchell

Bibliography. Aviation Week and Space Technology, March 19, 2001; H. Benedict, NASA: The Journey Continues, 2d ed., 1992; Columbia Accident Investigation Board Report, 6 vols., Government Printing Office, 2003; D. R. Jenkins, Return-to-Flight Space Shuttle Discovery, 2006; D. R. Jenkins, Space Shuttle, 3d ed., 2001; NASA, Shuttle Flight Data and Inflight Anomaly List, Rep. JSC-19413, Revision M, January 1990; J. R. Wertz and W. J. Larson (eds.), Space Mission Analysis and Design, 3d ed., 1999.

Space station

A complex physical structure specifically designed to serve as a multipurpose platform in low Earth orbit. Functioning independently and often without a crew actively involved in onboard operations, a space station contains the structures and mechanisms to operate and maintain such support systems as command and data processing, communications and tracking, motion control, electrical power, and thermal and environmental control. Evolving together with technology and increasing in scope and complexity, the space station has a history in which each program was based upon the developments and achievements of its predecessor.

Salyut. The Soviet Union began construction of the world's first space station, called *Salyut*, in 1970. It was the primary Soviet space endeavor for the next 15 years. The design was retained not only through a series of Salyuts that were launched within that decade and later, but also through the development of *Mir*, the most famous, long-lived Russian achievement in space. The Salyuts were cylindrical in shape and contained compartments with specialized functions for operating a space station. The docking compartment was designed to accept the Soyuz spacecraft that transported the cosmonauts. The transfer

compartment gave the cosmonauts access to the various work compartments. The work compartments contained the mechanisms that operated and controlled the station, as well as the laboratories in which the cosmonauts performed experiments while they were onboard. *Salyut 1* initially carried a crew of three, but after three cosmonauts died when a valve in the crew compartment of their descent module burst and the air leaked out, the crews were reduced to two cosmonauts, and both crew members were outfitted with pressurized space suits.

During the Cold War, Soviet Salyuts were launched as missions to serve both civilian and military purposes. These missions never were linked in a flight, and many of the details about the flights were shrouded in secrecy. Since the end of the Cold War those details have been declassified, revealing the extent and scope of Soviet activity in space, which included astronomical and medical experiments; testing of a broad range of communications and telemetry, guidance, and navigation systems for orbiting spacecraft; Earth observation using cameras and satellites; manufacture of space-processed products; and, perhaps most importantly, evaluations such as the effects of zero gravity over an extended period of time on a human body.

The Soviet Salyuts set the record of 8 months for extended duration of human space flight. At the end of their usefulness, all Salyut spacecraft were deorbited by Mission Control Moscow and permitted to burn up during reentry into the Earth's atmosphere.

Skylab. Starting later than the Soviets, the United States also had plans for a space station. The program, known as Skylab, was developed by the National Aeronautics and Space Administration (NASA) building on the success of its heavy-lift rocket, the Saturn, which had boosted the Apollo rockets and helped place the first human being on the Moon. Skylab weighed just less than 100 tons (90 metric tons). It was launched on May 14, 1973, from the Kennedy Space Center aboard a Saturn 5 rocket, which enjoyed a reputation for never having failed. Although the launch was flawless, a shield designed to shade Skylab's workshop deployed about a minute after liftoff and was torn away by atmospheric drag. That began a series of problems, most involving overheating, that had to be overcome before Skylab was safe for human habitation. Eventually three crews served aboard Skylab throughout 1973 for periods of 28, 59, and 84 days, respectively. The single greatest contribution made by each crew was the extravehicular activities that restored Skylab's ability to serve as a space station.

Apollo-Soyuz Test Program. In 1975, during the period of dètente, plans were made for a joint United States-Soviet space venture, known as the Apollo-Soyuz Test Program (ASTP). For the first time, United States astronauts and Soviet cosmonauts became members of each other's flight teams, training together, touring their launch sites, and working in their mission control rooms. For 15 months, they trained and prepared for their historical joint mission

in space. In July 1975, the astronauts of *Apollo 18* linked with the cosmonauts of *Soyuz 22*; their docked configuration functioned as a miniature international space station.

Mir. Throughout the 1970s, the goal of the Soviet space program appeared to be the development of a permanently deployed space station. It would have both civilian and military applications, and it would be based upon the technology gleaned from the experience with the Salyut and the Soyuz vehicles. A major task that the Soviets sought to accomplish in their quest for a permanent space station was the optimization of conditions that would enable cosmonaut crew members to live in space for extended periods of time. The crew then would be able to operate at maximum efficiency and effectiveness when performing space-based experiments. Soviet decision makers sought to take advantage of the unique opportunities presented by a permanent space station to exploit the space environment and benefit their economy. For example, Soviet scientists studied methods to direct solar power by means of equipment onboard a space station to Soviet facilities for consumption at home.

Mir (which in Russian means either "peace" or "world") was the name of the vehicle for the world's first multipurpose, permanently operating crewed space station. Based upon the Salyut design and configuration, Mir essentially was an extension and expansion of the shell of the basic Soviet space vehicle (Fig. 1). It would incorporate the standard Salyut/ Soyuz/Progress profile. External ports were added for docking of Soyuz vehicles which would carry crew members, and Progress (resupply) vehicles which would bring foodstuffs, drinking water, extra equipment and apparatus, sanitary requisites, medical apparatus, and propellant. These same Progress vehicles would return space "junk" to Earth. Construction was based upon a modular design, permitting the Soviets to replace modules whenever significant improvements in technology made the earlier modules obsolete. Two cylindrical modules formed the basic shape of Mir and served as the living and central control compartments for the crews. Additional modules were used for scientific experiments

The Mir core station was a 49-ft (15-m) module. Its launch aboard a Soviet Proton rocket on February 20, 1986, was televised internationally for the first time in Soviet space history. Mir was assembled in space and was composed of six modules. In addition to the core, the Kvant-1 module (launched in April 1987) was a 19-ft (6-m) pressurized lab used for astrophysics research. The Kvant-2 module was launched in December 1989, and was used for biological research and for Earth observation. The Kristall module was launched in August 1990, and provided the docking port for the United States space shuttles that visited Mir. Between 1994 and 2001, there were nine dockings between space shuttle vehicles and Mir. The final two modules, Spektr and Priroda (launched in June 1995 and April 1996, respectively) were remote sensing modules used to study the Earth. Mir



Fig. 1. Mir complex, in orbit 1986–2001. (After Commercial Uses of Space Exploration [brochure], Glavcosmos, 1989)

provided a platform for long-term microgravity research and increased knowledge about living and working in space. By the end of its function in March 2001, the 143-ton (130-metric-ton) *Mir* had spent 15 years in orbit and had orbited the Earth more than 87,600 times with an average orbiting speed of 17,885 mi/h (28,800 km/h). Its superstructure, 109 ft (33 m) long and 90 ft (27 m) wide, burned up upon reentry in to the Earth's atmosphere, and scraps were scattered into the Pacific Ocean just east of New Zealand. *See* AEROSPACE MEDICINE; SPACE BI-OLOGY; SPACE PROCESSING; SPACE SHUTTLE.

International Space Station. Through the late 1980s, the United States continued to compete with the Soviet Union for dominance in space. The Soviet successes with Mir prompted the United States to respond with what became known as Space Station Freedom. Moved from the NASA drawing board in 1993, components of the space station were tested by shuttle crews during a variety of missions. Supporters of this space station recognized that the benefits of space exploration are far-reaching and longterm, and they advocated such unique opportunities as manufacturing drugs, scientific materials research in microgravity, and studying the health and status of the Earth's environment from outer space. At the direction of President Clinton, the United States transformed the single-nation concept for Space Station Freedom into a multinational partnership with the European Space Agency and the Russian Space Agency to create what is now known as the International Space Station (ISS). See SPACE PROCESSING.

Five space agencies [NASA in the United States, the Canadian Space Agency (CSA), the European Space Agency (ESA), the Japan Aerospace Exploration Agency (JAXA), and the Russian Federal Space Agency (RSA)] and sixteen nations have united to build the *International Space Station*. More complex than any ground-based civil engineering project built to date, the *ISS* will take approximately 12 years to construct (**Fig. 2**).

The design evolved over more than a decade. When completed, the ISS will be the largest humanmade object ever to orbit the Earth. It will include six laboratories built by a consortium of nations. The internal research space will be equivalent to the cabin space inside a 747 aircraft. Approximately 80 United States space shuttle, Russian Soyuz, and partner-nation heavy transport vehicle flights will transport the structures and mechanism necessary to construct the station over those 12 years. The United States space shuttle must deliver most ISS modules and major components. The ISS will weigh approximately 1 million pounds (almost 450,000 kilograms); it will orbit the Earth at an inclination of 51.6° to either side of the Equator; it will fly at an altitude between 230 and 285 mi (370-460 km); and it will move at a speed of 17,500 mi/h (28,000 km/h), orbiting the Earth every 90 min. It will host a maximum crew of six.

As of January 2007, there had been 18 United States space shuttle, 14 Russian Soyuz, and



Fig. 2. Artist's rendering of how the *International Space Station* may look when its assembly sequence is completed in 2010. (Changes to the final station configuration will likely occur before then.) The 1-million-pound (450,000-kg) station will have a pressurized volume equal to two jumbo jets, and an acre of solar panels. (*NASA*)

25 Russian Progress flights to build and support the International Space Station. The first flight, from the Baikonur Cosmodrome (a launch facility in Kazakhstan), was a Russian Proton rocket in November 1998, which carried the Zarya module or Functional Cargo Block (FGB). This was a selfcontained, independently operating power and communications module. It formed the base to which further components were added. A month later, the space shuttle Endeavour carried up the Unity module, which was attached to Zarya. Additional flights added components and conducted repairs. The Russian service module (SM), Zvezda, was launched and docked with the FBG in July 2000. It provided the station's early living quarters and systems to support power, data processing, flight control, propulsion, and life support. Consistent with the Russian historical design and architecture of the Salyut and Mir platforms, Zarya and Zvezda featured automated operation.

Expedition flights support the ISS. Astronauts and cosmonauts assigned to flights to the ISS compose an Expedition crew. As of January 2007, there had been 14 Expedition crews posted to the ISS. Expedition crews are composed of either two or three members: a Commander and either one or two Flight Engineers. The primary missions of each Expedition crew are to assemble the station and to learn how to live and work in space. These are essential tasks if crewed space programs are to venture deeper into the solar system for the purposes of human exploration and scientific discovery. To this end, Expedition crew members serve as research scientists. They use the *ISS* as an orbiting laboratory and as a test platform for future technologies.

Expedition One begins building the ISS. The first permanent crew of the International Space Station also launched from the Baikonur Cosmodrome, on October 31, 2000, aboard a three-stage, 310-ton (280-metric-ton) Soyuz rocket. Known as Expedition One, the crew was composed of one American and two Russians. U.S. Navy Captain William (Bill) Shepherd served as the commander in what was his fourth trip into space. The Russians were Air Force Lieutenant Colonel Yuri Gidzenko, who served as Soyuz commander, and Sergei Krikalev, who served as flight engineer. This was the second flight for Gidzenko, whose first flight was as the commander for Mir on which he served for 179 days. Krikalev had been in space four times previously, logging 484 days in space. One of his flights was to the International Space Station aboard space shuttle Endeavour for the mission which mated the Zarya and Unity modules. Shortly after their arrival, the Expedition One crew named the International Space Station the "Space Station Alpha."

During flight the Expedition One crew used Greenwich Mean Time (GMT) as its reference. This

was designed to provide the best combination of communications coverage for the American and Russian flight control teams.

While aboard the International Space Station, the Expedition One crew hosted three visiting space shuttle crews. Those crews arrived with the mission of continuing the construction of the station. The vast array of photovoltaic cells that provide the bulk of the power for the station arrived and was installed by one crew. An additional crew brought the American laboratory, Destiny, which will be the center for scientific research. Within the laboratory are housed a variety of scientific racks, containing hardware and equipment for research and study. The primary mission of the Expedition One crew was to prepare the station for long-term habitation, and there were several significant tasks that had to be accomplished. For crew command and control, all onboard computer control systems and all means to identify emergencies, warnings, and cautions were inspected and tested for function and reliability. For crew comfort and safety, all air purification and revitalization devices, heaters, and water supply systems were inspected and made serviceable. For communications within the modules and down to Mission Controls in Houston and Moscow, all internal and external communications, command and data handling, and telemetry systems were tested and made operational. In addition, the Expedition One crew was to refurbish the International Space Station. That was accomplished by unloading three crewless Russian Progress resupply vehicles, all of which arrived and docked automatically to the ports available on the Russian Zarya and Zvezda modules.

Because of the amount of time required by the Expedition One crew to prepare the station for nominal operations, there was not much time planned for scientific activities and experiments. However, some of the experiments planned for the first expedition included research into protein crystal growth, analysis of seed growth, study of the Earth by means of high-resolution images taken from the space station, spatial differences in carbon dioxide concentration, physiological tests involving a resistive exercise device, and treadmill and medical evaluations of locomotion and heart rate. Scientific activities prepared independently by the Russian Space Agency involved research in the fields of biotechnology, biomedical research, geophysical research, and engineering.

Expedition Two continues construction. Expedition One was succeeded by Expedition Two, which launched aboard space shuttle *Discovery* on March 8, 2001, and docked with the *International Space Station* the following day. The Expedition One crew returned to Earth aboard *Discovery*, which undocked from the *International Space Station* on March 18, 2001, and landed at Edwards Air Force Base 2 days later. Their highly successful mission had lasted 138 days 17 hours 39 minutes.

Expedition Two also had a three-member crew. The commander was Russian cosmonaut Yuri Usachev, and the two flight engineers were U.S. Air Force Lieutenant Colonel Susan Helms and U.S. Army Colonel James Voss. On their arrival flight, the space shuttle *Discovery* carried the first multipurpose logistics module, which was built by the Italian Space Agency and named Leonardo. The logistics modules are reusable cargo carriers designed specifically for space-flight operations. The space shuttle *Endeavour* with a crew of seven launched from Kennedy Space Center on April 19, 2001, to deliver additional station-building materials and instrumentation. Unique among the cargo in this flight was the space station's remote manipulator system or robotic arm, and another multipurpose logistics module, called Rafaello.

Follow-on Expeditions. Through Expedition Fourteen and continuing with Expeditions until the *ISS* is no longer viable as a platform for exploration and research, the Expedition crews will serve onboard the *ISS* for stays between 4 and 6 months. In addition to the tasks of helping build, maintain, and sustain the *ISS*, the Expedition crews conduct science activities daily in a variety of fields such as human research, physical and life sciences, technology demonstrations, Earth observations, and education activities. These efforts were interrupted briefly after the loss of the space shuttle *Columbia* in February 2003, and United States space shuttle flights to the *International Space Station* were suspended until July 2005.

Given the ISS status as an advanced orbiting laboratory operating in zero gravity, the opportunities to study human endurance in space and to test new technologies and techniques for space exploration are endless. Human medical research, which develops the knowledge essential to send space explorers beyond Earth's orbit and deep into the solar system, focuses both on the effects of living in space and zero gravity and the countermeasures necessary to reduce the health risks associated with space travel. The goal is to defeat the harmful effects of space on human health. Research in the physical and life sciences is intended to use the environment of microgravity to study the growth of proteins, cells, crystals, and other biological structures. Technology demonstrations help develop new technologies used in spacecraft materials and systems that function in the microgravity environment. Finally, Earth observations and education activities link the global population with the work and mission on board the station, inspiring humankind to continue to explore and use space as a resource for improving life on Earth.

Future development. In addition to serving as a test bed and springboard for future space exploration, the *ISS* will serve as a research laboratory for innovation in science and technology. Early research on the *ISS* has focused on such topics as biomedical research and countermeasures to understand and control the effects of space and zero gravity on crew members; biological study of gravity's influence on the evolution, development, and internal processes of plants and animals; biotechnology to develop superior protein crystals for drug development; and fluid physics. Advanced research will be oriented toward topics such as human support technology, materials science, combustion science, and fundamental physics.

Construction of the *International Space Station* is projected to be completed in 2010. The ISS Program will continue to coordinate the training and operations of international flight crews, use vehicles built in multiple countries and launched at multiple sites, integrate a worldwide communications network that extends into deep space, and unite the international scientific and engineering communities.

When the *International Space Station* is completed, it will host an international crew of six members. Training for these crews is projected for a period between 3 and 5 years. Expedition crews will live and work aboard the station for a period of 3 to 6 months. The flight records of these expeditions and the results of the research conducted by the crews will be available to the international community. *See* SPACE FLIGHT; SPACE TECHNOLOGY; SPACE-CRAFT STRUCTURE. James J. Pelosi

Bibliography. V. Gorshenin, History of Soviet Space Exploration in Manned Orbiting Facilities, Soviet Academy of Sciences Press, Moscow, 1994; V. Grenowitz, The triumph of Mir: A Russian national treasure, Space Tech. J. (Moscow), no. 33, April 2001; M. Lechliter, Whence the Journey into Whirlwind: Cosmonaut Challenges in Space, Moscow State University Press, 2000; Library of Congress, Federal Research Division, History of Skylab, Item no. 1167, November 1975; N. R. Pellis and R. M. North, Recent NASA research accomplishments, aboard ISS, 54th International Astronautical Congress of the International Astronautical Federation, the IAA, and the International Institute if Space Law, Bremen, Germany. IAC-03-T.4.07, Sept. 29-Oct. 3, 2004; J. L. Rhatigan, J. A. Robinson, and C. F. Sawin, Exploration-Related Research on ISS: Connecting Science Results to Future Missions, NASA/TP-2005-213166, 1-34, 2005; J. A. Robinson and D. A. Thomas, NASA utilization of the International Space Station and the Vision for Space Exploration, ISTS 2006-0-1-06, International Symposium on Space Technology and Science, Kanazawa, Japan, June 4-10, pp. 1-6, 2006.

Space technology

The systematic application of engineering and scientific disciplines to the exploration and utilization of outer space. Since 1957, space satellites and probes have been launched beyond the Earth's atmosphere to orbit the Earth (velocity of 7 km/s or 4.4 mi/s) or to escape from it (velocity 11 km/s or 6.9 mi/s). Due to the severe environmental conditions in space, such as high vacuum, intense solar radiation, and extremes of temperature, a variety of disciplines are required to design, build, test, and launch a successful spacecraft. *See* ESCAPE VELOCITY.

Space technology developed so that spacecraft

and humans could function in this environment that is so different from the Earth's surface. Conditions that humans take for granted do not exist in outer space. Objects do not fall. There is no atmosphere to breathe, to keep people warm in the shade, to transport heat by convection, or to enable the burning of fuels. Stars do not twinkle. Liquids evaporate very quickly and are deposited on nearby surfaces. The solar wind sends electrons to charge the spacecraft, with lightninglike discharges that may damage the craft. Cosmic rays and solar protons damage electronic circuits and human flesh. The vast distances require reliable structures, electronics, mechanisms, and software to enable the craft to perform when it gets to its goal-and all of this with the design requirement that the spacecraft be the smallest and lightest it can be while still operating as reliably as possible.

Basic Technology

All spacecraft designs have some common features: structure and materials, electrical power and storage, tracking and guidance, thermal control, and propulsion.

Structure and materials. The spacecraft structure is designed to survive the forces of launch (3 g for a space shuttle launch) and to be lifted by cranes, attached to the vehicle, and otherwise handled during ground processing. The structure is made of metals (aluminum, beryllium, magnesium, titanium) or a composite (boron/epoxy, graphite/epoxy). It must also fit the envelope of the launcher. For the space shuttle this is 5 m (16.5 ft) in diameter and the length of a bus. For expendable launchers, the spacecraft is enclosed in a fairing, a cylindrical volume that varies with the launcher. The volume of the fairing and shuttle is not large enough to permit fully deployed solar arrays and other devices during launch. The spacecraft must also have structures that extend out so that sensors will be far enough away from sources of interference, such as radioisotope thermal electric generators (RTGs), to be useful, or outside the magnetic fields of the spacecraft. See SPACE SHUTTLE.

The use of fluids and gases presents unique problems. In the case of the Hubble Space Telescope, the optics are so sensitive that traditional thrusters cannot be used-they would deposit contaminants on the mirrors and ruin the clarity of the optics. Batteries, heat pipes, ammonia coolant loops, and thrusters must be analyzed to determine how their leakage would affect the mission. For long space missions, care must be taken to have exterior surfaces that are resistant to atomic oxygen, high solar intensity, or extremes of heat and cold. Sliding surfaces or mechanical bearings requiring lubricants are usually placed inside a sealed container. When they are not, liquid lubricant replenishment is provided, or special solid lubricants may be used. Graphite, which is an excellent lubricant on Earth, becomes a harsh lubricant in space when absorbed gases leave its surface. Special dry lubricants or magnetic bearings need to be used. *See* SPACE TELESCOPE, HUBBLE; SPACECRAFT STRUCTURE.

Thermal control. The average temperature of a spacecraft is determined by equating the heat input with the heat radiated away. The major power input for spacecraft located between Earth orbit and the orbit of Mercury is solar radiation. This is equal to the product of the solar constant (approximately 1370 W/m² near the Earth and 5900 W/m² near Mercury), the area facing the Sun, *a*, and the average absorptivity of the surface, α (0 to 1). The heat radiated away is the product of the emissivity of the surface, ϵ (0 to 1), the surface area, *A*, the Stefan-Boltzmann constant σ [5.67 × 10⁻⁸ W(m²/K⁴)], and the fourth power of the absolute temperature. For a sunlit spacecraft that does not generate heat, the temperature can be obtained from Eq. (1). Note that the

$$\sigma T^4 = \frac{\alpha}{\epsilon} \frac{a}{A} S \tag{1}$$

temperature varies with the properties of the surface (α/ϵ) , the orientation of the spacecraft (a/A), and the intensity of the solar radiation (*S*). As the spacecraft's time in space increases, the surface characteristics change due to ultraviolet exposure, atomic oxygen attack, and micrometeoroid impact. *See* SOLAR CONSTANT.

Spacecraft generally radiate heat internally; thus the above equation is a simplification. Radio transmitters generate many watts of heat, as do other electrical devices on board. Externally, the reflected light from the Earth or other celestial bodies adds to the external heat absorbed. The spacecraft's temperature will drop radically when these power inputs drop, as when the spacecraft passes into the shadow of the Earth and the solar radiation is eclipsed. For example, the temperature of an eclipsed deployed solar array of a spacecraft in geostationary orbit may drop to $-100 \text{ K} (-280^{\circ}\text{F})$.

Temperatures inside the spacecraft depend on thermal radiation between surfaces, thermal conduction through solid components, and heat dissipation of electrical components at various locations. These in turn depend on the surface coating (paint, metallic surface, or other coating). To maintain temperatures at acceptable limits, various active and passive devices are used: coatings or surfaces with special absorptivities and emissivities, numerous types of thermal insulation, such as multilayer insulation (MLI) and aerogel, mechanical louvers to vary the heat radiated to space, heat pipes, electrical resistive heaters, or radioisotope heating units (RHUs).

Multilayer insulation has been used extensively on spacecraft. It consists of layers of insulation between layers that reflect radiation. The atmosphere between the layers is removed in the vacuum of space, which prevents the conduction of heat. It is tens to hundreds of times more effective than foam or fiberglass batting.

Heat pipes are sealed tubes or rectangular pipes in which a liquid is evaporated at the hot end and the vapor flows to the cold end where it condenses and travels back to the hot end via a wick. They have the advantage of no moving parts and transferring 200 times the heat of a solid pipe of the same crosssectional area. *See* HEAT PIPE; TEMPERATURE.

Tracking and positioning. The location of a spacecraft can be measured by determining its distance from the transit time of radio signals or by measuring the direction of received radio signals, or by both. Data are usually taken over a period of hours or days, and the best orbit fit is determined numerically by computer. Distances are often measured by transmitting tones of different frequencies and finding the phase relations of the returned signal. The Intelsat ranging system uses four frequencies from 35 to 27,777 Hz, while the Goddard Space Flight Center range and range rate (GRARR) system transmits tones from 8 to 500,000 Hz. The direction of a spacecraft can be determined by turning the Earth station antenna to obtain the maximum signal, or by other equivalent and more accurate methods. See ASTRONAUTICS; CELESTIAL MECHANICS; GUIDANCE SYSTEMS; ORBITAL MOTION; SPACE NAVI-GATION AND GUIDANCE.

Propulsion. The velocity of a spacecraft is changed by firing thrusters. Solid propellant thrusters are rarely used. Liquid propellant thrusters are either monopropellant or bipropellant. Monopropellant thrusters use a fuel such as hydrazine or hydrogen peroxide that decomposes when it comes in contact with a catalyst. Bipropellant thrusters use a fuel and oxidizer that ignite when the two combine in the thruster. Electric thrusters, such as mercury or cesium ion thrusters, have also been used. Electric thrusters have the highest efficiency (specific impulse) but the lowest thrust. The Deep Space 1 spacecraft, which visited an asteroid and was expected to visit Comet Borrelly in late September 2001, uses electric thrusters to accelerate xenon ions to 109,435 km/h (68,000 mi/h). See HYDRAZINE; INTER-PLANETARY PROPULSION; ION PROPULSION; PROPEL-LANT; SPACECRAFT PROPULSION; SPECIFIC IMPULSE.

Attitude control. Most spacecraft are spin-stabilized or are three-axis body-stabilized. The former uses the principles of a gyroscope; the latter uses sensors and thrusters to maintain orientation. Some bodystabilized spacecraft (such as astronomical observatories) are fixed in inertial space, while others (such as Earth observatories) have an axis pointed at the Earth and rotate once per orbit. A body-stabilized spacecraft is simpler than a spinner but requires more hardware. The orientation of a spacecraft is measured with Sun sensors (the simplest method), star trackers (the most accurate), and horizon (Earth or other body) or radio-frequency (rf) sensors (usually to determine the direction toward the Earth). Attitude corrections are made by small thrusters or by reaction or momentum wheels; as the motor applies a torque to accelerate or decelerate the rotation, an equal and opposite torque is imparted to the spacecraft. Torques can also be exerted by turning the axis of a momentum wheel. When a wheel reaches its mechanical limit of rotation speed,

momentum is "dumped" by firing thrusters; that is, as the wheel decelerates, the thrusters are fired to exert a torque equal and opposite to the torque of the decelerating wheels. Momentum wheels are used on the Hubble Space Telescope because thrusters would deposite contaminants on its mirror. To reduce size and weight, momentum wheels are being considered for both attitude control and power storage. *See* STAR TRACKER.

Electrical power. Primary electrical power is most often provided by solar cells made from a thin section of crystalline silicon protected by a thin glass cover. Gallium arsenide solar cells are beginning to replace silicon, despite their increased weight, because they are more efficient. The crystal contains dopants to produce extra electrons on one side and holes on the other side, forming a pn junction. The *n*-type material is on the sunlit surface, since it is more resistant to the damage caused by the radiation in the Van Allen belts. Sunlight creates hole-electron pairs in the junction, and the electric field caused by the space charge layer in the junction sends the electrons toward a metal grid, and the holes toward the base electrode. A 2×4 cm (0.8 \times 1.6 in.) cell provides about 150 mW of power at 10-18% efficiency for silicon cells and up to 23% for gallium arsenide cells. See SOLAR CELL.

Excess power from the solar cells is stored in rechargeable batteries so that when power is interrupted during an eclipse, it can be drawn from the batteries. Nickel-cadmium batteries were used widely but are being replaced by nickel-hydrogen batteries, which provide higher energy per weight and higher energy per volume, properties that are very desirable for spacecraft. Other batteries with even better properties are being tested, such as lithium ion batteries, which up to now have been considered too dangerous, since they can explode or catch fire. Electrical power can be stored with momentum wheels. On the International Space Station, a flywheel energy storage system was being designed to replace the nickel-hydrogen batteries. The lifetime of batteries is limited by the temperature control maintained and the depth of discharge during operation. The latter is the fraction of power taken from the battery during each charge-discharge cycle. Flywheels are not limited by temperature or depth of discharge. They can be charged and discharged indefinitely. See BATTERY.

Other sources of power generation include fuel cells, radioisotope thermoelectric generators (RTGs), tethers, and solar dynamic power. Fuel cells have been used on the Apollo and space shuttle programs and produce a considerable amount of power, with drinkable water as a by-product. Fuel cells use the same principles as a battery, but employ a fuel and oxidizer to produce electricity and water at very high efficiencies (70%). RTGs use the heat of decay of radioisotopes to produce electricity. They are very useful for spacecraft at distances of Mars and beyond where the solar flux would require impractically huge arrays of solar cells to produce the same amount of power. RTGs have low efficiencies (9%), and the radioisotope of choice, plutonium, is very expensive. Tethers use a kilometers-long conducting wire moving through the magnetic field of the Earth to generate electricity, the same principle as a generator. They have been tested on the TSS-1R mission and produced voltages as high as 3500 V and current levels of 480 mA. Solar dynamic power is being considered for the space station. It uses the Sun's energy to heat a fluid to turn a turbine, much like terrestrial power plants. These devices have the advantage of high power and high efficiency (30%) while being small and not producing as much atmospheric friction as the large solar arrays. See ELECTROMAGNETIC INDUCTION; FUEL CELL; NUCLEAR BATTERY; SOLAR ENERGY; SPACE POWER SYSTEMS.

Telemetry and command. The status and condition of a spacecraft are determined by telemetry. Temperatures, voltages, switch status, pressures, sensor data, and many other measurements are transformed into voltages, encoded into pulses, and transmitted to Earth. This information is received and decoded at the spacecraft control center. Desired commands are encoded and transmitted from the control center, received by the satellite, and distributed to the appropriate subsystem. Commands are often used to turn equipment on or off, switch to redundant equipment, make necessary adjustments, and fire thrusters and pyrotechnic devices. *See* SPACE COM-MUNICATIONS; SPACECRAFT GROUND INSTRUMENTA-TION; TELEMETERING.

Special-Purpose Technology

Many spacecraft missions have special requirements and hence necessitate special equipment.

Space probes. Satellites that leave the Earth's gravitational field to travel around the Sun and visit other planets have special requirements. Distances are greater (on the order of 1.8×10^8 mi or 3×10^8 km), mission times are longer (years), and the intensity of the Sun's radiation may change markedly (inversely proportional to the square of the distance from the Sun). With greater distances, communications are more difficult, and information must be transmitted at a slower rate. For probes that go closer to the Sun, the higher intensity of sunlight poses a severe thermal problem. For probes that go to the outer planets (Jupiter, Saturn, and beyond), the decrease in solar intensity makes solar cells less effective, requiring other power sources such as radioisotope thermoelectric generators. See SPACE PROBE.

Reentry vehicles. Spacecraft that return to Earth require special protection. Retro-rockets are used to reduce the orbital velocity so that the satellite falls and enters the Earth's atmosphere. Most of the decrease in velocity is produced by atmospheric friction; in the last stage this may be augmented by deploying parachutes. On entering the atmosphere, the velocity is extreme (4.4–6.9 mi/s or 7–11 km/s), and the friction generates high temperatures (thousands of degrees). One type of protection is an ablative

shield. This shield, which was used on *Apollo* capsules, was designed to melt and vaporize, dissipating heat in the process. On the shuttle, a special thermal shield made of glass fibers can withstand the high temperatures, with enough insulation to protect the interior.

The shape of a reentry spacecraft is designed to maintain a stable and desired attitude during reentry. Communications during part of the reentry are cut off by the high-density plasma sheath formed around the spacecraft, through which radio-frequency signals will not penetrate. Final recovery may be accomplished in the air (film reconnaissance satellite, as in the Discoverer program), in the ocean (Apollo astronauts), or on land (shuttle and Russian cosmonauts).

For missions to other planets, aerobraking has been used. Aerobraking uses the atmosphere of the planet instead of using retrorockets to slow the spacecraft. The *Magellan* spacecraft used aerobraking in 1994 to achieve orbit of Venus, and the *Mars Global Surveyor* used aerobraking in the much thinner atmosphere of Mars to achieve orbit. *See* ATMO-SPHERIC ENTRY; NOSE CONE.

Crewed spacecraft. Technology must be developed to meet the many needs associated with human survival in space. Oxygen equivalent to one-fifth of an atmosphere is needed for breathing; the remaining nitrogen is desirable but not essential. The early Soviet cosmonauts had an atmospheric pressure essentially equivalent to that on Earth; United States astronauts had much less nitrogen, with a total pressure of only one-third of an atmosphere. This lower pressure is satisfactory for breathing, but increases the fire hazard and does not carry sound as far (only 16 ft or 5 m for normal conversation). Provision must be made for adding oxygen and removing carbon dioxide from the air, for supplying food and water, and for disposing of human wastes. For extravehicular activity (EVA), a special suit is required to provide oxygen for breathing and to maintain adequate pressure.

The lack of gravity produces a number of technological requirements for astronauts living in space. Physically, it is just as easy (or just as hard) to walk on the ceiling as to walk on the floor, but an astronaut feels more comfortable in a room with a local vertical, that is, with one surface that looks like a floor and another that looks like a ceiling. Objects tend to float in space so that restraints are necessary: mechanical, elastic, or magnetic. Showers require special vacuum hoses to remove the water, since it does not run down the drain. Astronauts can be weighed by rocking in a special chair fitted with springs; the astronaut's mass is determined from the frequency of rocking. *See* SPACE BIOLOGY; SPACE FLIGHT; SPACE STATION; WEIGHTLESSNESS.

Rendezvous. In some missions one spacecraft must find, approach, and make contact with another spacecraft. One active spacecraft performs the maneuvering by firing thrusters, while the other spacecraft (target) remains passive. During the far approach, the orbits of the two spacecraft are deter-

mined by conventional means, and maneuvers are calculated to bring the spacecraft closer together. During the near approach, the position of the target spacecraft is determined directly from the active spacecraft, sometimes by radio and sometimes by visual contact. For final approach, a closing velocity is maintained with the proper attitude, so that contact is made with the desired surfaces. Usually a probe mates with an appropriate drogue, hooks hold the two spacecraft together, and mechanical devices then firmly attach one spacecraft to the other.

For the space shuttle, the Ku-band antenna also serves as a radar antenna for final approach and rendezvous. When the shuttle docks with the *International Space Station*, it uses docking hardware manufactured in Russia.

Reliability

Space is distant not only in kilometers but also in difficulty of approach. Large velocity changes are needed to place objects in space, which are then difficult to repair and expensive to replace. Therefore spacecraft must function when they are launched, and continue to function for days, months, or years. The task is similar to that of building a car that will go 125,000 mi (200,000 km) without requiring mechanical repair or refueling. Not only must space technology build a variety of parts for many missions, but it must achieve a reliability far greater than the average. This is accomplished by building inherent reliability into components and adding redundant subsystems, supported by a rigorous test schedule before launch. Efforts are made to reduce the number of single points of failure, that is, components that are essential to mission success and cannot be bypassed or made redundant.

Electronics. The failure rate for an electronic subsystem is often estimated by adding up the failure rates of its components. The failure rate of each component is assumed to be independent of that of other components, and the failure of a subsystem. While electronic parts are highly reliable, a subsystem has so many parts that there may be an appreciable chance of failure. Component failure rates are minimized by special manufacture, quality control, and derating, that is, operation at voltages or temperatures which are lower than the maximum permitted.

If an electronic subsystem does not have the desired reliability, a redundant unit can be included. If either of the two subsystems continues to operate, the mission can still be accomplished. Mathematically, the probability that at least one of the subsystems is operating, R_{12} , is given by Eq. (2), where *R* is

$$R_{12} = 1 - (1 - R)^2 \tag{2}$$

the reliability of the single subsystem. For example, if each subsystem has a reliability, R, of 0.98, then the probability that at least one subsystem is working is 0.9996, which is a significant improvement. The equation above is based on the assumption that failures of the two subsystems are not related; this may not be true when two units are built by the

same manufacturer at the same time. *See* RELIABIL-ITY, AVAILABILITY, AND MAINTAINABILITY.

Mechanical systems. While there are fewer mechanical parts than electrical parts on most spacecraft, mechanical parts tend to be less reliable than electrical ones. Deployment mechanisms usually work only once, but must work correctly the first time. Bearings must be lubricated for their expected lifetimes, and special lubricants must be used. Torque margins are increased, so that even if unusually high torques are encountered, the motors will still keep the device turning. Electromechanical relays should have currents reduced before they are opened, to prevent arcing and welding of contacts in vacuum.

Testing. The key to a reliable spacecraft is extensive testing before launch. This ensures not only that every spacecraft part is adequately functioning, but also that the spacecraft is subjected to the differing environments expected in space, and in some cases is overstressed to find the weak components. Tests are performed first on components, then subsystems, and finally, when feasible, on fully assembled spacecraft. The testing program requires many weeks and consists of three main parts: electrical performance, vibration, and thermal-vacuum tests. Electrical testing includes evaluation of the spacecraft's ability to perform a variety of operations which will be required in space. Antennas may be separately tested on an antenna range, where the effect of reflections from nearby objects is minimized. Steady-state acceleration may be evaluated with a static test, in which mechanical loads are placed on various parts of the structure, or with a large centrifuge. Vibration tests, simulating the launch environment, subject the spacecraft to low-frequency vibration (5-100 Hz), acoustic vibration (20-10,000 Hz), and pyrotechnic shock conditions. Many of these tests can be accomplished with an electromechanical vibrator. Thermalvacuum tests consist of placing the spacecraft in a high-vacuum chamber (less than 10^{-5} torr or 10^{-3} pascal) and simulating the temperatures that will occur in space. See ENVIRONMENTAL TEST; INSPEC-TION AND TESTING; MECHANICAL VIBRATION; QUAL-ITY CONTROL.

Sometimes the first spacecraft tested (prototype) remains on the ground, and only subsequent (flight) spacecraft are launched. The prototype spacecraft is often less reliable because of modifications made, tests which must be repeated, and planned overstressing in tests. Most spacecraft perform their missions successfully, but often unexpected anomalies occur. Redundant components may fail, and subsystems, although performing adequately, may not quite meet design specifications. Telemetered performance is analyzed, problems are diagnosed, and changes are proposed in the design and testing of future spacecraft. In a few cases, space technology improves the spacecraft performance in space, and appropriate rf commands are sent to modify its operation, to reset parameters, to modify software, or to fire thrusters which will reorient the spacecraft or shake it to expel a bubble in a fuel line.

Testing must be timed correctly during the assembly process. The Tethered Satellite System 1 was fully tested to ensure that the satellite with its tether would deploy properly. A modification was then approved that ultimately caused the tether not to deploy fully, resulting in a partially failed mission. The International Space Station has the unique problem that some of the components are not connected to each other until they are in orbit. NASA has a system using state-of-the-art computer-aided-designand-manufacturing (CAD/CAM) software and photography to determine that the components will mate to each other mechanically. NASA also decided to conduct tests to ensure that the electrical connections from one component to another were compatible and that the software would command and control each of the elements. NASA and Boeing had to build emulators of the components already in orbit, and will continue to do so as the station is assembled. These multielement integrated tests add further expense, but are necessary to prove that the space station will work when fully assembled. See ASTRONAU-TICAL ENGINEERING; COMPUTER-AIDED DESIGN AND MANUFACTURING; SATELLITE (SPACECRAFT); SPACE. Jeffrey C. Mitchell; Gary D. Gordon

Bibliography. J. A. Angelo, Jr., and D. Buden, *Space Nuclear Power*, Orbit Book Company, 1985; G. D. Gordon and W. L. Morgan, *Principles of Communications Satellites*, 1993; M. D. Griffin and J. R. French, *Space Vebicle Design*, 2d ed., 2004; C. M. Harris and A. G. Piersol, *Harris' Sbock and Vibration Handbook*, 5th ed., 2002; M. H. Kaplan, *Modern Spacecraft Dynamics and Control*, 1976; G. Maral and M. Bousquet, *Satellite Communication Systems*, 4th ed., 2002; W. L. Morgan and G. D. Gordon, *Communications Satellite Handbook*, 1989; G. P. Sutton and O. Biblarz, *Rocket Propulsion Elements*, 7th ed., 2001.

Space-time

A term used to denote the geometry of the physical universe as suggested by the theory of relativity. It is also called space-time continuum. Whereas in newtonian physics space and time had been considered quite separate entities, A. Einstein and H. Minkowski showed that they are actually intimately intertwined. I. Newton's ideas on space and time are summarized in the following list:

1. Given two events, each of which is clearly localized in space and lasts only for an instant in time, such as two strokes of lightning striking small targets, all observers of the events will be in agreement as to which of the two events took place earlier in time, or whether they were actually simultaneous.

2. If the events were not simultaneous, the interval of time between them is an absolute entity, agreed on by all competent observers.

3. The spatial distance between two simultaneous events is an absolute entity, agreed on by all competent observers.

Of these three, the first assumption, concerning the concept of simultaneity of distant events, is the crucial one, the other two depending on it. Simultaneity, however, can be given an unambiguous meaning only if there is available some instantaneous method of signaling over finite distances. Actually, according to the theory of relativity, the greatest speed of transmission of intelligence of any kind is the speed of light *c*, equaling about 3×10^{10} cm/s. Moreover, any signal traveling precisely at the speed c appears to travel at that same speed to all conceivable observers, regardless of their own states of motion. This is the only reasonable interpretation of the results of the Michelson-Morley experiment and the effect of aberration. Accordingly, the question of whether two given events are simultaneous or not can be decided only with the help of signals that at best have traveled from the sites of these events to the station of the observer at the speed of light. See ABERRATION OF LIGHT; LIGHT; RELATIVITY.

Under these circumstances, Einstein showed that in general two observers, each using the same techniques of observation but being in motion relative to each other, will disagree concerning the simultaneity of distant events. But if they do disagree, they are also unable to compare unequivocally the rates of clocks moving in different ways, or the lengths of scales and measuring rods. Instead, clock rates and scale lengths of different observers and different frames of reference must be established so as to assure the principal observed fact. Each observer, using his or her own clocks and scales, must measure the same speed of propagation of light. This requirement leads to a set of relationships known as the Lorentz transformations. *See* LORENTZ TRANSFORMATIONS.

In accordance with the Lorentz transformations, both the time interval and the spatial distance between two events are relative quantities, depending on the state of motion of the observer who carries out the measurements. There is, however, a new absolute quantity that takes the place of the two former quantities. It is known as the invariant, or proper, space-time interval τ and is defined by Eq. (1), where

$$\tau^2 = T^2 - \frac{1}{c^2} R^2$$
 (1)

T is the ordinary time interval, *R* the distance between the two events, and *c* the speed of light in empty space. Whereas *T* and *R* are different for different observers, τ has the same value. In the event that Eq. (1) would render τ imaginary, its place may be taken by σ , defined by Eq. (2). If both τ and σ are

$$\sigma^2 = R^2 - c^2 T^2 \tag{2}$$

zero, then a light signal leaving the location of one event while it is taking place will reach the location of the other event precisely at the instant the signal from the latter is coming forth.

The existence of a single invariant interval led the mathematician Minkowski to conceive of the totality of space and time as a single four-dimensional continuum, which is often referred to as the Minkowski universe. In this universe, the history of a single space point in the course of time must be considered as a curve (or line), whereas an event, limited in both space and time, represents a point. So that these geometric concepts in the Minkowski universe may be distinguished from their analogs in ordinary threedimensional space, they are referred to as world curves (world lines) and world points, respectively.

Minkowski geometry. The geometry of the Minkowski universe in some respects resembles the geometry of ordinary (euclidean) space but differs from it in others. The Minkowski universe has four dimensions instead of the three dimensions of ordinary space; that is to say, for a complete identification a world point requires four pieces of data, for instance, three space coordinates and a time reading. But there are in the Minkowski universe world points, world lines, two-dimensional surfaces (including planes), three-dimensional surfaces (often called hypersurfaces), and four-dimensional domains. A hypersurface may, for instance, be a spatial domain (volume) at one instant in time, or it may be a two-dimensional (ordinary) surface for an extended period of time. Thus one may form all the geometric figures that are also possible in a four-dimensional euclidean space.

The indefinite metric. In a euclidean space there are the cartesian coordinate systems, those rectilinear systems of coordinates that are mutually perpendicular and whose coordinate values correspond to real lengths. The distance S between two points whose coordinate differences are X, Y, and Z, respectively, is given by Eq. (3). The coordinate transformations

$$S^2 = X^2 + Y^2 + Z^2 \tag{3}$$

that lead from one cartesian coordinate system to another are called orthogonal coordinate transformations. Formally, they are the coordinate transformations that preserve the precise form of Eq. (3). Likewise, the Lorentz transformations preserve the precise form of Eqs. (1) and (2), respectively. A coordinate system in which these two equations hold is called a Lorentzian frame of reference.

Whereas the form on the right of Eq. (3) is positive definite, that is, always greater than or equal to zero, the right-hand sides of Eqs. (1) and (2) are indefinite; that is, they may be positive or negative. This fact represents the single but all-important difference between a four-dimensional euclidean space and a Minkowski universe. The forms (1), (2), and (3) are called metrics. Hence the Minkowski universe is said to possess an indefinite metric.

As a result of the indefinite character of the metric, a triangle in the Minkowski universe may possess one side that is longer than the sum of the two others; conversely, one of the three sides may have the length zero. Depending on whether τ or σ is real, or both vanish, an interval is classified as timelike, spacelike, or a null-interval (see **illus.**)

Improper Lorentz transformations. In the Minkowski universe there are three different types of Lorentz transformations involving some kind of reflection, in addition to the more usual Lorentz transformations

(called proper Lorentz transformations). The first type of improper Lorentz transformation changes the sign of all three spatial coordinates but leaves the sense of the time axis unchanged. This transformation changes a right-handed screw into a left-handed screw; it is also called a parity transformation. The second type of improper Lorentz transformation interchanges the future with the past but leaves the space coordinates unchanged. This transformation is called time reversal. The third improper Lorentz transformation reflects both the space and the time coordinates; it bears no special name of its own. The original arguments which led to the formulation of the special theory of relativity all support the proposition that the laws of nature are invariant under proper Lorentz transformations. They are inconclusive as to whether the laws of nature should also be invariant under the improper Lorentz transformations. The laws of mechanics and of electrodynamics have this property; the second law of thermodynamics distinguishes between past and future in a nonsymmetric manner, but it is usually assumed that this feature of thermodynamics is to be explained by its statistical nature and that it has no bearing on the properties of the underlying basic dynamics. See SYMMETRY LAWS (PHYSICS).

Until recent times, it had therefore been taken for granted that the basic laws of nature should make no distinction between right-handed and left-handed screws, and that they should not discriminate between future and past. Certain difficulties in the interpretation of the decay of *K* mesons led T. D. Lee and C. N. Yang to suspect that these assumptions might not be tenable, and they suggested some experiments on meson decay and on radioactive beta decay, which showed that in these particle transformations nature certainly discriminates between left-handed and right-handed screws. *See* PARITY (QUAN-TUM MECHANICS).

Neutrinos are electrically neutral particles which travel at the speed of light and are endowed with an intrinsic spin equal to that of electrons. It appears that the neutrino always spins counterclockwise when viewed in the direction of its travel, while its antiparticle, the antineutrino, always spins clockwise. (As far as is known, this holds true of each of the presently known varieties of neutrinos.) The exact role of the improper Lorentz transformations in nature is under intensive theoretical and experimental investigation. *See* NEUTRINO.

Curved space-time. Whereas the Minkowski universe is the appropriate geometric model for the special theory of relativity, the general theory of relativity makes use of a further generalization. In the Minkowski universe a particle that is not subject to external forces and which therefore travels along a straight line (in ordinary space) and at a uniform speed is represented by a straight world line. In general relativity, an external gravitational force is indistinguishable from an inertial force, which in special relativity would arise if a non-lorentzian frame of reference were to be employed. Accordingly one requires a four-dimensional space in which it is im-



Two-dimensional sketch of three possible classes of intervals. The null directions actually form a three-dimensional surface having the shape of a double cone, called the light cone. An alternative Lorentz frame is indicated by broken lines.

possible to distinguish between a lorentzian and a non-lorentzian frame of reference whenever a gravitational field is present. Such a space cannot be flat, as the Minkowski universe is, but must be curved. Such an aggregate is called a riemannian space. In a general riemannian space there are no straight lines but only curves. *See* FRAME OF REFERENCE; GRAVITA-TION; RIEMANNIAN GEOMETRY. Peter G. Bergmann

Spacecraft ground instrumentation

Instrumentation located on the Earth for monitoring, tracking, and communicating with satellites, space probes, and crewed spacecraft. Radars, communication antennas, and optical instruments are classified as ground instrumentation. They are deployed in networks and, to a lesser extent, in ranges. Ranges are relatively narrow chains of ground instruments used to follow the flights of missiles, sounding rockets, and spacecraft ascending to orbit. Some ranges are a few miles long; others, such as the U.S. Air Force's Eastern Test Range, stretch for thousands of miles. Networks, in contrast, are dispersed over wide geographical areas so that their instruments can follow satellites in orbit as the Earth rotates under them at 15° per hour, or space probes on their flights through deep space.

Networks are of two basic kinds: networks supporting satellites in Earth orbit, and networks supporting spacecraft in deep space far from Earth. A third concept was added in the 1980s with the Tracking and Data Relay Satellite System (TDRSS), also called the Space Network (SN). TDRSS replaced most of the ground stations used for Earth orbital support. The *Tracking and Data Relay Satellites (TDRS)* are
placed in geosynchronous orbits to relay signals to and from other orbiting spacecraft during more than 85% of each orbit, to and from a single ground station.

Functions. Ranges and networks have various technical functions:

1. Tracking: determination of the positions and velocities of space probes and satellites through radio and optical means.

2. Telemetry: reception of telemetered signals from scientific instruments and spacecraft house-keeping functions.

3. Voice reception and transmission: provision for communication with the crew of a spacecraft, such as the space shuttle.

4. Command: transmission of coded commands to spacecraft equipment, including scientific instruments.

5. Television reception and transmission: provision for observation of the crew, spacecraft environment, and so on.

6. Ground communications: telemetry, voice, television, command, tracking data, and spacecraft acquisition data transmission between network sites and the central mission control center, and payload information to user facilities.

7. Computing: calculation of orbital elements and radar acquisition data prior to transmission to users; also, computation of the signals that drive visual displays at a mission control center. *See* TELEMETERING.

Mission control. From the mission control center, highly trained mission personnel, observing computer-generated displays, dispatch commands and analyze acquired data through ground communication with the appropriate network stations. From these stations, command data are dispatched to orbiting satellites.

Tracking can be accomplished by a great variety of instruments, broadly classified as optical or radiofrequency (rf). In the optical class are cinetheodolites, ballistic cameras, Baker-Nunn cameras, telescopes, and laser ranging systems. Radio-frequency tracking systems are even more varied and include radar, active and passive radio interferometers, radio Doppler systems, and range-and-range-rate systems, which resemble Doppler radars except that they do not measure spacecraft direction. *See* DOPPLER RADAR; OPTICAL TRACKING SYSTEMS; RADAR.

Ranges. The best-known range in the United States is the Eastern Test Range (ETR; formerly called the Atlantic Missile Range), which begins at Cape Canaveral, Florida. In addition, the Department of Defense operates the Western Test Range (WTR) and several others. The National Aeronautics and Space Administration (NASA) launches most of its spacecraft from the Eastern Test Range, but also launches some sounding rockets and small satellites from its Wallops Island facility on the Virginia east coast. During the shuttle era, the Eastern Test Range performs the functions of both the Eastern Test Range and the Western Test Range for the shuttle. Each of the United States ranges possesses tracking equipment, telemetry receivers, and electronic equipment to perform its functions. Many other countries have missile and sounding rocket ranges.

The Eastern Test Range is one of the longest ranges in the world. The range head begins with the NASA and military launch pads and blockhouses, on Cape Canaveral proper, and those of NASA on nearby Merritt Island, Florida. A missile or spacecraft launched from one of these areas climbs vertically, then pitches toward downrange tracking and communication antennas. As the rocket moves over the Atlantic, it passes within range of stations instrumented to track, command, and receive telemetry from the rising launch vehicle and its payload. *See* LAUNCH COM-PLEX.

Cape Canaveral itself is well supplied with instruments such as the C-band AN/FPS-16 radar, many tracking telescopes, and several large Doppler and interferometer tracking systems. Similar instruments along the range watch each flight. The range stations are connected to each other by submarine cable and radio and satellite communication links. Geodetic surveys, timing signals from radio station WWV of the National Institute of Standards and Technology, and standardized operating procedures also unify the range stations.

Spacecraft networks. NASA operates two groundbased networks (Fig. 1) and the TDRSS. The groundbased networks are the Spaceflight Tracking and Data Network (STDN), which tracks, commands, and receives telemetry from United States and foreign satellites in Earth orbit; and the Deep Space Network (DSN), which performs the same functions for all types of spacecraft sent to explore deep space, the Moon, and solar system planets. The TDRSS provides the same support to Earth orbital spacecraft as the STDN. The U.S. Department of Defense operates two classified networks: the Satellite Control Facility (SCF); and the National Range Division Stations, which include those of all United States military ranges. Russia and European Space Agency (ESA) also maintain similar networks.

Spaceflight Tracking and Data Network (STDN). The STDN is a complex of ground-based stations designed to support NASA's Earth-orbiting scientific and applications spacecraft and crewed spaceflight programs such as the space shuttle. The operations control center and the central computing facility for operation and analysis are located at the Goddard Space Flight Center in Greenbelt, Maryland (Fig. 1).

The STDN, operational since May 1971, combined the former Space Tracking and Data Acquisition Network (STADAN) and the Manned Space Flight Network (MSFN). The crewless STADAN tracked spacecraft in Earth orbits from ground facilities distributed over much of the globe. The minimum tracking (minitrack) facilities, implemented in 1958, used antennas composed of steel rails parallel to the ground. The minitrack's narrow fan pattern received radio transmissions from the spacecraft as it passed overhead. The STADAN stations used parabolic (dish) antennas whose conical beams swept the sky to follow the orbiting spacecraft.



Fig. 1. World map showing the tracking networks for space operations of the National Aeronautics and Space Administration (NASA).

The MSFN provided two-way contact between the ground and space for the astronauts in the Mercury, Gemini, and Apollo programs. It consisted of both ground stations and tracking and support ships that could be positioned where ground stations were unavailable or impractical. The stations were configured to support the individual requirements of each mission whether the mission involved Earth orbit or Moon landing.

In general, the systems needed at each STDN station are determined by the support that the station provides and whether the support is directed primarily toward crewed or crewless space efforts. Within the STDN, the unified S-band (USB) system is the prime equipment dedicated to tracking, telemetry, and two-way communications. It is so named because voice and data can be handled on one S-band carrier. Stations that use the USB system are further identified by the size of their USB antenna systems: 85-ft (26-m) or 30-ft (9-m) dish diameter.

The significant disadvantage of the STDN is that the antennas are bound to the surface of the Earth. This creates a limited field of view because the antenna can usefully transmit and receive only when a spacecraft is overhead. Thus, the average coverage that a low-altitude spacecraft may expect to receive from the tracking stations is limited to approximately 15% of its orbit. This amount of coverage is insufficient to support newer, more sophisticated spacecraft with instruments that generate much more information. It became impractical to store this information on board the spacecraft during the long periods when the spacecraft is out of contact with ground stations. The most practical solution was to construct the TDRSS.

The TDRSS network has replaced most of the STDN stations, substituting the space segment and two ground stations for most of the ground stations within the STDN. The remaining STDN stations provide support to older spacecraft, which are not compatible with the TDRSS, and the special support,

such as launch and landing support to the shuttle.

Deep Space Network (DSN). NASA, through the Jet Propulsion Laboratory of the California Institute of Technology in Pasadena, has established a Deep Space Network (DSN), consisting of three Deep Space Communication Complexes (DSCC) located at Goldstone, California; Madrid, Spain; and Canberra, Australia. The complexes are approximately 120° apart in longitude, which makes it possible to provide continuous communications for planetary spacecraft as the Earth rotates. From its beginning in 1959, the DSN has supported all the deep space probes such as Mariner, Pioneer, Viking, and Voyager, as well as assisting during the Apollo crewed lunar missions. While TDRSS began replacing STDN in 1985, DSN services continue to be used for Earth orbiters above 7500 mi (12,000 km) because TDRSS antenna pointing is designed only for spacecraft up to this altitude.

The Network Control Center at the Jet Propulsion Laboratory is the single point of control for the DSN. The Network Control Center handles messages transmitted to and from the spacecraft, and validates DSNgenerated radio navigation data. The data are then sent to Mission Operations for flight path control. Additionally, the Network Control Center performs research, development, and engineering for the DSN.

Each DSCC has three or four antennas. A main data-capture instrument is the 230-ft (70-m) transmit-receive antenna with a parabolic surface. The precisely contoured parabolic surface permits the antenna to operate efficiently at short-wavelength radio frequencies (for example, at the X-band frequency, principally used to receive pictures from the outer planets, the wavelength is 3.5 cm). The combination of large area and high-frequency operations gives the 230-ft (70-m) antenna an extremely narrow field of view (one ten-millionth of the sky). The receiver system that receives such weak signals from deep space has a minuscule noise temperature (20 K), hundreds of

times less than that in the best high-fidelity home FM radio receivers. The total power received by each 230-ft (70-m) antenna in a typical outer planet encounter, such as Voyager at Uranus, can be as weak as 10^{-15} W. Yet, the noise-combating codes allow high-resolution pictures to be reconstructed with few errors. For example, in January 1986, the NASA Voyager 2 spacecraft flew by Uranus. A good data rate was maintained by arraying or simultaneously combining signals from the DSN 230-ft (70-m) antenna and the two 112-ft (34-m) DSN antennas at Canberra and the Australian Parkes Observatory 210-ft (64-m) radio astronomy antenna 200 mi (320 km) away. This provided 29,000 bits of information per second, equivalent to a color picture every 5 min. The result was hundreds of color pictures during the brief encounter. Because Uranus was in the southern part of the sky in 1986, the Australian stations viewed the spacecraft longer than the Goldstone or Madrid stations.

In addition to receiving data, stations must send commands to the distant spacecraft to instruct the on-board computers what maneuvers to perform and what experiments to conduct. The antennas at each DSCC can simultaneously send commands to and receive data from the spacecraft. Different frequencies are used for transmitting and receiving; therefore, even though hundreds of kilowatts can be transmitted, the effect on the extremely weak signals being received from the spacecraft is negligible.

While the antenna is receiving data and transmitting commands, radio measurements of spacecraft paths are conducted for navigation. Because the very distant spacecraft often must be targeted to pass within small (5–10-mi or 8–16-km) regions near the target planets or their satellites or rings, extreme directional accuracy is required, in some cases near a ten-millionth of a degree. By receiving signals from the spacecraft at two DSCCs simultaneously, the time delay between reception at the two stations can be measured. Since the speed of light and the distance and direction between the two stations are known, the position of the spacecraft can be determined from trigonometry.

The DSN is also used as a scientific facility in radio science experiments. In these experiments, the changes in the received signal are carefully measured as they pass through a planet's atmosphere or ionosphere, interplanetary space, the solar corona, or the rings around planets. Using these measurements, radio scientists can determine the atmospheric pressure as a function of altitude above the surface. Coupled with spacecraft instruments and Earth-based measurements, such radio science has been a very powerful tool for the exploration of the solar system.

The DSN planetary radar is also used to characterize the bodies in the solar system. Its noteworthy achievements include calibrating the length scale of the solar system (the astronomical unit) to greatly improve radio navigation of planetary spacecraft, discovering the slow backward rotation of the planet Venus, finding craters on Venus, measuring altitudes on Mars, selecting Martian landing sites for the *Viking* Lander, determining that Saturn's rings were composed of large icelike particles, and studying asteroids and comets to determine surface properties and rotation.

By using techniques similar to two-station precision radio navigation, changes in relative positions of particular locations on Earth are measured. This provides a deeper understanding of crustal plate motion and, ultimately, of earthquake mechanisms. The DSN is also used as a radio astronomy observatory, either alone or in cooperation with other observatories. *See* GEODESY; RADIO ASTRONOMY.

Much advanced technology has arisen from the DSN, including the first field-operational amplifying masers for low-noise, weak-signal reception; the first field-operational hydrogen maser clocks for frequency and timing in navigation; power-efficient coded communications, including the first use of block and convolutional coding; on-line signal processing by general-purpose digital computers; and high-speed determination of signal strength as a function of frequency across a very wide frequency band. *See* ATOMIC CLOCK; MASER.

In addition to providing communications for United States deep-space craft, joint deep-space missions with other countries are supported on a cooperative basis. As the only worldwide network that can work with extremely weak signals, the DSN is a valuable resource for other countries.

Laser Tracking Network. The Laser Tracking Network consists of a series of laser systems used for ranging to retroreflector-equipped satellites in highly stable orbits. Laser stations obtain ranging data for these satellites by bouncing a highly concentrated pulse of laser light off the retroreflector corner cube installed on the spacecraft exterior. The exact position of the spacecraft in orbit can then be mathematically determined for a given point in time. By comparing several ranging operations, orbital predictions can be interpolated and extrapolated.

The resultant data have a variety of applications, such as precise prediction of satellite orbit and measurement of the Earth's gravitational field, polar motion, earth tides, Earth rotation, tectonic plate motion, and crustal motion to accuracies within the centimeter range. The Laser Tracking Network is a multinational cooperative effort with over 30 laser sites located in North and South America, Europe, China, Japan, and Australia.

The Laser Tracking Network consists of both fixed and mobile laser ranging systems. The mobile stations can be deployed worldwide to prepared sites. A typical laser mobile system consists of a laser transmitter, a receiver subsystem, the optical equipment, a computer, a timing subsystem, and a surveillance radar. The total system is housed in mobile vans, with a power-generation van required at some sites.

TDRSS. In the plan developed for 1990 and beyond, the DSN provides support to high-altitude Earthorbiting spacecraft. Other than for launch support from the stations at Merritt Island, Florida; Bermuda; and Dakar, Senegal, all support to low-altitude Earthorbiting satellites is from the TDRSS. The TDRSS



Fig. 2. Tracking and Data Relay Satellite System (TDRSS) typical user configuration.

(**Fig. 2**) consists of two operational geostationary satellites, an in-orbit spare, and two colocated ground terminals with associated operations control facilities at White Sands, New Mexico. The two active satellites, called *TDRS-East* and *TDRS-West*, are in geosynchronous orbit roughly above the Equator and at 41 and 171°W longitude, respectively. The TDRSS spare is located at 83°W longitude. The TDRSs relay signals between the ground station and satellites orbiting generally below 7500 mi (12,000 km) above Earth during nearly all of each orbit, except for a zone of exclusion which varies with spacecraft altitude.

Communication between the ground terminals and TDRSS is at Ku-band frequencies via ground terminal antennas. TDRS on-orbit tracking, telemetry, and command are provided by the same facilities. A separate antenna is provided for S-band operations during transfer orbits and as a backup to the Ku-band capability. Each TDRS can support two high-data-rate users simultaneously at either S- or Ku-band with the steerable single-access antennas. Multiple access permits up to 20 users at a time, but with much lower data rates.

The advantage of the TDRSS is that support of low-orbiting spacecraft operations is no longer restricted to short and infrequent periods of mutual STDN-spacecraft visibility. Orbital coverage for lowaltitude Earth-orbiting spacecraft is increased from a maximum of 15% of each orbit, using conventional tracking stations, to 85% with TDRSS. This increase, in turn, made possible the elimination of the majority of the STDN stations and consolidation of their remaining functions with those of the DSN, which can take over responsibilities of both.

The space shuttle *Challenger* carried the first of the three satellites into space in April 1983. Since that time the two other satellites required to complete the constellation were launched, and the system is operational.

Worldwide network. The common element connecting all networks is the worldwide NASA communication system (NASCOM), which links all NASA stations, on all continents except Antarctica, through the use of submarine cables, land lines, microwave data links, and satellite and fiber-optic transmissions. The trend is progressively toward satellite communications as the data rates to be handled continue to increase. The use of satellites permits an Earth terminal to be placed at the user's facility and thus permits the transfers of data in fewer hops. Ground instrumentation provides the vital link between the spacecraft and mission controllers and scientists on the ground. See SATELLITE (SPACECRAFT); SPACE COMMUNICATIONS; SPACE FLIGHT; SPACE NAV-IGATION AND GUIDANCE; SPACE PROBE; SPACE STA-TION H. William Wood

Bibliography. M. Gerard and P. W. Edwards (eds.), *Space Information Systems in the Space Station Era*, 1988.

Spacecraft propulsion

A system that provides control of location and attitude of spacecraft by using rocket engines to generate motion. Spacecraft propulsion systems come in various forms depending on the specific mission requirements. Each exhibits considerable variation in such parameters as thrust, specific impulse, propellant mass and type, pressurization schemes, cost, and materials. All of these variables must be considered in deciding which propulsion system is best suited to a given mission. Typical spacecraft applications include communications satellites, science and technology spacecraft, and Earthmonitoring missions such as weather satellites. Orbital environments range from low-Earth to geosynchronous to interplanetary. See ASTRONAUTI-CAL ENGINEERING; ROCKET PROPULSION; SATELLITE (SPACECRAFT); SPACE FLIGHT; SPACE PROBE; SPECIFIC IMPULSE; THRUST.

Design parameters. The two fundamental variables that define the design of spacecraft propulsion systems are the total velocity change to be imparted to the spacecraft for translational purposes, and the impulse necessary to counteract the various external torques imposed on the spacecraft body. From these, the required quantity of a given propellant combination can be specified by Eqs. (1) and (2).

$$M_{p1} = \left[1 - \exp\left(\frac{-\Delta V}{I_{sp} g}\right)\right] M_{s/c} \tag{1}$$

$$M_{p2} = \frac{I}{I_{sp}} \tag{2}$$

In Eq. (1), M_{p1} is the mass of propellant that is required to be expended to perform the translational maneuver of velocity increment ΔV ; I_{sp} is the specific impulse, which is dependent on the propellant combination and engine design; g is the acceleration of gravity; and $M_{s/c}$ is the mass of the spacecraft prior to the rocket engine burn. In Eq. (2), M_{p^2} is the mass of propellant required to counteract the total disturbance impulse I. The total propellant requirement for a given mission is the sum of these results. A typical propellant budget for a communications satellite is given in the table. Propellant accounts for almost 60% of the lift-off mass of a communications satellite. In this case, the translational ΔV propellant constitutes over 95% of the spacecraft's needs, and stationkeeping (translational) burns use over 95% of the fuel

required during the operational life of the spacecraft (over 10 years).

Most spacecraft require translation in one or more axes as well as rotation about the spacecraft center of mass. The ΔV requirements are dictated by the mission design, launch-site selection, and in-orbit perturbations. For instance, some spacecraft require repositioning throughout the mission. A launch from, for example, French Guiana in South America results in a similar mission to a launch from Cape Canaveral, Florida, but the propellant requirements are substantially affected because the launch vehicles leaving from each site put the spacecraft in different orbits. The spacecraft must then perform translations of different magnitude to place themselves in the final location.

Importance of specific impulse. From Eqs. (1) and (2) the significant effect of the specific impulse I_{sp} on the total propellant load that a spacecraft must carry to perform its assigned mission can be seen. Since a massive satellite must be boosted into space by the use of expensive launch vehicles, such as the space shuttle and Ariane, significant cost savings may be gained if smaller, less expensive launch vehicles may be used. The size of the required launch vehicle is directly proportional to the mass of the payload. Since most of the other components that make up spacecraft are relatively fixed in weight, it is critical to utilize propellant combinations that maximize specific impulse.

Propellant combinations. For modern spacecraft the choices are either bipropellants, which utilize a liquid oxidizer and a separate liquid fuel; solid propellants, which consist of oxidizer and fuel mixed together; or monopropellants, which are liquid fuels that are easily dissociated by a catalyst into hot, gaseous reaction products. High specific impulse is offered by bipropellants, followed by solid propellants and monopropellants. For spacecraft the specific impulse values range from 315 s down to 220 s. The liquid bipropellants are nitrogen tetroxide (N_2O_4) as the oxidizer and either hydrazine (N_2H_4) or monomethylhydrazine (CH₃N₂H₃) as the fuel. Solid rocket motors use hydroxy-terminated polybutadiene (HTPB) as a binding agent for the ammonium percholorate (NH₄ClO₄) oxidizer and aluminum fuel. Monopropellants are presently limited to hydrazine, which is catalyzed on a bed of iridium-coated pellets. See PROPELLANT.

Propulsion system designs. Spacecraft attitude control schemes play an important role in defining the

Typical propellant budget for a geosynchronous communications satellite					
Function	ΔV , ft/s (m/s)	Fuel, lbm (kg)	Oxidizer, Ibm (kg)		
Transfer orbit burns	_	4 (2)	_		
Apogee injection	4927 (1502)	1200 (544)	1320 (599)		
Preoperational		20 (9)			
Station acquisition	43 (13)	25 (11)	_		
Station keeping	2240 (683)	900 (408)	_		
Attitude control	<u> </u>	20 (9)	_		
Total	7210 (2198)	2169 (984)	1320 (599)		

detailed characteristics of spacecraft propulsion systems. Essentially, there are three methods for stabilizing a spacecraft: three-axis control, spin control, and gravity gradient. In three-axis systems the body axes are inertially stabilized with reference to the Sun and stars, and utilize rocket engines for control in all six degrees of freedom. Spin-stabilized spacecraft use the inertial properties of a gyroscope to permanently align one of the axes by rotating a major portion of the spacecraft body about this axis. This approach significantly reduces the number of thrusters needed for control. Gravity gradient control is a nonactive technique that relies on the Earth's tidal forces to permanently point a preferred body axis toward the Earth's center. See GYROSCOPE; INERTIAL GUIDANCE SYSTEM; SPACECRAFT STRUCTURE.

Translation of a spacecraft, independent of its control technique, requires thrusters aligned parallel to the desired translational axis. Usually, all three axes require translational capability. Combining the two requirements for attitude control and translation results in the minimum number of rocket engines required to perform the mission. These are supplemented with additional thrusters to allow for failures without degrading the performance of the propulsion system. Thrust levels vary and are dependent on spacecraft maximum and minimum acceleration levels, which in turn are dependent on allowable burn efficiencies, power availability, and the structural capability of deployed appendages, such as solar arrays or scientific instruments. In spacecraft applications, bipropellant and monopropellant engines are significantly lower in thrust by several orders of magnitude compared to solid rocket motors. This implies that burn durations, which are inversely proportional to thrust, are much longer for liquid propulsion systems than solid systems for a given velocity increment ΔV . Solid rocket motors are almost completely limited to performing large velocity changes. Bipropellant rocket engine thrust levels are between 100 lbf (445 newtons) for large- ΔV burns and 2 lbf (10 N) for attitude and translational maneuvers. Monopropellant thrusters range from 100 lbf (445 N) down to 0.05 lbf (0.2 N). Figure 1 shows the schematic thruster layout for a typical three-axis controlled bipropellant propulsion system for a communications system. The propellant budget shown in the table is based on the same spacecraft mission.

Minimization of propulsion system mass. As discussed above, a critical parameter in the ultimate design of a spacecraft propulsion system is the mass of the propellant. This is dictated by the engine specific impulse and the spacecraft launch weight. Simplistically then, it would be reasonable to assume that the propellant-engine combination with the highest specific impulse would be the preferable choice. However, the ultimate requirement is the lowest possible mass for the entire propulsion system. The complexity of the system is greatly influenced by, and is roughly proportional to, the specific impulse, since bipropellants require more tanks, valves, and so forth than either solid systems or monopropellant systems. This is primarily due to the differences in density



Fig. 1. Typical communications satellite thruster layout. Thrust is provided to control rotation about all three axes and translation, in six degress of freedom.

between liquids and solids, and the fact that bipropellants require high-pressure gas sources to expel the fluid from the tanks and into the rocket engine chamber. Typical three-axis stabilized bipropellant and monopropellant systems are compared in Fig. 2. The definition of the mission requirements, in the form of total spacecraft launch mass and the mission total velocity increment ΔV , are translated into the hardware elements needed to control the system, and the mass of these elements is estimated. Subtracting the propellant and system hardware mass from the total launch mass provides an estimate of the spacecraft mass that is available for distribution among all the other systems that constitute a spacecraft (for example, power, attitude control, and payload). For communications satellites in the lift-off weight range of 3000 lbm (1360 kg), the trade-off between specific impulse and system mass dictates the use of a solid rocket motor for the main-orbit circularizing burn and a monopropellant propulsion system for on-orbit attitude control and translation. Spacecraft launch masses above about 5000 lbm (2268 kg) require the use of all-bipropellant systems.

Blowdown versus pressure-regulated systems. From Fig. 2 it can be seen that there are two approaches to expelling propellant from spacecraft in a zero-g environment. The simple monopropellant systems use a pressurant gas stored within the same tank as the hydrazine fuel. Operation of the rocket engine valves allows the fuel to flow by virtue of the pressure difference between the tanks and the spacecraft exterior (vacuum). As the fluid is expended, the tank pressure decreases exponentially until all the fluid is removed. This technique is called a blowdown system. It is lightweight but requires a significant gas volume to ensure that the pressure remains at some reasonable value, since the specific impulse decreases with feed pressure. For most monopropellant systems the ratio of initial to final tank pressure is 4:1. By using the ideal gas law, it can be shown that the



Fig. 2. Typical propulsion system designs for communications, Earth observation, and interplanetary spacecraft missions. (a) Monopropellant system. (b) Bipropellant system.

initial gas volume must be 25% of the tank volume. *See* GAS.

In systems with larger propellant mass, which must be bipropellant systems, the tank volumes required for a simple blowdown system become impractical. Therefore, bipropellant systems usually employ separate gas reservoirs, which store the gas (helium) required to expel the propellant at very high pressures; 4500 lb/in.² absolute (31 megapascals) is typical. The storage-tank pressure is reduced to the engine operating pressure of 220 lb/in.² ab-

solute (1.5 MPa) by a pressure regulator, hence the name pressure-regulated system. Although the storage-tank mass is high, this is offset by the gain in propellant-tank mass compared to that in the blowdown approach and the ability to operate the system at the optimum engine pressure to maximize specific-impulse performance.

At the propellant loads typical of current spacecraft propulsion systems (2000 lbm or 907 kg), the use of turbopumps, which are often used in launch vehicles, has not proved to be weight-competitive with the blowdown or the pressure-regulated systems. In the future, as propellant mass continues to increase as spacecraft mass increases, pumps are likely to be utilized.

Implications of zero-g. A primary feature that distinguishes spacecraft propulsion technology from that utilized on launch vehicles is the need to cope with the extremely low acceleration values that exist in orbit. (This environment is termed zero-g, although this is not really accurate, since a constant acceleration level of the order of 10^{-5} g is permanently experienced by Earth-orbiting spacecraft.) Such a low-acceleration environment primarily affects the behavior of any fluid on board (generally propellant), due to the domination of surface-tension forces. In 1-g, Earth-bound experience, fluid geometry is controlled totally by gravitational forces. Hence, water in a cup exhibits essentially a flat surface. Any difference in fluid height is removed by fluid flowing under the influence of gravity to a common surface. The relationship defining whether surface tension or gravitational forces are dominant is provided by the dimensionless Bond number B_0 , defined by Eq. (3).

$$B_0 = \frac{\rho a L^2}{\sigma} \tag{3}$$

Here, ρ is the density of the fluid, *a* is the local acceleration, *L* is a characteristic length of the system (for example, the tank diameter), and σ is the fluid surface tension. For environments where B_0 is much greater than 1, the system is said to be gravity-dominated. When B_0 is much less than 1, the system is surface-tension-dominated. In those cases where the Bond number is close to 1, the system is intermediate and neither force clearly dominates. *See* SUR-FACE TENSION.

Clearly, in a spacecraft environment gravity is no longer significant, and the fluid internal forces define the steady-state liquid surface. In this situation, fluids attempt to minimize their total energy by reducing their surface area, and therefore the surface tension, resulting in spherical propellant shapes. The fluid spheres can attach anywhere on the internal tank surface. In particular, the propellant may not be located over the outlet of the tank, and hence gas may be expelled instead of propellant, which is clearly undesirable.

To keep the propellant "puddle" from locating itself in arbitrary locations, the principle of surface tension is often used to preferentially locate the propellant over the outlet. One way is by using channels constructed of metal with a single open face covered with a fine-weave metal screen (**Fig. 3***a*). The screen, which is woven like cloth, has pores of only about 15 micrometers that are wetted by the propellant. From Eq. (3) it follows that the Bond number is very low, and the pores are surface-tensiondominated. Figure 3*b* shows schematically the liquid configuration within a pore. When the channel is surrounded by fluid, the liquid flows through the pores with little frictional loss. In this case, gas-free propellant is provided to the propulsion system since only



Fig. 3. Surface-tension propellant management device. (a) Typical device. Liquid propellant flows through the wire screens and through the channels to the circular device at the center, where it is expelled from the spacecraft. (b) Configuration of liqud and gas within one of the pores (radius = R) of the screen.

liquid is flowing out of the tank. In the more general case of the channel and the screen exposed to gas, the surface-tension forces restrict the ability of the gas to penetrate the liquid surface within the pore. The gas pressure required to ingest bubbles across the fluid barrier is termed the bubble point and is defined by Eq. (4), where ΔP is the bubble point

$$\Delta P = \frac{\sigma}{2R} \tag{4}$$

for the given configuration, σ is the surface tension of the propellant, and R is the characteristic radius (in this application it is the radius of the pore). For nitrogen tetroxide, the bubble point is about 2 lb/in.² (14 kilopascals). Hence, the flow losses in the fluid moving through the screen must equal this value before the gas can break through the fluid surface and enter the propulsion system tubing. By proper design, it is possible to make channel devices that can expel 99.9% of the fluid content of a tank before the gas breaks through. When the gas does enter the tubing, and finally the engine, the spacecraft can no longer be predictably controlled, and this is the definition of the end of the mission. Typical mission durations are 10-15 years of operation for communications satellites and 1-5 years for low-Earth missions. Interplanetary spacecraft life requirements often exceed 10 years. Keith Davies Bibliography. B. N. Agrawal, *Design of Geosynchronous Spacecraft*, 1986; P. G. Hill and C. R. Peterson, *Mechanics and Thermodynamics of Propulsion*, 2d ed., 1992; G. P. Sutton, *Rocket Propulsion Elements*, 7th ed., Wiley, 2001.

Spacecraft structure

The supporting structure for systems capable of leaving the Earth and its atmosphere, performing a useful mission in space, sometimes returning to the Earth and sometimes landing on other bodies. Among the principal technologies that enter into the design of spacecraft structures are aerodynamics, aerothermodynamics, heat transfer, structural mechanics, structural dynamics, materials technology, and systems analysis. In applying these technologies to the structural design of a spacecraft, trade studies are made to arrive at a design which fulfills system requirements at a minimum weight with acceptable reliability and which is capable of being realized in a reasonable period of time.

The structural aspects of space flight can be divided into six broad regions or phases: (1) transportation, handling, and storage; (2) testing; (3) boosting; (4) Earth-orbiting flight; (5) reentry, landing, and recovery on the Earth; (6) interplanetary flight with orbiting of or landing on other planets. Each phase has its own structural design criteria requiring detailed consideration of heat, static loads, dynamic loads, rigidity, vacuum effects, radiation, meteoroids,

acoustical loads, atmospheric pressure loads, foreign atmospheric composition, solar pressure, fabrication techniques, magnetic forces, sterilization requirements, accessibility for repair, and interrelation of one effect with the others. Heavy reliance is placed on computer-generated mathematical models and ground testing.

The basic spacecraft structural design considerations apply equally well to both crewed and crewless spacecraft. The degree of reliability of the design required is, however, much greater for crewed missions. Also, the spacecraft structures in the case of crewed missions must include life-support systems, and reentry and recovery provisions. In the case of lunar or planetary missions where landing on and leaving the foreign body are required, additional provisions for propulsion, guidance, control, spacecraft sterilization, and life-support systems must be realized, and the structure must be designed to accommodate them.

Various factors must be considered when selecting spacecraft structural materials for the three principal phases of space flight—launch, space journey, and reentry (**Table 1**).

Transportation, handling, storage. For ground transportation and storage, special load isolation systems and hermetically sealed carriers are often built to control shock and prevent contamination. The spacecraft is monitored continuously to prove that it has remained within its design envelope as it travels to the launch site by truck, airplane, or barge.

Factors	Phase I: launch	Phase II: space journey	Phase III: reentry
Technological factors	Rocket power	Satellite all-up weight	Resistance to high kinetic heating
	Thrust alignment	Impact by meteoroids,particles, atoms, ions, and electrons	Dissipation and redistribution ofheat
	Structural stiffness		Compatibility of structural materials under thermal shock
	Operation of the launcher stages	Electromagnetic radiation	
		Radiation heating and cooling	Operation of control devices
Materials factors	Elastic modulus-density ratio	Behavior of solids at low pressures	I hermal stresses
	Rocket nozzle erosion, particularly nonuniform erosion	Radiation damage, effect on the physical properties of materials	Material survival under thermal shock at temperatures of ~8000 K (14,000°F)
	Low-temperature conditions in fuel tanks	Transparencies	Erosion
	Construction of solid-propellant	Emissivity	Ablation
	rocket-motor casings		
		Radiation shielding	Thermal insulation
	Acoustic fatigue	Resistance to impact: surface damage, spalling	Heat sinks
			Protective coatings
		Behavior of bearing surfaces, friction	
		Seals	
		Behavior of electrical materials,	
		semiconductors	
		Lightweight vehicle construction	
		Composite assemblies	
Materials of special interest	Graphite Refractory oxides	Light alloys: aluminum, magnesium, titanium, beryllium	Graphite Plastics
	Carbides, nitrides, and borides Structural plastics	Polymers and elastomers	Ceramics, particularly reinforced ceramics Molybdenum
	Nickel alloys		NIODIUITI
	Some structural low-density alloys		

*After A. J. Kennedy, Materials Background to Space Technology, George Newnes Limited Tower House, London, 1964.



Fig. 1. Finite element model of the SPOT 5 spacecraft in its stowed launch configuration. (European Space Agency/ Matra Morconi)

Testing. To ensure that spacecraft structures will meet mission requirements criteria in general requires testing to levels above the expected environmental conditions by a specific value. The test level must be set to provide for variations in materials, manufacture, and anticipated loads. In cases where structures are required to perform dynamic functions repeatedly, life testing is required to ensure proper operation over a given number of cycles of operation. *See* INSPECTION AND TESTING.

A major function of the test program is to validate the structural mathematical models. A typical model (**Fig. 1**) is constructed with a powerful finite element computer program such as NASTRAN or PATRAN. The model is used to predict dynamic loads, stress levels, natural frequencies in both the stowed and deployed configurations, on-orbit jitter and stability, temperature extremes and deformations, and operation loads for mechanisms. *See* COMPUTER-AIDED ENGINEERING; ENGINEERING DESIGN; FINITE ELEMENT METHOD.

Boost. The purpose of the boost phase is to lift the vehicle above the sensible atmosphere, to accelerate the vehicle to the velocity required, and to place the spacecraft at a point in space, heading in the direction required for the accomplishment of its mission. For space missions, the required velocities range from 26,000 ft/s (8 km/s) for nearly circular Earth orbits to 36,000 ft/s (11 km/s) for interplanetary missions. Achievement of these velocities requires boosters many times the size of the spacecraft itself. Generally, this boosting is accomplished by a chemically powered rocket propulsion system using liquid or solid propellants. Multiple stages are required to reach the velocities for space missions. Vertical takeoff requires a thrust or propulsive force that exceeds the weight of the complete flight system by approximately 30%. An example of a multiplestage booster is the Delta launch vehicle used for the crewless missions (**Fig. 2**). The Delta II 7925 vehicle has the capability to place 4000 lb (1800 kg) into a geosynchronous transfer orbit. *See* IN-TERPLANETARY PROPULSION; PROPELLANT; ROCKET PROPULSION.

The boost trajectory (**Fig. 3**) can be controlled to some extent by proper selection of the thrust-toweight ratio and launch angle. The higher thrust-toweight ratios are more efficient, but they result in a higher dynamic pressure which in turn results in increased spacecraft loading and heating.

Boost environments. The basis for design of the primary spacecraft structure is the ability to withstand the boost environment. Typical launch vehicles produce longitudinal acceleration loadings in excess of five times the Earth's gravitational field, or 5g. Lateral load factors are typically less than half of the longitudinal load factor, but cause the greatest portion of stress induced by the spacecraft base bending moment. A coupled-loads analysis, which is a dynamic simulation of the critical transient loading events, is performed in order to predict actual component



Fig. 2. Exploded view of typical Delta II 7925 three-stage structure. (After Commercial Delta II Payload Planners Guide, McDonnell Douglas Commercial Delta Inc., 1990)



Fig. 3. Typical flight profile trajectory of the Ariane 4. (Ariane 4 User's Manual)

acceleration, force, stress, or displacement response for each unique launch configuration. Sophisticated finite element modeling techniques are used to represent the launch vehicle and spacecraft in such analyses.

Vibroacoustic loads. Spacecraft structures are exposed to an acoustic environment throughout the boost phase of flight until the vehicle is out of the sensible atmosphere. The highest acoustical sound-pressure levels occur immediately after liftoff because of impingement on the payload fairing of sound originating from engine exhaust energy which is reflected off the launch pad. Another significant acoustic event occurs during transonic flight and is caused by aerodynamic shock waves and a highly turbulent boundary layer. The sound-pressure level within the fairing creates random vibration of the spacecraft components. This acoustically generated random vibration environment is the most significant loading for small components which are typically mounted on panel-type secondary structures sensitive to high-frequency excitation. See BOUND-

ARY LAYER FLOW; SHOCK WAVE; SOUND; TRANSONIC FLIGHT; VIBRATION.

Space phase and design considerations. The space phase begins after the boost phase and continues until reentry. In this phase, the structures that were stowed for launch are deployed. Important design considerations include control system interaction, thermally induced stress, and minimization of jitter and creaks.

Deployment of loads. In order to fit within the tight confines of the boost vehicle, the spacecraft must be tightly folded and stowed. When in Earth orbit, it must extend and deploy into its operational shape (**Fig. 4**). This operation places high reliability requirements on mechanisms, motors, and joints. In space, the temperature changes can be extreme. The deployment may be complicated and involve pyrotechnic devices, coiled springs, redundantly driven motors, gearboxes, latches, and cables.

Extensive testing of the deployment is required because there are probably more failures of satellites from this phase than from any other. Of course, some space structures will not support themselves in Earth's gravitational field, so accurate deployment simulation can be challenging.

If an astronaut is involved, then the deployment operation must meet restricted force and access requirements. To assist the task, the astronaut is provided with a special tool that produces torque at several preset levels.

Control system. The spacecraft control system imparts inertial loads throughout the structure. In the zero gravity environment, every change in loading or orientation must be reacted through the structure.

The structural stiffness of the spacecraft must be considered, especially as it affects the control system. The control system sensor location and the higherorder modes must be considered so that control system coupling, which may result in overall spacecraft instability, is avoided. The liquid propellant rocket systems, using fairly low pressurization, result in thinshell spacecraft structures which are not as rigid as solid propellant cases, which must withstand high internal pressures.

For orbiting telescopes, line-of-sight jitter and control is just as important as stability. As larger space structures are developed, they are increasingly difficult to control. This is due to two main factors: (1) their large size creates very low natural frequencies, which are slow to respond and are usually lightly damped, and (2) the structures often change configuration when telescopes point to new targets or solar arrays track the Sun. *See* CONTROL SYSTEMS.

Pressure vessels. Spacecraft structural design usually requires that part of the principal structure be a pressure vessel. Efficient pressure vessel design is therefore imperative. An important material property, especially in pressure vessel design, is notch sensitivity. Notch sensitivity refers to the material's brittleness under biaxial strain. This apparent brittleness contributed to premature failure of some early boosters. *See* PRESSURE VESSEL.



Fig. 4. INTELSAT VII in its stowed and deployed configurations. (Loral, LM Astro)

In pressure vessel design the variation of ultimate tensile strength with temperature is important. The exposure to elevated temperature is short.

Tank structures for spacecraft differ with the type of propellant used. With solid propellants the tank is primarily a pressure vessel; it must be able to contain the high pressures (500-1000 lb/in.² or 3.4-6.9 megapascals) generated by combustion. Other loads are usually of secondary importance, and the bursting strength of the tank is usually the most significant factor. Care must be taken in design to avoid stress concentrations, particularly near joints, which might cause premature failure.

Tanks for liquid propellants, on the other hand, need to contain relatively low pressures (10-100 lb/in.² or 69-690 kilopascals). However, flight loads in the form of thrust and bending moment become important in the design of the tank. One approach is to use the tank pressure to assist in carrying these loads, and to stabilize the light-gage unstiffened tank shell. The skin gage is chosen to be just adequate for containing the tank pressure stresses; the pressure is chosen on the basis of providing adequate stability under flight loads. Such pressures are generally in excess of those required to stabilize the tank to the point where the buckling strength is developed. Alternative approaches are to provide stiffening in the form of stringers or frames to stabilize the tank, and lowering tank pressures only to that level required by the propulsion system.

Interstages between tanks are generally designed for flight loads; their construction is similar to that of airplane fuselages. They may be monocoque unstiffened shells or may have longitudinal (stringer) or circumferential (frame) stiffening.

Gravity gradient loads. Gravity gradient loads are particularly significant in long flexible structures located within a significant gravity field such as the Earth's. Properly considered, these loads can work for the spacecraft designer by providing a method of passive spacecraft control. An excellent example of the use of this technique was the radio astronomy explorer spacecraft (*RAE-Explorer 38*) which gravity-stabilized by use of its 750-ft (229-m) antennas (**Fig. 5**).

Solar radiation pressure. Solar radiation pressure can cause long flexible structural members, such as the *RAE* antennas and booms, to deflect quite significantly. Solar pressure can also cause significant disturbing torques on spacecraft that have large surface areas. This effect cannot be ignored in spacecraft structural design or in control systems design.

Space material and debris. Meteoroids are natural bodies remaining from the formation of the solar system or from collisions of other natural bodies. These particles may have extremely high velocities relative to the spacecraft (up to 225,000 ft/s or 68 km/s).



Fig. 5. Radio Astronomy Explorer spacecraft with antennas extended in comparison to the Empire State Building in New York City.

Orbital debris comprises residual particles resulting from human space-flight activities. Collisions involving these bodies and a space station and other long-duration orbiting spacecraft are inevitable. The worst-case effects of such collisions include the degradation of performance and the penetration of pressure vessels, including high-pressure storage tanks and habitable crew modules. An essential parameter in the design of these structures is the mitigation of these effects. *See* METEOR.

The probability of an object 0.4 in. (1 cm) or larger impacting any part of a structure with a 53,820-ft² (5000-m²) exposed surface area (roughly one football field) and a 250-mi (400-km) operating altitude is predicted to be one impact in 71 years. However, an impact would not necessarily cause a failure in a space station since the core components of the largest space station designed cover only about 21,500 ft² (2000 m²) and the design life is only 15 years. Further, space-station design features shielding to protect against objects as large as 0.55 in. (1.4 cm). This represents protection from 99.8% of the objects that would otherwise penetrate an unshielded component (roughly 0.04 in. or 0.1 cm). These factors make a catastrophic penetration extremely unlikely.

The great majority of the debris is far smaller than the millimeter scale. Impacts from these objects, most of which are the size of a grain of sand, too small to cause significant structural damage, are the most frequent. These impacts cause the degradation of sensitive surfaces such as optical components and solar panels. Mitigation of these effects can be dealt with by routine maintenance operations.

Another important consideration is protection for crew members during extravehicular activities (EVA). Presently the risk appears very small because of the relatively small exposed area of the space suit and short duration of the exposure. However, increases in the debris environment could pose operational and design constraints on extravehicular activity.

Energetic particles. Radiation shielding is required for some vehicles, particularly those operating for extended times within the Earth's magnetically trapped radiation belts or during times of high sunspot activity. The shielding may be an integral part of the structure. Computer memories are particularly susceptible to radiation and cosmic-ray activity and must be shielded to survive. This is a minor concern for scientific spacecraft but a major concern for defense satellites. Effects of radiation on most metallic structures over periods of 10-20 years is not severe. The durability of composite structures in space is a major concern for long life. Based on available data, the synergistic effects of vacuum, heat, ultraviolet, and proton and electron radiation degrade the mechanical, physical, and optical properties of polymers. See MAGNE-TOSPHERE; RADIATION HARDENING; VAN ALLEN RADI-ATION

Thermal extremes. Temperatures in space can vary from 212° F (100°C) on sunlit surfaces to -292° F (-180° C) on dark surfaces. As a result, temperature

is a major design consideration for all moving parts, such as motors, hinges, and valves. When a satellite enters or exits the Earth's shadow, the rapid temperature change can induce a phenomenon called thermal jitter, where parts of the structure deform so fast that they oscillate about their neutral positions and shake the entire spacecraft. This happened in the early days of the *Hubble Space Telescope*. It was corrected with updated control software and new solar arrays. *See* HUBBLE SPACE TELESCOPE.

Temperature extremes in the structure and the enclosed environment are controlled by several techniques. Passive control is accomplished by surface coatings and multilayer thermal blankets which control the radiation transfer from the spacecraft to space and vice versa. Because incident solar radiation varies inversely with the square of distance from the Sun, means of adjusting surface conditions are required for interplanetary missions. Heat generated by internal equipment or other sources must be considered in the heat balance design. Other techniques used to actively control spacecraft temperatures are thermal louvers and heat pipes. *See* HEAT PIPE.

Thermal gradients must be considered in spacecraft design, especially when the spacecraft has one surface facing the Sun continuously. In some cases it is desirable to slowly rotate the spacecraft to eliminate such gradients.

Inertial loads. An important consideration in spacecraft structural design is the distribution of weight which determines the balance and inertia ratios of the spacecraft. In general the structural weight of spacecraft ranges 15-30% of the total spacecraft weight. By proper design, additional weight needed to balance the spacecraft and obtain the proper inertia ratios can be minimized. For spin-stabilized spacecraft the spin moment of inertia should be greater than the principal lateral moment of inertia by approximately 5% in order to ensure stability about the spin axis. Consideration must be given to expendables such as jettisoned hardware and gas depletion. As spacecraft become larger and the equipment contained therein becomes more massive, inertia loading is increasingly important.

Sloshing. Sloshing in the case of liquid propulsion systems must be carefully evaluated because it affects the structural integrity of the thin shell and the overall stability of the spacecraft. The fuel mass movements result in shifts of the center of gravity, causing pitch or yaw of the system. Control system coupling with either pitch or lateral sloshing can occur unless proper baffling or damping is provided.

Material considerations. The materials of general interest used in spacecraft structures are alloys of aluminum, magnesium, and titanium; steel; stainless steel; superalloys; plastics; and graphite-epoxy, Kevlar, and fiberglass composites. The fabrication and design techniques that are employed to obtain the most efficient structure for a given job often include honeycomb sandwich construction (**Fig. 6**), helical wound fiberglass and metal, expanded plastic sandwich construction, and woven fabrics (**Fig. 7**).



Fig. 6. Honeycomb sandwich construction provides strong and stiff but lightweight structures. (*Hexcel Corp.*)

Spacecraft structures usually are required to be lightweight and rigid, which results in the selection of high-modulus materials.

Titanium and beryllium have low densities and relatively high modulus-density ratios. Alloys of these metals are relatively difficult to fabricate, and



Fig. 7. Woven Kevlar fabric is light and strong.

therefore their application is quite limited. The more common aluminum, magnesium, and stainless steel alloys are basic spacecraft structural materials. They are easy to fabricate, relatively inexpensive, and in general quite suitable for use in the space environment.

Plastics are used in spacecraft structures when radio-frequency or magnetic isolation is required. They are also used in situations where some structural damping is desired. The thermal blankets, mentioned earlier as passive temperature control methods, are primarily Mylar plastic.

Composite materials. The modern requirements for low weight, high strength, high stiffness, and low thermal expansion (for precision optical pointing) have prompted the use of composite materials for spacecraft structure. These materials consist of highstrength reinforcement fibers which are supported by a binder material referred to as the matrix (Fig. 7). The fibers are typically made of glass, graphite, or carbon, and the matrix is an epoxy resin. The material is usually produced with the fibers configured in a unidirectional ply which has been pre-impregnated with the matrix. Such plies exhibit tremendous strength in the fiber direction but are relatively weak in the transverse direction. As a result, the plies are bonded together to form a laminate in which the fiber angle in each unique ply can be specified to provide the desired combined properties in each direction. The combined laminate is cured in an autoclave under stringent temperature and pressure conditions to ensure constant properties. With knowledge of the loading due to a unique application, the material properties can be tailored by optimizing the ply angles within the laminate to produce a material which may be substantially less dense than an equivalentstrength metallic structure.

An example of a spacecraft structure relying on composite materials is the Earth Observation System (EOS-AM1). The graphite-epoxy tubes composing the primary structure have been optimized for the substantially compressive loading induced by the Atlas launch vehicle. The tubes are made from a 70-ply rolled laminate consisting of 80% plies in the axial direction for high tensile and compressive strength and 20% plies in the 45° orientation for high torsional strength. Further, the EOS-AM1 instruments require precision pointing accuracies in the range of 0.1° . Under the extreme temperature gradients imposed on the spacecraft because of its polar orbit, the thermally induced expansion would preclude the use of typical metals for the primary structure. The thermal expansion coefficient for the graphite-epoxy tubes is 1.8 microstrain/°F (3.2 microstrain/°C) in comparison to 12.0 microstrain/°F (21.6 microstrain/°C) for typical aluminum alloys. Graphite-epoxy laminate is also used in the Hubble Space Telescope, where the laminates are arranged to minimize the coefficient of thermal expansion, preventing deformation when temperature changes occur (Fig. 8). In the Hubble, only 1 milliarcsecond $(3 \times 10^{-7}$ degree) of line-of-sight angular deformation is allowed for thermal expansion. For spacecraft requiring high precision in pointing angle, graphiteepoxy composites are the materials of choice. *See* COMPOSITE MATERIAL.

Reentry phase. Although the atmospheric layer of the Earth is relatively thin, it is responsible for the reduction of vehicle velocity and the resulting deceleration loads, as well as for the severe heating experienced by reentering vehicles. A body entering the Earth's atmosphere possesses a large amount of energy. This energy must be dissipated in a manner which allows the reentering vehicle to survive. Most of the vehicle's original energy can be transformed into thermal energy in the air surrounding the vehicle, and only part of the original energy is retained in the vehicle as heat. The fraction that appears as heat in the vehicle depends upon the characteristics of the flow around the vehicle. In turn, the flow around the vehicle is a function of its geometry, attitude, velocity, and altitude. Several different approaches are used for reentry vehicles with different trajectories (Fig. 9). See ATMOSPHERIC ENTRY.

Atmospheric drag. Spacecraft are seldom designed to reenter the Earth's atmosphere (the space shuttle being an exception), but may be designed to enter extraterrestrial atmospheres (**Fig. 10**). In either case, the structural design is similar.

High-speed reentry causes extreme friction and heat buildup on spacecraft that must be dissipated by using high-temperature ceramic or ablative materials. The space shuttle has reentered the Earth's atmosphere over a hundred times without accident (**Fig. 11**). It is covered with special thermal insulating tiles that allow the structural elements to remain cool when the surface reaches $1200^{\circ}F$ (650°C), and its leading edges are protected by a carbon-carbon reinforced material that can withstand temperatures as high as $2300^{\circ}F$ ($1260^{\circ}C$).

Satellites whose orbits decay into the Earth's upper atmosphere become flaming objects as they rapidly descend. Generally, most or all of the satellite is consumed before it reaches the surface, but there are exceptions such as the March 22, 2001, reentry of the Russian space station, *Mir*.

Ablation. The normal application of ablation is to limit the temperature of the structural shell or internal components of the payload, in order to withstand the extreme thermal environment during ballistic zero-lift reentry. This temperature limitation is the principal criterion. The most desirable material, therefore, is one which has a high heat of ablation and low thermal conductivity. Material properties, however, do not necessarily possess these characteristics simultaneously, to the same degree (Table 2). Therefore, specific investigations or trade studies must be made to determine the minimum weight of ablation material required to maintain the structural temperature within design limits. Care must be exercised to match the ablation material to the environment. For example, if the maximum temperature that would be reached without ablation is below 3000° F (1650°C), quartz would not be considered because it does not ablate appreciably below 3000°F. Many materials will rapidly deteriorate without ablating if



Fig. 8. Internal telescope of the *Hubble Space Telescope*, whose structural members are made of graphite-epoxy. (NASA/Lockheed)

exposed to temperatures below their ablation temperatures, and thus will not provide the needed protection for the structure.

Cooling and transpiration. Cooling the vehicle by circulation of water, lithium, or other media, or by transpiration of gas or vapor is limited by system complexity, reliability, and adverse weight penalties.

Thermal protective systems are time-limited because the heat-absorption capacity of the protection system is continuously expended during the process.



Fig. 9. Comparison of various space reentry and ballistic missile trajectories.

These systems, therefore, apply primarily to shortduration (5-10 min) flights.

Lifting reentry. In crewed applications, vehicles employing aerodynamic lift during reentry have several advantages over zero-lift ballistic bodies: (1) The use of lift allows a more gradual descent, thus reducing the deceleration forces on both vehicle and occupants. (2) The vehicle's ability to glide and maneuver within the atmosphere gives it greater accuracy in either hitting a target or landing at a predetermined spot. (3) It can accommodate greater errors of guidance systems because for a given deceleration it can tolerate a greater range of entry angles. (4) Greater temperature control is afforded because aerodynamic lift may be varied to control altitude with velocity.

The total time required for a lifting vehicle to descend through the atmosphere depends on the drag characteristics of the vehicle. With high drag coefficients the time may be as low as 30 min, and with low drag it may be approximately 1–2 h. In the case of grazing reentries, several orbital loops may be required to bleed off the energy, and reentry duration can be measured in days. The relatively long times required for descent make time-dependent ablation or cooling impractical.



Fig. 10. Mars Pathfinder entry, descent, and landing. (NASA/Jet Propulsion Laboratory)

Thermal stress during reentry. One of the most severe structural problems is thermal stress, which arises as a result of temperature gradients. These temperature gradients result from causes such as heat sinks, transient heating, and crewed-compartment cooling. Thermal stress is a function of the coefficient of expansion, proportional to the modulus of elasticity and the temperature gradient.

Thermal stresses will be produced in sandwich or conventional sheet-stringer construction if the thermal deformations are externally constrained. Thermal stresses will also be produced in the absence of external constraints, solely because of the incompatible deformations of the different parts of the body, for example, a rectangular anisotropic sandwich panel with a temperature gradient through its thickness.

The elastic thermal stresses are computed for the typical temperature gradients associated with the maximum temperatures. If the vehicle is intended for repeated use, the compressive yield point cannot be exceeded. As an example, sandwich construction of high-nickel alloy M-252 cannot be used above 1660°F (900°C) because the thermal stress equals the compressive yield stress at that temperature. The air-load stresses should be added to the thermal stresses, thus reducing even further the allowable temperature for material and type of construction under these conditions. The same curves are drawn for a titanium-molybdenum alloy. Its more favorable lower coefficient of thermal expansion and its higher allowable stress in the higher-temperature region result in a better sandwich. However, there are many practical problems to be solved in the brazing of molybdenum alloys.

If the vehicle were designed for a single shot, the thermal stresses could exceed the yield point considerably without seriously jeopardizing the mission. However, tremendous benefits accrue if thermal stresses are eliminated.

High-temperature materials. The extreme thermal environment requires materials with higher temperature



Fig. 11. The space shuttle (shown with cargo bay doors open) can return to Earth with help from insulation tiles that keep it from burning up. (NASA/Johnson Space Center)

capability than standard aircraft steels and aluminum alloys. These materials are the superalloys, refractory alloys, and ceramics.

The superalloys which contain high percentages of either nickel or cobalt have a maximum useful temperature range, under low stresses, of 1600-2000°F (870-1090°C). These alloys are readily available in thin gages and can be formed, riveted, brazed, and fusion- or spot-welded with essentially standard procedures. The surface of these alloys is rapidly oxidized above 1300°F (700°C). However, the oxide forms an adherent coating that protects the metal from further oxidation and provides a high surface emissivity (0.8–0.9).

The refractory alloys (niobium, molybdenum, tantalum, and tungsten) have higher densities than the superalloys but retain their useful strength from 3300° F or 1820° C (niobium) to over 4000° F or 2200° C (tungsten). All refractory metals are oxidized catastrophically at high temperatures, and surface protection is required. Coatings have been developed which protect molybdenum for approximately 300 h at 3000° F (1650° C) and niobium for several hours at 2600° F (1430° C). Niobium alloys are available in thin gages and can be formed by standard procedures. The quality of thin-gage molybdenum alloys is not consistent and requires hot forming because of their low ductility at room temperature. Tungsten alloy, available in sheet, bar, and wire form, has the highest melting point of all the metallic elements. This material offers extremely high strength but at a high mass density of 19 g/cm³. Tantalum is similar to niobium except that it has a much higher density. The refractory alloys can be joined by riveting, but fusion welding requires such special procedures as inert atmosphere. Spot welding is extremely difficult.

Conventional ceramics retain fairly high strengths up to 4000° F (2200°C), and the more exotic ceramics are potentially usable at over 8000° F (4430°C). Their high melting points and relative inertness to oxidizing atmospheres make them attractive materials for leading edges and nose cones of glide vehicles. However, the lack of ductility presents serious design problems.

Cost is not necessarily the criterion for material selection. Availability, strength, stiffness, weight, producibility, and cost of fabrication must also be considered. The final material selected will be the one which results in the least expensive end item

TABLE 2. Typical ablation materials and properties					
Material	Ablation	Thermal conductivity,	Specific heat,	Density,	
	temperature,	Btu/(in.)(s)(°F)	Btu/(lb)([°] F)	Ib/in. ³	
	°F (°C)	[J/(m)(s)(°C)]	[kJ/(kg)([°] C)]	(g/cm ³)	
Quartz	4400 (2430)	$\begin{array}{c} 0.1292 \times 10^{-4} \ (0.9660) \\ 0.4033 \times 10^{-3} \ (30.15) \\ 2.896 \times 10^{-6} \ (0.2165) \\ 3.475 \times 10^{-6} \ (0.2598) \end{array}$	0.25 (1.05)	0.08391 (2.323)	
Acrylic	250* (120)		0.38 (1.59)	0.04282 (1.185)	
Poly(tetrafluoroethylene)	1000* (540)		0.25 (1.05)	0.07523 (2.082)	
Refrasil–11% phenolic resin	2000 [°] (1090)		0.23 (0.96)	0.06366 (1.762)	
*Estimated.					



Fig. 12. ATS 7. (a) Structural assemblies. (b) Side view and (c) bottom view of deployed configuration. (d) Side view and (e) section A-A of launch configuration. (f) View of deployed spacecraft. 1 in. = 25 mm; 1 ft = 0.3 m.



Fig. 13. The International Space Station is an example of a large structure of changing configuration that cannot be completely assembled on Earth. (NASA)

and best meets the design requirements. *See* HIGH-TEMPERATURE MATERIALS.

Erectable structures. These take many and varied shapes. They are sometimes relatively simple hinged booms, while on other occasions they become quite large and massive. One type of unique erectable structure is the antennas on the radio astronomy explorer (RAE) spacecraft (Fig. 5). The four 750-ft (229-m) V antennas were erected in flight, as were two 300-ft (91-m) damper tubes and two 60-ft (18-m) dipole antennas. Each of these antenna booms was stored as a 0.002-in.-thick (50-micrometer) beryllium copper alloy tape approximately 2 in. (5 cm) wide in a mechanism within the spacecraft. Upon commands from the Earth, the booms were deployed from the mechanism to form $\frac{1}{2}$ -in.-diameter (13-mm) tubes. The weight of the 1/2-in.-diameter (13-mm) tube is approximately $^{1}/_{4}$ -oz/ft (23 g/m); therefore the 3750 ft (1143 m) of elements aboard the RAE weighed only 60 lb (27 kg). The 750-ft (225-m) antennas were designed so that upon deployment the flat tape interlocked into a torsionally rigid structural member. In addition to serving the function of radio astronomy antennas, the long booms provided gravity stabilization so that one axis of the spacecraft always faced the Earth.

Rigid spacecraft structures. Many more spacecraft structures are rigid than erectable or inflatable. Some examples of rigid spacecraft structures are the *Orbiting Astronomical Observatory* (*OAO*), the *Nimbus* weather spacecraft, the *Telstar* communications spacecraft, the *Syncom* communications spacecraft, the *Orbiting Solar Observatory* (*OSO*), and the lunar and command service modules.

Typical of a rigid spacecraft structure is the *Applications Technology Satellite* (Fig. 12), which includes a flexible, erectable, parabolic reflector 30 ft (9 m) in diameter. It is used to broadcast television signals into remote areas.

Another spacecraft structure is the space shuttle (Fig. 11) or Space Transportation System (STS), which can carry 65,000 lb (30,000 kg) of cargo to and from low Earth orbit.

The *International Space Station (ISS)* is a cooperative, 16-nation effort (**Fig. 13**). It will include six laboratories and weigh a million pounds when

assembled as currently projected in 2010. *See* SPACE STATION.

The *Viking* lander is a prime example of an interplanetary spacecraft designed for orbiting and soft landing on Mars. *See* COMMUNICATIONS SATELLITE; METEOROLOGICAL SATELLITES; MILITARY SATELLITES; SATELLITE NAVIGATION SYSTEMS; SCIENTIFIC AND AP-PLICATIONS SATELLITES.

William Haile; Peter S. Weinberger Bibliography. K. G. Budinski and M. K. Budinski, Engineering Materials: Properties and Selection, 1998; V. A. Chobotov, Spacecraft Attitude Dynamics and Control (Orbit, a Foundation Series), 1991; P. Conley (ed.), Space Vehicle Mechanisms: Elements of Successful Design, 1997; D. Jenkins, Space Shuttle: The History of the National Space Transportation System: The First 100 Missions, 2001; R. Meyer, Elements of Space Technology for Aerospace Engineers, 1999; R. Neri, Manned Space Stations: Their Construction, Operation, and Potential Applications, 1990; T. Sarafin (ed.), Spacecraft Structures and Mechanisms, from Concept to Launch, 1995.

Spallation reaction

A nuclear reaction that can take place when two nuclei collide at very high energy (typically 500 MeV per nucleon and up), in which the involved nuclei are either disintegrated into their constituents (protons and neutrons), light nuclei, and elementary particles, or a large number of nucleons are expelled from the colliding system resulting in a nucleus with a smaller atomic number. This mechanism is clearly different from fusion reactions induced by heavy or light ions with modest kinetic energy (typically 5 MeV per nucleon) where, after formation of a compound nucleus, only a few nucleons are evaporated. A spallation reaction can be compared to a glass that shatters in many pieces when it falls on the ground. The way that the kinetic energy is distributed over the different particles involved in a spallation reaction and the process whereby this results in residues and fluxes of outgoing particles are not well understood. See NUCLEAR FUSION.

Reactions in space. Spallation reactions take place in interstellar space when energetic cosmic rays (such as high-energy protons) collide with interstellar gas, which contains atoms such as carbon, nitrogen, and oxygen. This leads to the synthesis of light isotopes, such as ⁶Li, ⁹Be, ¹⁰Be, and ¹¹B, that cannot be produced abundantly in nucleosynthesis scenarios in the big bang or stellar interiors. *See* NU-CLEOSYNTHESIS.

Laboratory reactions. In terrestrial laboratories spallation reactions are initiated by bombarding targets with accelerated light- or heavy-ion beams, and they are used extensively in basic and applied research, such as the study of the equation of state of nuclear matter, production of energetic neutron beams, and radioactive isotope research.

Heavy-ion reactions. When a spallation reaction occurs as the result of a heavy nucleus colliding at high impact energy with another heavy nucleus (for example, two lead nuclei with 82 protons and 126 neutrons), the kinetic energy of the bombarding nucleus is transformed into compression, thus increasing the density of the nuclear matter, and into thermal motion of the nucleons. New particles are thereby created, some of which escape the nuclear medium. By studying these escaping particles, information can be extracted concerning the equation of state of nuclear matter. As in ordinary gases, liquids, and solids, this equation of state gives information about properties of nuclear matter, such as the compressibility, as a function of density and temperature. The equation of state of nuclear matter is relevant to better understanding the creation of neutron stars and black holes. At extremely high bombarding energy, so high a temperature might be reached that the nucleons (protons and neutrons) would disintegrate into their constituents, known as quarks, and, together with the gluons that carry the strong interaction between the quarks-like small strings-form the so-called quark-gluon plasma. At present it is assumed that such a state was present at the very beginning of the universe, immediately after the big bang. See BIG BANG THEORY; QUARK-GLUON PLASMA; RELATIVISTIC HEAVY-ION COLLISIONS.

Light-ion-induced reactions. When a light nucleus (such as a proton) hits a heavy nucleus (such as lead) with an energy of a few hundred megaelectronvolts or more, spallation can also take place. The resulting neutrons that escape the reaction are intensively used in condensed-matter, chemistry, geology and biology research at different spallation neutron source facilities. *See* NEUTRON; NEUTRON DIFFRAC-TION; RADIATION DAMAGE TO MATERIALS; SLOW NEU-TRON SPECTROSCOPY.

The light-ion-induced spallation reaction is less violent, and there is a nonnegligible probability that only a number of nucleons (10-40) will be removed from the target nucleus, resulting in a radioactive isotope. Such spallation reactions are used intensively to produce new isotopes. On the chart of nuclides, a representation of the different isotopes as a function of number of neutrons and protons, these radioactive isotopes can lie far away from the line of stable isotopes and have very short half-lives, on the order of a few milliseconds. Their study yields information about the manifestation of the electromagnetic, strong, and weak interactions in the nuclear medium, and their properties determine to a certain extent the nucleosynthesis processes taking place in explosive stellar environments. However, spallation of heavy atoms can lead to the production of more then 100 different nuclei. Therefore the desired isotopes must be separated from the debris. Electromagnetic mass spectrometers combined, for example, with laser ion sources are used, and numerous new isotopes have been discovered and studied. Several projects around the world have been set up to accelerate radioactive isotopes that are produced in proton-induced spallation reactions (among

others), in order to carry out studies of nuclear reactions that, in turn, involve these reaction products. *See* MASS SPECTROSCOPE; NUCLEAR REACTION; RADIOACTIVE BEAMS; RESONANCE IONIZATION SPEC-TROSCOPY. Piet Van Duppen

Bibliography. H. L. Ravn and B. W. Allardyce, On-line mass separators, in D. A. Bromley (ed.), *Treatise on Heavy Ion Science*, vol. 8, p. 363, Plenum Press, New York, 1989; C. E. Rolfs and W. S. Rodney, *Cauldrons in the Cosmos*, pp. 508–514, University of Chicago Press, 1988; R. Silberberg and C. H. Tsao, Spallation process and nuclear interaction products of cosmic rays, *Phys. Rep.*, 191:351–408, 1990.

Sparganosis

An infection by the plerocercoid, or sparganum, of certain species of the genus Spirometra. Members of this genus have a life cycle much like that of Diphyllobothrium latum. The adult normally occurs in the intestine of dogs and cats, the procercoid in copepods, and the plerocercoid in the musculature of frogs, snakes, or aquatic mammals, but not in fish. The whitish, unsegmented plerocercoid usually is an inch or so in length, but may reach a length of a foot or more. Humans may become infected accidentally by drinking water with infected copepods or by eating hosts containing plerocercoids. Since humans are not suitable hosts for the adult stage, the plerocercoids leave the gut and enter the musculature, body cavities, or other sites and remain as plerocercoids. In Eastern countries, freshly opened frogs are sometimes used as poultices on sores, particularly around the eyes, and if plerocercoids are present, they may invade the human tissues. Human sparganosis is rare in North America. See PSEUDO-Reino S. Freeman PHYLLIDEA.

Spark gap

The region between two electrodes in which a disruptive electrical spark may take place. The gap should be taken to mean the electrodes as well as the intervening space. Such devices may have many uses. The ignition system in a gasoline engine furnishes a very important example. Another important case is the use of a spark gap as a protective device in electrical equipment. Here, surges in potential may be made to break down such a gap so that expensive equipment will not be damaged. The dependence of spark-gap operation on such factors as pressure, length, electrode characteristics, and time dependence of potential is quite complicated. *See* BREAK-DOWN POTENTIAL; ELECTRIC SPARK. Glenn H. Miller

Spark knock

The sound and related effects produced in sparkignited internal combustion engines by instantaneous ignition and combustion (autoignition) of the gaseous fuel-air mixture ahead of the advancing flame front. Spark knock is a complex physicochemical phenomenon.

After spark ignition, a flame travels outward from the spark plug and, under normal combustion, will progressively burn the entire fuel-air charge. The burned gas liberates heat and expands, leading to increased pressure and temperature in the unburned gas ahead of the flame front. In addition, cylinder pressure and temperature are increasing due to the upward, compressive motion of the piston. The result is that the unburned gas may be raised above its autoignition temperature. However, before autoignition occurs, time must be allowed for the chemical reactions that precede knock. If the flame front passes through the unburned gas before these reactions are completed, normal combustion results. If these reactions occur too quickly or if the flame front velocity is too small, the unburned gas spontaneously ignites and burns instantaneously (see illus.). While this process is often called detonation, it is not identical to the detonation of explosives. The instantaneous combustion results in a very intense pressure wave that is internally reflected within the cylinder at a characteristic frequency (5-9 kHz) related to the velocity of sound within the cylinder and the dimensions of the combustion chamber. The flexing of the cylinder wall and cylinder head produces the audible, high-frequency (pinging) sound known as spark knock. See COMBUSTION CHAMBER; EXPLO-SIVE.

Besides sound, spark knock can result in pitting, or erosion, of the combustion chamber, damage to spark plug electrodes, and possible structural damage to the engine. Spark knock also leads to loss of engine efficiency by inducing spark plug preignition, resulting in overly advanced spark timing. Spark knock also causes intense turbulence within the cylinder, aggravating heat loss from the burned gas to the colder cylinder and head surfaces and reducing efficiency.

There are three general factors affecting the existence of spark knock: the effect of engine design, the effect of fuel, and the effect of engine operating variables.

Effect of engine design. Engine design variables influencing spark knock are compression ratio, combustion chamber design, spark plug location, and incylinder gas velocity. Designs that increase the temperature, pressure, and chemical residence time of the unburned gas (end gas) increase spark knock. Increased compression ratio, off-center plug location, and slow-burn combustion chambers all lead to increased spark knock. A faster-burning chamber, due to higher in-cylinder gas velocity, and central plug location increase knock resistance. Characteristics of faster-burning chambers include rotational motion (swirl) of the charge (due to off-cylinder axis charge admission); in-cylinder jetting action at the top of the compression stroke caused by nonuniform squeezing of the charge (due to proximity of portions of the piston and cylinder head); and use of two or more spark plugs. A spark plug should be



Sequence of spark knock. (a) Initial flame growth around the end of the spark plug. (b) Beginning of combustion and progression of flame through unburned gas. (c) Autoignition in the unburned end gas ahead of the flame. (d) Spark knock, or detonation. (Champion Spark Plug Co.)

located closer to the exhaust valve than the intake valve, since the former is hotter. Turbocharged and supercharged engines have increased spark-knock tendency because of higher intake-manifold air pressure. Stratified charge spark-ignited engines, having inhomogeneous air-fuel ratio distribution in each cylinder, have decreased knock tendency. *See* IG-NITION SYSTEM; SPARK PLUG; SUPERCHARGER; TUR-BOCHARGER.

Effect of fuel. The fuel octane number represents its resistance to spark knock. Higher octane numbers are due to higher autoignition temperatures or longer end-gas chemical reaction times. Spark knock can be eliminated by substituting a high-octane fuel for one of lower octane. *See* OCTANE NUMBER.

Either fuel structure or fuel additives determine octane number. More compact hydrocarbon molecules have higher octane numbers than do long-chain molecules. For many years, one of the most popular and effective antiknock additives was tetraethyllead. However, tetraethyllead use for new automobile engines sold in the United States has been substantially eliminated because of lead's ability to poison platinum catalysts. Platinum-based catalysts are required for aftertreatment of HC, CO, and NO_x exhaust emissions to render them harmless, as required by federal regulations. With the introduction of catalysts, engine compression ratios were reduced to permit extensive use of lower-octane nonleaded gasoline. Increasingly, nonleaded octane improvers, such as ethanol, methanol, and methyl butyl tertiary ether, are added to refinery gasoline. However, considerably larger amounts of these than of tetraethyllead are required to provide equivalent antiknock. See AL-COHOL FUEL; GASOLINE.

Effect of operating variables. Engine operating variables, such as spark advance, revolutions per minute, throttle angle, coolant temperature, intake air temperature and humidity, and air/fuel ratio, can have a major influence on spark knock. To reduce or eliminate spark knock, spark advance can be reduced (retarded), revolutions per minute reduced, throttle angle decreased, coolant temperature decreased, intake air temperature reduced, intake air temperature reduced, intake air temperature reduced.

Many automotive engines, including most supercharged and turbocharged spark-ignition engines, have a piezoelectric knock sensor. When detonation occurs, the sensor sends a small signal voltage to the detonation-control system or the electronic engine control system. This retards the ignition timing until the detonation stops. On some supercharged or turbocharged engines, the detonation-control system reduces the boost pressure in the intake manifold. On other supercharged or turbocharged engines, detonation is controlled by first retarding ignition timing and then reducing boost pressure. *See* AUTOMOTIVE ENGINE; COMBUSTION; INTERNAL COM-BUSTION ENGINE. Joseph R. Asik; Donald Anglin

Bibliography. E. A. Avallone and T. Baumeister III (eds.), *Marks' Standard Handbook for Mechanical Engineers*, 10th ed., 1996; A. By, B. Kempinski, and J. M. Rife, *Knock in Spark Ignition Engines*, Soc. Automot. Eng. Pap. 810147, 1981; E. F. Obert, *Internal Combustion Engines and Air Pollution*, 1973; C. F. Taylor, *The Internal Combustion Engine in Theory and Practice*, 2d rev. ed, vol. 2, 1984.

Spark plug

A device that screws into the combustion chamber of an internal combustion engine to provide a pair of electrodes between which an electrical discharge is passed to ignite the combustible mixture. The spark plug consists of an outer steel shell that is electrically grounded to the engine and a ceramic insulator, sealed into the shell, through which a center electrode passes (see **illus.**). The high-voltage current jumps the gap between the center electrode and the ground electrode fixed to the outer shell.

The electrodes are made of nickel and chrome alloys that resist electrical and chemical corrosion. Some center electrodes have a copper core, while others have a platinum tip. Many spark plugs have a resistor in the center electrode to help prevent radiofrequency interference. The parts exposed to the combustion gases are designed to operate at temperatures hot enough to prevent electrically conducting deposits but cool enough to avoid ignition of the



Cross section of a typical spark plug. (Champion Spark Plug Co.)

mixture before the spark occurs. See IGNITION SYS-TEM. Donald L. Anglin

Spatangoida

An order of exocyclic Eucchinoidea in which the posterior ambulacral plates form a shield-shaped area behind the mouth. The plates are arranged in two



Echinocardium cordatum, showing diagnostic characters of spatangoids. (a) Aboral view. (b) Posterior aspect. (c) Amphisternous plastron. (d) Apical system.

similar parallel longitudinal series. The structure is termed an amphisternous plastron. Teeth are lacking; phyllodes develop, but without bourrelets. Thus there is no floscelle. The apical system is compact. The families are defined mainly by reference to the fascioles, which are ribbonlike bands of minute, close-set uniform ciliated spinules on various parts of the test (see illus.). According to their position, they are defined as anal, subanal, inner (surrounding the apical system), peripetalous (surrounding the petals), or marginal. The earliest forms were the Toxasteridae of the Lower Cretaceous, which lacked fascioles and petals. The surviving Palaeopneustidae are deep-sea forms with an oval test, long spines, and weakly developed fascioles and petals. The Spatangoida are more specialized, heart-shaped forms which live in mud or sand. The order reached its maximum in the mid-Tertiary, but is still richly represented today in most seas. See ECHINODERMATA; ECHINOIDEA; EUECHINOIDEA. Howard B. Fell

Spearmint

Either of two vegetatively propagated, clonal cultivar species of mints of the family Lamiaceae (Labiatae). They are grown primarily in Idaho, Indiana, Michigan, Washington, and Wisconsin as a source of essential oil of spearmint.

Cultivation. In agricultural production the dormant stolons, or leafless underground stems (often misnamed roots), of these perennial species are planted in a 6-in.-deep (15-cm) trench in rows 38 in. (97 cm) apart from February to April and chemically weeded with a terbacil herbicide; the leaf crop is mowed and windrowed in July or August at the time of flowering and plant maturity, partly dried, and then steam-distilled. Fields are shallow-plowed in early November, and the second-year, third-year, and subsequent crops are a solid stand, or meadow mint, in all areas that do not use rill irrigation.

Cultivars. The original cultivar "Native spearmint" with flowers in terminal inflorescences (see illus.), introduced from Europe into New England in pioneer days, is a sterile allotriploid hybrid having 36 somatic chromosomes, an example of a botanical mule. Native cultivar is usually called Mentha spicata (2n = 48) but is actually a hybrid between this species and *M. longifolia* (2n = 24). Native spearmint comprises about one-third of the production in the United States. Scotch spearmint cultivar (illus. b) with flowers in the axils of the leaves, M. cardiaca (2n = 72), some strains have 60 chromosomes), was introduced in 1910, apparently from Scotland. Mentha cardiaca has hybrid vigor and usually yields 50-65 lb (23-30 kg) of oil per acre in the Midwest as opposed to 30-40 lb (14-18 kg) per acre for Native, with yields up to 80-125 lb (36-57 kg) in the Far West. Both varieties have similar but not identical oil composition, with 60-69% carvone, 7-20% limonene, and 1-2% cineole as main compounds.



Spearmint. (a) Native spearmint (Mentha spicata). (b) Scotch spearmint (M. cardiaca). (Courtesy of A. M. Todd Co.)

Research has shown that Scotch is a very specific, very select, sterile F_1 hybrid between *M. arvensis* (2n = 96 or 72) and *M. spicata*. Both cultivars are subject to Puccinia rust and powdery mildew, and Scotch is also susceptible to Verticillium wilt. Mutation breeding is in progress to obtain resistant strains for these major diseases.

Uses. Principal uses of the oil are in flavoring gum, toothpaste, and candy. Chopped fresh leaves of *M. spicata* preserved in vinegar are used as a condiment served with lamb, especially in England, and dried or freeze-dried leaves of several strains are used in

flavoring soups, stews, tea, or sauces. Sprigs of the decorative curly mint *M. crispa* (or *M. spicata* var. *crispata*) are often used in mixed drinks such as mint juleps.

Other mint species, as *M. niliaca*, *M. rotundifolia*, and *M. cordifolia*, may have carvone and a spearmint odor, but the presence of too much dihydrocarvone or of the alcohols and esters of carvone and dihydrocarvone (as carveol and carvyl acetate) or other hydrocarbons such as myrcene makes the commercial utilization of their oils unlikely. Although the cultivar species have been introduced throughout the world, in the United States they grow best between 40 to 48° north latitude, as in northern Europe, or farther north to 55° in the milder climate of Great Britain. *See* LAMIALES; SPICE AND FLAVORING.

Merritt J. Murray

Special functions

Functions which occur often enough to acquire a name. Some of these, such as the exponential, logarithmic, and the various trigonometric functions, are extensively taught in school and occur so frequently that routines for calculating them are built into many pocket calculators. *See* DIFFERENTIATION; LOGARITHM; TRIGONOMETRY.

The more complicated special functions, or higher transcendental functions as they are often called, have been extensively studied by many mathematicians because they arose in the problems which were being studied. Among the more useful functions are the following: the gamma function defined by Eq. (1), which generalizes the factorial; the related beta function defined by Eq. (2), which generalizes

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt \tag{1}$$

$$B(x, y) = \int_0^1 t^{x-1} \left(1 - t\right)^{y-1} dt$$
 (2)

the binomial coefficient; and elliptic integrals, which arose when mathematicians tried to determine the arc length of an ellipse, and their inverses, the elliptic functions. The hypergeometric function and its generalizations include many of the special functions which occur in mathematical physics, such as Bessel functions, Legendre functions, error functions, and the classical orthogonal polynomials of Jacobi, Laguerre, and Hermite. The zeta function defined by Eq. (3) has many applications in number theory,

$$\zeta(s) = \sum_{n=1}^{\infty} n^{-s} \tag{3}$$

and it also arises in M. Planck's work on radiation. *See* BESSEL FUNCTIONS; ELLIPTIC FUNCTION AND IN-TEGRAL; GAMMA FUNCTION; HEAT RADIATION; HYPER-GEOMETRIC FUNCTIONS; LEGENDRE FUNCTIONS; NUM-BER THEORY; ORTHOGONAL POLYNOMIALS.

Bernoulli polynomials and numbers. Bernoulli polynomials, defined by the generating function in

Eq. (4), and Bernoulli numbers, $B_n = B_n(0)$, arise

$$\frac{te^{xt}}{e^t - 1} = \sum_{n=0}^{\infty} B_n(x) \frac{t^n}{n!} \qquad |t| < 2\pi \qquad (4)$$

in many applications. (Different notations are used, especially for Bernoulli numbers.) They occur in the Euler-Maclaurin summation formula (5), where *a*, *m*,

$$\sum_{j=a}^{n} f(j) = \int_{a}^{n} f(x) \, dx + \frac{1}{2}f(a) + \frac{1}{2}f(n) + \sum_{k=1}^{m} \frac{B_{2k}}{(2k)!} \left[f^{(2k-1)}(n) - f^{(2k-1)}(a) \right] + R_{m}(n) \quad (5)$$

and *n* are arbitrary integers, a < n, m > 0, and notation (6) applies. The symbol [x] stands for the largest

$$R_m(n) = -\int_a^n \frac{B_{2m}(x-[x])}{(2n)!} f^{(2m)}(x) \, dx \quad (6)$$

integer less than or equal to *x*. This formula is very useful in finding numerical values of some slowly convergent infinite series or partial sums of divergent series. An important sum which can be evaluated using Bernoulli numbers is given by Eq. (7).

$$\sum_{k=1}^{\infty} \frac{1}{k^{2n}} = \frac{(2\pi)^n B_{2n}}{2(2n)!} \tag{7}$$

Bernoulli polynomials and numbers also play an important role in number theory, combinatorial analysis, and the study of spline functions. Splines are piecewise polynomials which are spliced together in a smooth manner, up to a certain degree of differentiability. They are very useful in fitting curves to numerical data and are the solutions to a number of important extremal problems. *See* COMBINATORIAL THEORY.

More complicated functions. All of the special functions mentioned above can be given by explicit representations, either as a series or an integral. There are other functions, among them Lamé functions, spheroidal and ellipsoidal wave functions, and Mathieu functions, which arise as solutions to important differential equations, and for which explicit formulas are lacking. At least the formulas are not nearly as explicit as for the above functions. The ordinary differential equations satisfied by these functions arise from Laplace's equation or the wave equation when it is solved by separation of variables in certain systems of curvilinear coordinates. Instead of having explicit integral representations, these functions satisfy special integral equations. See DIFFERENTIAL EQUA-TION; INTEGRAL EQUATION; LAPLACE'S DIFFERENTIAL EQUATION; WAVE EQUATION.

For example, the prolate spheroidal wave functions of order zero, $\psi_n(x)$, n = 0, 1, ..., are the bounded continuous solutions of differential equation (8) and integral equation (9), where χ_n and

$$(1-x^2)\frac{d^2\psi_n}{dx^2} - 2x\frac{d\psi_n}{dx} + (\chi_n - c^2x^2)\psi_n = 0 \quad (8)$$

$$\lambda_n \psi_n(x) = \int_{-1}^1 \frac{\sin c \, (x - y)}{\pi \, (x - y)} \psi_n(y) \, dy \qquad (9)$$

 λ_n are eigenvalues. Differential equation (8) is the usual starting place for the study of these functions. It arises when solving the wave equation in prolate spheroidal coordinates, and the integral equation has classically been derived as a property of these functions. Integral equation (9) is of interest because of its applications to stochastic processes, the study of lasers, the uncertainty principle, antenna theory, and the statistical theory of energy levels of complex systems. *See* STOCHASTIC PROCESS; UNCERTAINTY PRINCIPLE.

Historical development. Typically, special functions are discovered in the course of working on one problem. Many properties are then discovered, not only to aid in the solution of the problem which gave rise to them, but as a mathematical exercise, and then some of the new properties are used to solve completely different problems.

Each generation of mathematicians has a way of looking at mathematics which is different from previous generations. In the nineteenth century, complex analysis was being developed, and many important properties of special functions were discovered in the course of this development. In the second half of the twentieth century, Lie group and Lie algebra methods have been applied to obtain other important properties of special functions. *See* COM-PLEX NUMBERS AND COMPLEX VARIABLES; LIE GROUP. Richard Askey

Bibliography. M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*, 10th ed., 1972; L. Andrews, *Special Functions of Mathematics for Engineers*, 2d ed., 1998; R. A. Askey, T. M. Koornwinder, and W. Schempp (eds.), *Special Functions: Group Theoretical Aspects and Applications*, 1984; A. Erdélyi et al., *Higher Transcendental Functions*, 3 vols., 1953-1955, reprint 1981; G. Szegö, *Orthogonal Polynomials*, 4th ed., 1975; Z. X. Wang and D. R. Guo, *Special Functions*, 1989.

Speciation

The process by which new species of organisms evolve from preexisting species. It is part of the whole process of organic evolution. The modern period of its study began with the publication of Charles Darwin's and Alfred Russell Wallace's theory of evolution by natural selection in 1858, and Darwin's *On the Origin of Species* in 1859.

There was no problem of speciation during the period when it was believed that species were divinely created and immutable. Belief in the fixity of species was almost universal before the middle of the nineteenth century. Then it was gradually realized that all species continuously change, or evolve; however, the causative mechanism remained to be discovered.

Darwin proposed a mechanism. He argued that (1) within any species population there is always some heritable variation; the individuals differ among

themselves in structure, physiology, and behavior; and (2) natural selection acts upon this variation by eliminating the less fit. Thus if two members of an animal population differ from each other in their ability to find a mate, obtain food, escape from predators, resist the ravages of parasites and pathogens, or survive the rigors of the climate, the more successful will be more likely than the less successful to leave descendants. The more successful is said to have greater fitness, to be better adapted, or to be selectively favored. Likewise among plants: one plant individual is fitter than another if its heritable characteristics make it more successful than the other in obtaining light, water, and nutrients, in protecting itself from herbivores and disease organisms, or in surviving adverse climatic conditions. Over the course of time, as the fitter members of a population leave more descendants than the less fit, their characteristics become more common.

This is the process of natural selection, which tends to preserve the well adapted at the expense of the ill adapted in a variable population. The genetic variability that must exist if natural selection is to act is generated by genetic mutations in the broad sense, including chromosomal rearrangements together with point mutations. *See* GENETICS; MUTA-TION.

If two separate populations of a species live in separate regions, exposed to different environments, natural selection will cause each population to accumulate characters adapting it to its own environment. The two populations will thus diverge from each other and, given time, will become so different that they are no longer interfertile. At this point, speciation has occurred: two species have come into existence in the place of one. This mode of speciation, speciation by splitting, is probably the most common mode. Two other modes are hybrid speciation and phyletic speciation; many biologists do not regard the latter as true speciation.

Splitting. There are four ways in which a species population can become separated into two (or more) parts that can undergo genetic divergence and evolve into separate species (**Fig. 1**). G. L. Bush has named them types Ia, Ib, II, and III.

In type Ia an extensive population becomes split into comparatively large parts (Fig. 1*a*). In terrestrial species this may happen because movement of the Earth's crustal plates rafts the parts away from each other; or else a barrier such as a mountain range or a tract of desert may form and split a once-continuous population. A marine population is split if a land bridge forms which cuts through it; for example, the fall in sea level that accompanied the Pleistocene glaciations created a land bridge between Alaska and Siberia.

A type Ib split occurs if a small, outlying segment of a population becomes detached from the rest (Fig. 1*b*). Such a peripheral isolate is usually a colony founded by a few emigrants from the main body of the species population. The colony can evolve into a new species even if it is not totally isolated from the parent population, provided its necessarily small gene pool chances to be well adapted to the habitat in which it finds itself. The "preadapted" possessors of these genes will be selectively favored within their small local habitat, and will quickly proliferate.

Population splits of both types Ia and Ib cause the diverging populations to be geographically separated, or allopatric. The resultant genetic separation is known as allopatric speciation.

In type II a population becomes split into geographically contiguous parts between which there is a limited amount of gene exchange. The environments of the daughter populations are assumed to differ sufficiently for natural selection to override the effects of the restricted gene flow, and speciation can therefore occur. This is parapatric speciation (Fig. 1*c*).

A type III split occurs if, within a single geographic region, parts of a population in some way become so isolated from each other that gene exchange between them falls below the level needed to maintain the integrity of the original population. The result is known as sympatric speciation (Fig. 1d). It is probably common among parasites and parasitoids, but is thought to be rare or nonexistent in free-living organisms. Even though closely related species adapted to different habitats, or with different modes of life, or with different resource requirements, often co-occur in the same geographic area, their co-occurrence should not be ascribed to sympatric speciation unless a mechanism can be found that could have initiated their divergence in the first place. A more likely explanation for such co-occurrences is that the species evolved allopatrically or parapatrically, that ecological divergence accompanied speciation, and that the new species then became secondarily sympatric by migrating into each other's areas after the geographic barrier that formerly separated them disappeared. See ISLAND BIOGEOGRAPHY.

Descriptions of allopatric, parapatric, and sympatric speciation often tacitly assume that in the



Fig. 1. Different modes of speciation; a-e are described in the text. The two tints represent the geographic ranges of diverging and speciation species populations.

absence of clear, unmistakable barriers, gene exchange occurs freely throughout very extensive populations; often an entire species is treated as a single breeding population. However, P. R. Ehrlich and P. H. Raven have argued that usually the total population of a whole species is broken into a large number of small breeding populations (demes) that are fairly isolated from each other and among which gene exchange is comparatively rare. Thus a deme rather than a species population is the unit of evolution. As a result, sympatric speciation will sometimes seem to have occurred if a deme near the center of a species' range chances to diverge from the surrounding demes; the process may be called quasisympatric speciation (Fig. 1e). The fact that all demes do not diverge from one another, giving as many species as there are demes, presumably results from stabilizing selection. Stabilizing selection causes groups of demes to persist in their resemblance to each other, and such groups are the "populations" constituting the visible units of evolution. See POPULATION GE-NETICS.

Isolating mechanisms. Two genetically diverging populations are not separate species so long as it remains possible for them to lose their distinctness, by interbreeding, if they chance to become sympatric. For speciation to be complete, the populations must evolve isolating mechanisms that prevent gene exchange between them; this allows them to maintain their genetic integrity.

There are many kinds of isolating mechanisms. The diverging populations may become genetically incompatible, so that hybrids either are never produced or are sterile. Even when the diverging populations can be shown to be infertile, they may be effectively isolated by a variety of mechanisms; they may breed at different times or in different habitats; they may have evolved different patterns of mating behavior, different mating calls, or different sex attractants (pheromones), so that individuals of the two populations never attempt to copulate; copulation may be mechanically impossible because of the structure of the genitalia (this occurs in some insects); and in insect-pollinated plants, diverging populations sometimes become adapted to different species of insect pollinators.

Hybrid speciation. In plants, new species are sometimes formed abruptly by the following mechanism. The accidental crossing of two diploid plants of different, though related, species yields a healthy but sterile diploid hybrid. The hybrid is sterile because the chromosomes it has received from its parents cannot form normal bivalent pairs; meiosis is upset, and functional gametes cannot be produced. But now chromosome doubling occurs in the sterile hybrid, making it tetraploid; normal meiosis becomes possible; fertile gametes are produced, and a new species has appeared.

In symbols, let the genomic constitution of the parents be *AA* and *BB* respectively. That of their sterile hybrid is *AB*, in which none of the chromosomes has a homologous partner to pair with at meiosis. After chromosome doubling, however, the hybrid,

now tetraploid, has the constitution *AABB*; meiosis can proceed normally, and the gametes are fertile. The new species is called an amphidiploid (or allotetraploid).

A whole range of similar, but more elaborate, variants of this process occur as well, yielding hybrids with different numbers of chromosome sets. The general name for the process, with numbers unspecified, is amphiploidy (or allopolyploidy or alloploidy).

The process happens because chromosome doubling is a common occurrence in plants; doubling results either from somatic doubling at mitosis (the paired chromosomes fail to separate) or from nonreduction at meiosis. Amphiploids may be produced repeatedly, with new lineages being added in a succession of outcrossings, backcrossings, and chromosome doublings. The result is a polyploid complex, an array of interrelated diploid and polyploid species. *See* MEIOSIS.

Speciation in uniparental organisms. Many plants reproduce uniparentally for long periods of time. Uniparental reproduction occurs in two ways: by autogamy (self-fertilization) and by apomixis (asexual reproduction); the latter term includes both vegetative multiplication and agamospermy (the asexual formation of viable seeds). A uniparental plant is not a member of a "biological species" as the term is usually defined, namely, a reproductively isolated collection of interbreeding individuals. This has led to the definition, by G. G. Simpson, of the "evolutionary species," a more inclusive category, defined so as to apply to both uniparental and biparental organisms. An evolutionary species is a collection of genotypically similar individuals of common ancestry whose similarity is maintained by stabilizing selection in a given environment.

In describing speciation in uniparental organisms, one must consider how a new uniparental lineage arises in the first place, and how its members can subsequently speciate. A uniparental lineage arises when two sexual species hybridize and the hybrids, though sexually sterile, reproduce uniparentally. If many crossings occur between the two sexual species, the result is a collection of hybrid individuals that are similar, but not identical, genetically. The collection constitutes an evolutionary species. Suppose now that each individual becomes the ancestor of a succession of uniparentally produced descendants. If the progenitors of these lineages begin growth in a number of slightly different habitats, natural selection will lead to differentiation among them: the survivors in each habitat will be adapted to that habitat. Thus the plants in any one habitat are likely to differ slightly from those in other habitats. The differences are genetic, and they persist. Each distinguishable group of plants is called a microspecies. An array of microspecies, descended from the same ancestral pair of sexual species, is an agamic complex.

Microspecies that reproduce asexually (apomictically) are usually highly heterozygous; but autogamous microspecies lose the heterozygosity of their hybrid origin after several generations of



Fig. 2. Evolution of two lineages. Lineage 1 (species A) persists unchanged. Lineage 2 changes with time and has been divided, at an arbitrary boundary, into two chronospecies, B and C. (*After P. D. Gingerich, Univ. Mich. Pap. Paleontol.*, 15:1–140, 1976)

self-fertilization and become complete, or nearly complete, homozygotes.

Phyletic evolution and chronospecies. Consider a series of fossils spanning a long interval of time (long in the geological sense) and believed to represent the successive forms of a single lineage. If there has been appreciable evolution within the lineage, the change is described as phyletic evolution or anagenesis.

Fossils from a single lineage but of different ages may differ from each other so conspicuously that even though the younger is directly descended from the older, paleontologists find it convenient to treat them as taxonomically different species. Such species are called chronospecies (or successional species, paleospecies, or evolutionary species). They are not species in the usual sense; questions as to



Fig. 3. Two contrasted modes of species evolution: (a) punctuated equilibrium model and (b) phyletic gradualism model. (*After S. M. Stanley, Macroevolution, Pattern and Process, W. H. Freeman*, 1979)

whether or not they can interbreed do not arise, since they are never alive at the same time. In a lineage evolving gradually, decisions on where to put the dividing line between each chronospecies and its successor are arbitrary. A "vanished" chronospecies, one that has disappeared and been replaced by a successor chronospecies of the same lineage, is said to have undergone pseudoextinction.

Figure 2 contrasts the behavior of a lineage evolving phyletically (lineage 2) with one persisting unchanged (lineage 1). The "species" labeled B and C in lineage 2 are chronospecies. At an arbitrary point on the time scale, B becomes pseudoextinct and is succeeded by C.

The transition from a chronospecies to its successor is not usually regarded as speciation in the strict sense. Speciation as usually defined entails the splitting of lineages.

Quantum speciation versus phyletic gradualism. When a lineage splits, the two descendant lineages may diverge slowly or rapidly. There has been much debate on whether one or other of these modes has dominated the course of evolution and, if so, which. The two modes to be contrasted are shown diagrammatically in **Fig. 3**.

The tree on the right in the figure represents evolution and speciation according to the model known as phyletic gradualism. It supposes that all lineages undergo continuous gradual change (that is, phyletic evolution as described above) at a fairly constant rate. When splits occur, the descendant lineages diverge gradually from each other. There is no appreciable change in evolutionary rate at the time a split takes place. While the separated lineages are still capable of interbreeding to a limited extent, though with reduced fertility, they are known as semispecies. Their morphological divergence may sometimes be so slow that they remain practically indistinguishable even after they have lost the capacity to interbreed and have therefore reached specific rank; if so, they are known as cryptic species.

In Fig. 3a the tree represents the punctuated equilibrium model of evolution and speciation, so called by N. Eldredge and S. J. Gould. It supposes that the splitting of a parent species population is always (or nearly always) accompanied by quantum speciation, the extremely rapid (in the geological sense) and pronounced divergence of one of the descendant lineages from the ancestral form. It is assumed that the divergent species originates most often by the chance splitting off of a small peripheral isolate from an extensive parental population; this is allopatric speciation of type Ib (Fig. 1). If the isolate is small, its small gene pool may not be at all representative of the parent gene pool from which it is drawn. It may, simply by chance, deviate markedly. Further, because of its peripheral location, the environment of the isolate is likely to differ from that of the main body of the species population, and natural selection will then tend to magnify the deviation.

Evolutionary changes associated with the appearance of new species are thus, according to this model, concentrated in speciation events of very short duration; they are represented by the horizontal segments of the tree's branches (Fig. 3*a*). During the long time intervals between successive speciation events, a species remains almost unchanged, as shown by the vertical segments of the branches. While a species is in this static condition, phyletic evolution may bring about minor adjustments; this within-species evolution is known as microevolution. But the magnitude of these phyletic changes is negligible compared with that accompanying macroevolution, the formation of new species. Ernst Mayr introduced the terms maintenance evolution and switch evolution to describe these two evolutionary modes.

Many students of evolution are of the opinion that most groups of organisms evolve in accordance with the punctuated equilibrium model rather than by phyletic gradualism. There are two chief arguments for this view. First, it is clear from the fossil record that many species persist without perceptible change over long stretches of time and then suddenly make large quantum jumps to radically new forms. Second, phyletic gradualism seems to be too slow a process to account for the tremendous proliferation of species needed to supply the vast array of living forms that have come into existence since life first appeared on Earth. *See* ANIMAL EVOLUTION; SPECIES CONCEPT. E. C. Pielou

Bibliography. G. L. Bush, Modes of animal speciation, *Annu. Rev. Ecol. Systemat.*, 6:339-364, 1975; T. Dobzhansky, *Genetics of the Evolutionary Process*, 1972; D. H. Erwin and R. L. Anstey, *New Approaches to Speciation in the Fossil Record*, 1995; V. Grant, *Plant Speciation*, 2d ed., 1981; M. L. McKinney, *Evolution of Life: Processes, Patterns, and Prospects*, 1993.

Species concept

The idea that the diversity of nature is divisible into a finite number of definable species. In general, species concepts grow out of attempts to understand the very nature of biological organization above the level of the individual organism. There are two basic questions: (1) What does it mean to be a species in general? Do all species have certain characteristics, such as forming genealogical lineages, just as all atoms have certain characteristics, such as the ability to undergo chemical reactions? (2) What does it mean to be a particular species? The first question addresses species concepts. The second question addresses how to apply a species concept to living organisms of the world. Does the name Homo sapiens apply to a group of organisms existing in nature? If so, does it belong, as a member, to a natural kind that can be characterized by some set of properties? The difference between the questions of what it is to be a species in general versus what does it mean to be a particular species represents the dividing line between what it is to be a natural kind and what it is to be a natural individual. To understand this distinction, we must first take up the more general question of the nature of kinds. We can then return to the question of species concepts. *See* SPECIATION.

Kinds and particulars. Kinds are attempts to divide the world of particulars according to a definition that specifies properties that separate individuals into two classes: those particulars that have the properties are members of the kind, and those particulars that do not have the properties belong to another kind. What do we mean when we say "particulars"? Particulars are individual things in nature. We can usually point to particulars, such as pointing to a cat or pointing to a truck. A particular belongs to a kind if the particular has the properties that define the kind. For example, we might define the kind "black cat" as those cats that have the property of being covered with black fur. My black cat belongs to this kind, but my gray cat does not since it lacks the property of the kind. In fact, black cats in general (every black cat that has ever lived or will live in the future) belong to the kind "black cat."

As it turns out, there are at least two types of kinds: nominal kinds and natural kinds. The difference between nominal and natural kinds lies in how the kind functions. Natural kinds function in scientific theories of how nature operates, because the theory asserts that certain particulars (individuals) have certain properties. For example, the kind "atom" functions in theories of physics and chemistry as the smallest kind of matter whose members (particular individual atoms) can take part in chemical reactions (the property). In turn, hydrogen is a kind of atom whose members have one proton (property), while helium is a kind of atom whose members have two protons (another property). Kinds such as hydrogen and helium are natural kinds because members of each kind (individual atoms of hydrogen and helium) are predicted to exist under the right circumstances and have their properties as an assertion of our theories of chemistry and physics. In contrast to natural kinds, most of our language is filled with nominal kinds. Nominal kinds are not predicted by scientific theories, but function to allow us to communicate ideas. For example, no scientific theory predicts the kinds "truck" and "bicycle." Yet, nominal kinds have properties (trucks have four wheels, bicycles have two) that allow us to distinguish trucks from bicycles. Thus, nominal kinds may have useful functions in communication, but they are not directly derived from scientific theories. See SCIENTIFIC METHODS.

Concepts and theories. Since Aristotle, biologists have attempted to develop a concept of "species" that will function in biology as a natural kind in a manner similar to natural kinds in other disciplines such as "quark" in physics and "atom" in chemistry. This search was inhibited because philosophers have traditionally thought of particular species (*Homo sapiens, Pinus ponderosa,* etc.) as natural kinds, or sets; or they thought that particular species were nominal—arbitrary but useful inventions of taxonomists to organize the real entities, that is, individual organisms. However, work by Michael

Ghiselin, David Hull, and others asserted that particular species were, in fact, particulars that function in evolutionary theory in the same manner as individual atoms function in chemical theory—as interactors or participants in process. If this is so, then each of these individual species would be members of a natural kind, "species," that had properties predicted from evolutionary theory. Attempts to capture the properties of the kind "species" result in a species concept.

The ideas that organisms could be grouped into more or less discrete units, and the idea that some of these units were more similar to each other than to other units, long predated ideas about evolution. That there was a pattern to life's diversity gave rise to the first species concept: members of one species are different in their physical characteristics from members of other species, and their offspring look like them (like begets like). This concept was the basis for both folk taxonomies (names of different groups of individual organisms in common language) and scientific taxonomies (formal names meant to have scientific import). As the practice of scientific taxonomy developed, species were gathered into larger groups based on similarities and differences. These groupings were hierarchical and seemed to reflect patterns of similarities and differences seen in nature. Current evolutionary theories seek to explain why life's diversity seems to be arranged in the hierarchical manner that we see in nature. See TAXONOMIC CATEGORIES; TAXONOMY; ZOOLOGICAL NOMENCLATURE

Like all scientific theories, theories of evolution predict natural kinds with attendant properties that are found in individuals. Theories of evolution can be generally divided into two basic themes. Microevolutionary theory is concerned with evolutionary change within and between populations of individual organisms. Processes include natural selection, gene flow, and stochastic processes such as genetic drift. These theories do not address the hierarchical patterns that predate Darwin and have fascinated biologists since Linnaeus. Rather, they predict that populations of individual organisms will change through mutation, selection, and drift in predictable ways. Macroevolutionary theory is concerned with the origin and fate of species. It directly addresses the hierarchical nature of diversity. Process theories that purport to explain the hierarchy include speciation and extinction. These two themes, microand macroevolution, intersect at the level of species. From the macroevolutionary perspective, species are those particulars that originate from various processes of speciation and disappear through extinction. Macroevolutionary processes follow the "rules" of microevolutionary processes, but add an extra level of processes such as response to a physical barrier that divides a once continuous population into two populations. Thus, the natural kind "species" has members (particular species) that undergo such processes. See MACROEVOLUTION; ORGANIC EVOLU-TION.

The problem with species concepts derives from

the problem with changing ideas of the world in general. Our concepts of natural kinds, and indeed whether we think that a certain kind is real (has members in the world) or unreal (properties can be listed, but nothing exists that has the properties, like unicorns or fairies), depend on our current theories. The recent controversy concerning the definition of the kind "planet" is an example. Pluto has been rejected (at least for the moment) as being a planet based on current theory, and a new kind, dwarf planet, has been proposed that includes not only Pluto but other smaller bodies in the solar system, such as the asteroid Ceres. The hope is that the closer the theory is to nature, the more predictive will be the kinds and their properties. So it is with species concepts.

Types of species concepts. Species concepts, in one form or another, have been around for several hundred years. Each concept (and there are over 20 current ones) is an attempt to capture the properties of the largest (or smallest) kind and thus permit a search for particulars (Homo sapiens, etc.) that are members of the kind. Of the plethora of species concepts, evolutionary biologists (including systematists and taxonomists) seem to be converging on the concept of species-as-lineages, the Evolutionary Species Concept-species are those things that exist between speciation events and that originate through speciation events (Henning, 1996; Wiley and Mayden, 2000). Other concepts capture part of this more general concept. For example, the venerable Morphological Species Concept claims that different species have different or slightly different physical characteristics. This is true for many species (humans versus chimpanzees), but not true for others where speciation has resulted in behavioral or genetic changes with little or no detectable change in physical characteristics (as in some closely related fruit flies). The popular Biological Species Concept asserts that species have the property of reproductive isolation. This seems to be true for a great number of species, but not for all species and especially not for many species of recent origin. This has led workers such as Richard Mayden to assert that all currently valid species concepts (kinds for which examples can be found in nature) are different reflections of the more general natural kind, that is, the Evolutionary Species Concept. Of course, even if all evolutionary biologists agreed with Mayden, this does not end controversies as to whether a particular species is real. For example, is Pinus ponderosa (a pine tree) real? We might vigorously disagree as to whether or not a speciation event has occurred, based on the evidence presented. We might argue that Pinus ponderosa is not a species but is actually two or more species, or simply a population of another species. However, if we can agree, our agreement will be founded both on the empirical data and on a species concept that applies in an unambiguous manner, leading to a consilience (a synthesis of knowledge) between practice and theory.

E. O. Wiley

Bibliography. M. T. Ghiselin, A radical solution to the species problem, *Systematic Zool.*, 23:536-544,

1974; W. Hennig, *Phylogenetic Systematics*, University of Illinois Press, Urbana, 1966; D. L. Hull, Are species really individuals?, *Systematic Zool.*, 25:174-191, 1976; R. L. Mayden, A hierarchy of species concepts: The denouement in the saga of the species problem, pp. 381-424 in *Species, the Units of Bio-diversity*, ed. by M. F. Claridge, H. A. Dawah, and M. R. Wilson, Chapman and Hall, London, 1997; G. G. Simpson, *Tempo and Mode in Evolution*, Columbia University Press, New York, 1944; E. O. Wiley and R. L. Mayden, The evolutionary species concept, pp. 70-89 in *Species Concepts and Phylogenetic Systematics*, ed. by Q. D. Wheeler and R. Meier, Columbia University Press, New York, 2000.

Specific charge

The ratio of charge to mass, expressed as e/m, of a particle. The acceleration of a particle in electromagnetic fields is proportional to its specific charge. Specific charge can be determined by measuring the velocity v which the particle acquires in falling through an electric potential $V(v = \sqrt{2eV/m})$; by measuring the frequency of revolution in a magnetic field H (the so-called cyclotron frequency $\omega = eH/mc$, where c is the velocity of light); or by observing the orbit of the particles in combined electric and magnetic fields. In the mass spectrograph, the fields are arranged so that particles of differing velocities but of the same e/m are focused at a point. See ELECTRON MOTION IN VACUUM; ELEMENTARY PARTICLE; MASS SPECTRO-Charles J. Goebel SCOPE.

Specific fuel consumption

The ratio of the fuel mass flow of an aircraft engine to its output power, in specified units. Specific fuel consumption (abbreviated sfc or SFC) is a widely used measure of atmospheric engine performance. For reciprocating engines it is usually given in U.S. Customary units of pound-mass per hour per horsepower [(lbm/h)/hp or lbm/(hp·h)], and International System (SI) units of kilograms per hour per kilowatt [(kg/h)/kW]. *See* RECIPROCATING AIRCRAFT ENGINE.

For the gas turbine family of atmospheric aircraft engines, and for ramjets, performance is usually given in terms of thrust specific fuel consumption (abbreviated tsfc or TSFC) expressed as fuel mass flow per unit thrust output with Customary units of pound-mass per hour per pound-force [(lbm/h)/lbf] or SI units of kilograms per hour per newton [(kg/h)/N; 1 N equals approximately 0.225 lbf]. For high-supersonic and hypersonic ramjets, specific fuel consumption is sometimes given in poundmass per second per pound-force [(lbm/s)/lbf] or kilograms per second per newton [(kg/s)/N]. Specific thrust (or specific impulse), which is the inverse of specific fuel consumption, is often used especially for supersonic combustion ramjet (scramjet) performance at Mach numbers above 6 (the Mach number N_{Ma} equals 1.0 at the local speed of sound). *See* AIRCRAFT PROPULSION; TURBINE ENGINE SUBSYSTEMS; MACH NUMBER; PROPULSION; SPECIFIC IMPULSE; TURBINE PROPULSION; TURBOFAN; TURBOJET.

Since the combustion process in atmospheric aircraft engines is supported by oxygen in the ambient air, only the hydrocarbon fuel (gasoline for reciprocating engines and kerosine for turbine engines and ramjets), which must be carried on board the aircraft, needs to be accounted for in determining engine performance. Methane and hydrogen may be used in the future as fuels for atmospheric engines. To obtain a low specific fuel consumption, the fuel should have both a high heat of combustion (that is, energy content per unit mass of fuel) and high cycle efficiency. The cycle efficiency is related to the operating compression ratio of the engine. *See* AIRCRAFT FUEL; GASOLINE; KEROSINE.

A typical specific fuel consumption value for reciprocating engines and turboprops (in terms of equivalent shaft horsepower) is approximately 0.5 (lbm/h)/hp [0.3 (kg/h)/kW]. Gas turbines range between 0.6 (lbm/h)/lbf [0.06 (kg/h)/N] for turbo-fans and 0.8 (lbm/h)/lbf [0.08 (kg/h)/N] for turbo-jets. Ramjet engines with hydrocarbon fuel range upward from a minimum of about 2.0 (lbm/h)/lbf [0.20 (kg/h)/N] at flight speeds between Mach numbers 2 and 4. Values of specific fuel consumption for ramjets and scramjets with hydrogen fuel be-ginning at Mach 4 trend from approximately 0.9 to 2.4 (lbm/h)/lbf [0.09 to 0.24 (kg/h)/N] at Mach 16 and beyond. J. Preston Layton

Bibliography. D. G. Shepherd, Aerospace Propulsion, 1972.

Specific heat

A measure of the heat required to raise the temperature of a substance. When the heat ΔQ is added to a body of mass *m*, raising its temperature by ΔT , the ratio *C* given in Eq. (1) is defined as the heat capacity

$$=\frac{\Delta Q}{\Delta T} \tag{1}$$

of the body. The quantity c defined in Eq. (2) is called

С

$$c = \frac{C}{m} = \frac{1}{m} \frac{\Delta Q}{\Delta T} \tag{2}$$

the specific heat capacity, following the recommendation by the International Union of Pure and Applied Physics (IUPAP). However, this quantity is also often referred to with the shorter name specific heat. A commonly used unit for heat capacity is joule \cdot kelvin⁻¹ (J · K⁻¹); for specific heat capacity, the unit joule \cdot gram⁻¹ · K⁻¹ (J · g⁻¹ · K⁻¹) is often used. Joule should be preferred over the unit calorie = 4.18 J. As a unit of specific heat capacity, Btu \cdot lb⁻¹ · °F⁻¹ = 4.21 J · g⁻¹ · K⁻¹ is also still in use in English-language engineering literature. If the heat capacity is referred to the amount of substance in the body, the molar heat capacity *c_m* results, with the unit

 $J \cdot mol^{-1} \cdot K^{-1}$. The IUPAP recommends distinguishing this quantity from the specific heat capacity *c* through the subscript *m*, although this recommendation is usually disregarded in an attempt to conserve subscripts.

If the volume of the body is kept constant as the energy ΔQ is added, the entire energy will go into raising its temperature. If, however, the body is kept at a constant pressure, it will change its volume, usually expanding as it is heated, thus converting some of the heat ΔQ into mechanical energy. Consequently, its temperature increase will be less than if the volume is kept constant. It is therefore necessary to distinguish between these two processes, which are identified with the subscripts *V* (constant volume) and *p* (constant pressure): *C_V*, *c_V*, and *C_p*, *c_p*. For gases at low pressures, which obey the ideal gas law, the molar heat capacities differ by *R*, the molar gas constant, as given in Eq. (3) where

$$c_p - c_V = R \tag{3}$$

 $R = 8.31 \text{ J} \cdot \text{mol}^{-1} \cdot \text{K}^{-1}$; that is, the expanding gas heats up less. For monatomic gases, for example, $c_p = \frac{5}{2R}$, considerably smaller than $c_V = \frac{3}{2R}$.

For solids, the difference between c_p and c_V is of the order of 1% of the specific heat capacities at room temperature, and varies roughly linearly with temperature. It results from the thermal expansion. This small difference can often be ignored. *See* CALORIMETRY; CHEMICAL THERMODY-NAMICS; HEAT CAPACITY; SPECIFIC HEAT OF SOLIDS; THERMODYNAMIC PROCESSES. Robert 0. Pohl

Bibliography. International Union of Pure and Applied Physics, Document U.I.P. 20, 1978; G. N. Lewis and M. Randall, *Thermodynamics*, 2d ed., 1961; F. W. Sears and G. L. Salinger, *Thermodynamics, Kinetic Theory, and Statistical Mechanics*, 3d ed., 1975; K. Wark and D. E. Richards, *Thermodynamics*, 6th ed., 1999; M. W. Zemansky and R. H. Dittman, *Heat and Thermodynamics*, 7th ed., 1996.

Specific heat of solids

The specific heat (short for specific heat capacity) of a solid is the amount of heat required to increase the temperature of a unit mass of the solid by a unit amount. It is determined by the vibrations of its atoms or excitations of its electrons, and also by a variety of phase transitions. *See* SPECIFIC HEAT.

Contribution of atomic vibrations. If a solid object, say a piece of rock salt, is said to have a certain temperature, it means that its constituent atoms perform motions around their equilibrium positions. The larger the amplitude of these motions, the larger the temperature.

Classical case. In the classical picture of this thermal motion, each atom is considered to be a harmonic, three-dimensional oscillator with six degrees of freedom, three of kinetic and three of potential energy. Each degree of freedom contains on average the energy $(1/2)k_BT$, where *T* is the absolute temperature

[measured in kelvins (K)] and k_B is Boltzmann's constant ($k_B = 1.38 \times 10^{-23} \text{ J} \cdot \text{K}^{-1}$). Thus, an object containing 1 mole or $N_A = 6.02 \times 10^{23}$ atoms, contains the thermal energy Q in the form of atomic vibrations given by Eq. (1), and has the molar heat capacity c_v given by Eq. (2), where R is the molar gas constant,

$$Q = 6(1/2)RT$$
 (1)

$$c_V = 3R \tag{2}$$

and $R = N_A k_B = 8.31 \text{ J} \cdot \text{mol}^{-1} \cdot \text{K}^{-1}$. The molar heat capacity can be converted to specific heat capacity by dividing the molar heat capacity by the mass of 1 mole of the substance. Equation (2) is called the rule of Dulong and Petit, proposed in 1819 on the basis of measurements near room temperature. *See* DEGREE OF FREEDOM (MECHANICS); KINETIC THEORY OF MATTER; TEMPERATURE.

Quantized atomic vibrations. When H. F. Weber (1875) extended the measurements over a wide temperature range (220 to 1300 K or -84 to 1880° F), he found that Eq. (2) holds only in the limit of high temperatures. Figure 1 shows measurements on copper, extended to even lower temperatures: at ~20 K $(-424^{\circ}F)$ the specific heat capacity has dropped to \sim 3% of its room-temperature value which agrees with the value based on the rule of Dulong-Petit; the arrow on the right ordinate of Fig. 1 indicates the molar heat capacity of 3R. The explanation for the decrease at low temperatures was offered by A. Einstein (1907). He suggested that the atomic vibrations are quantized; the atoms cannot vibrate with continuously variable amplitude (or energy). Rather, the atoms can only have energies varying in integer steps of hv, where b is Planck's constant, b = 6.63×10^{-34} J · s, and v is the classical frequency of



Fig. 1. Specific heat capacity for copper (data points), and predictions according to theories of Einstein and Debye (curves), on logarithmic scales. Above \sim 300 K (80°F), the specific heat capacity approaches the value based on the rule of Dulong-Petit, which for copper is $J \cdot g^{-1} \cdot K^{-1}$. The molar heat capacity of 3*R* for copper (molecular weight 63.546) is indicated.



Fig. 2. Specific heat capacity of silicon dioxide (SiO₂) in its crystalline and amorphous (glass) phases. The Debye characteristic temperatures Θ_D , determined from elastic constants, are so similar that the Debye specific heat prediction is shown as a single curve for both solids.

vibration. As the temperature decreases, an increasing fraction of the atoms will not be excited at all; they will be frozen out. This leads to an exponential decrease of the specific heat capacity. Figure 1 shows Einstein's fit to the data, which he obtained by using the frequency $v_E = 4.6 \times 10^{12} \text{ s}^{-1}$ (Einstein frequency for copper) as an adjustable parameter. Through this work, the validity of the quantum concept was demonstrated for the motion of material particles. *See* PLANCK'S CONSTANT; QUANTUM ME-CHANICS.

Quantized elastic waves. In 1911, Einstein realized a shortcoming in his theory: not only did it lead to a drop-off of the specific heat capacity at low temperatures faster than observed experimentally (Fig. 1), but it also led to a thermal conductivity of the wrong magnitude and temperature dependence. He also realized the cause for this problem: he had assumed that each atom vibrates independently of its neighbors; however, every atom, as it vibrates, will push and pull on its neighbors, and this will lead to a coupling between neighboring atoms.

M. Born and T. von Kármám (1912) showed that this coupling will lead to elastic waves, which will propagate with the velocity of sound through the solid. Using the analogy with electromagnetic waves, P. Debye (1912) showed that these elastic waves lead to a slower temperature dependence of the specific heat capacity at low temperatures. According to Debye, the molar heat capacity is given at low temperatures ($T < 0.1\Theta_D$) by Eq. (3), where Θ_D is called

$$c = 234R \left(\frac{T}{\Theta_D}\right)^3 \tag{3}$$

the Debye characteristic temperature and is determined by the elastic constants of the solid, which are measured.

Figure 1 shows the specific heat capacity of copper predicted by Debye's theory. Another comparison is shown in **Fig. 2**, for crystalline quartz. The agreement below 10 K (-442° F) for quartz is excellent and demonstrates the validity of the theoretical model. The discrepancy between theory and experiment above this temperature is also well understood: Not all the elastic waves in quartz have the same speeds of sound; the high-frequency ones, which will be excited predominantly at high temperatures, are particularly variable. This variation of the speeds of sound leads to variations in the number density of the elastic waves that can be thermally excited, and hence will lead to variations in the specific heat.



Fig. 3. Molar heat capacity at constant pressure (c_p) of aluminum. The specific heat capacity is obtained by dividing the molar heat capacity by the molecular weight of aluminum (26.982).

See LATTICE VIBRATIONS; THERMAL CONDUCTION IN SOLIDS.

Difference of specific heats. While in gases the difference between the specific heat capacity at constant pressure and that at constant volume is significant, in solids the difference is relatively small. Therefore, the specific heat is usually measured under constant pressure p (usually p = 0, vacuum). In the example shown in Fig. 3, the molar heat capacity under constant pressure (c_p) of aluminum above room temperature does not saturate at the value given by the Dulong-Petit rule, which is strictly valid only for c_v ; the gradual increase of c_b above room temperature is the result of the thermal expansion of the solid. Above 700 K (800°F), there is also a sudden increase of the specific heat as the melting point T_m (933.2 K or 1220.1°F) is approached. See THERMAL EXPAN-SION.

Electronic specific heat. A typical metal like copper or aluminum contains roughly as many free electrons as atoms. According to classical theory, each electron should contain a thermal energy of $(1/2)k_BT$ for each degree of freedom, three in the case of translational motion. It is, therefore, remarkable that the Debye theory yields such excellent agreement with the experiment for copper, a metal (Fig. 1), considering that this theory completely ignores an electronic contribution. In fact, free electrons are noticed only at lower temperatures. In aluminum, the lattice specific heat according to Debye's theory, shown in Fig. 3, agrees well with the measurements above 20 K (-424° F). Below 1 K (-458° F), the experimental molar heat capacity approaches a linear temperature dependence, also shown in Fig. 3, that is given by Eq. (4). At 1.163 K, aluminum be-

$$c_p = 1.35 \times 10^{-3} \frac{\text{J}}{\text{mol} \cdot \text{K}^2} T$$
 (4)

comes superconducting: its electrical resistivity suddenly drops to zero. By applying a magnetic field (~100 gauss or 0.01 tesla), this transition can be suppressed, and the linear temperature dependence of the electronic specific heat can be observed to the lowest temperatures. The electronic specific heat in the normal (that is, nonsuperconducting) state can be understood on the basis of the Pauli exclusion principle: each quantum state can be occupied by at most one electron, and thus only electrons occupying the uppermost energy states have unoccupied energy states nearby, to which they can move by picking up thermal energy. According to a theory due to A. Sommerfeld, this leads to a molar specific heat given by Eq. (5), where T_F is a characteristic

$$c_V = \frac{\pi^2}{2} \frac{T}{T_F} R \tag{5}$$

temperature called the Fermi temperature, which is of the order of 10^5 K. In deriving this theory, it has been assumed that each atom in the metal supplies one conduction electron. *See* EXCLUSION PRINCIPLE; FERMI-DIRAC STATISTICS; FREE-ELECTRON THEORY OF METALS; SUPERCONDUCTIVITY. When the metal sample is cooled in the absence of a magnetic field below its superconducting transition temperature T_c (1.163 K in aluminum), the specific heat rises abruptly as the electronic energy states are rearranged. Upon further cooling, the specific heat decreases exponentially. However, even at 0.2 K the specific heat of the electrons exceeds that of the atomic vibrations in aluminum (according to the Debye theory) by at least tenfold.

Specific heat of phase transitions. The transition from the normal to the superconducting state shown in Fig. 3 is an example of a phase transition occurring among the metal electrons. Other phase transitions in the solid state are, for example, those from magnetic disorder (paramagnetic) to magnetic order (ferro-, ferri-, or antiferromagnetic), or from electric dipolar disorder (paraelectric) to an ordered phase (for example, ferroelectric). The transition from an ordered to a disordered phase ordinarily occurs at a critical temperature T_c , near which fluctuations between regions of order and disorder cause an increase of the specific heat capacity. Thus a graph of specific heat capacity versus temperature T has a sharp peak at $T = T_c$. See CRITICAL PHENOMENA; PHASE TRANSITIONS.

While much progress has been achieved in the understanding of the approach to phase transitions in the solid state, including the specific heat near the critical temperature T_c , the solid-liquid phase transition (for example, the rapid rise of the specific heat of aluminum near the melting point, shown in Fig. 3) is still only poorly understood.

Amorphous solids. In contrast to crystals, these solids lack long-range order. Their structure resembles that of frozen-in liquids. The specific heat of these solids exceeds that of the Debye prediction, even if they are electrical insulators, notably at the lowest temperatures; below 1 K, it varies approximately linearly with the temperature and has almost the same magnitude (to within a factor of 10) in all amorphous solids. (In amorphous metals, the lowtemperature specific heat will be increased further by the contribution by the electrons.) Figure 2 shows a comparison of the specific heat capacity of the electric insulator silicon dioxide (SiO₂) in the amorphous (silica) and the crystalline (α -quartz) phases. The Debye theory predicts nearly the same heat capacity for both phases, which is therefore shown as a single curve. Obviously, the Debye model is only a poor approximation in the amorphous phase. The physical nature of the excitations giving rise to the linear specific heat anomaly is not yet understood. A particular challenge is the universality of these excitations. See AMORPHOUS SOLID. Robert O. Pohl

Bibliography. N. W. Ashcroft and N. D. Mermin, Solid State Physics, 1976; A. Cezairliyan (ed.), Specific Heat of Solids, 1988; E. S. R. Gopal, Specific Heats at Low Temperatures, 1966; C. Kittel, Introduction to Solid State Physics, 8th ed., 2005; Y. S. Touloukian et al. (eds.), Thermophysical Properties of Matter, vols. 4–6, 1970; G. K. White and P. J. Meeson, Experimental Techniques in Low-Temperature Physics, 4th ed., 2002.

Specific impulse

The impulse produced by a rocket divided by the mass m_p of propellant consumed. Specific impulse I_{sp} is a widely used measure of performance for chemical, nuclear, and electric rockets. It is usually given in seconds for both U.S. Customary and International System (SI) units, but this is not strictly correct.

The impulse produced by a rocket is the thrust force F times its duration t in seconds. The specific impulse is given by Eq. (1). Its equivalent,

$$I_{sp} = \frac{Ft}{m_p} \tag{1}$$

specific thrust F_{sp} , that is sometimes used alternatively, is the rocket thrust divided by the propellant mass flow rate F/\dot{m}_p . See IMPULSE (MECHANICS); THRUST.

The rationale for using seconds as the units of specific impulse is the lack of distinction in U.S. Customary usage between pound-force (lbf) and poundmass (lbm) that are numerically equivalent, so that cancellation leaves units of seconds; however, this procedure cannot be justified in SI units, where the newton is the base unit of force and the kilogram the unit of mass, so that the unit of specific impulse is $N \cdot s/kg$. See UNITS OF MEASUREMENT.

Usage of the quantity effective jet velocity V_{je} is preferred to specific impulse, since it is a more fundamental and better conceived measure of performance. It is related to specific impulse by Eq. (2),

$$V_{je} = g_c I_{sp} \tag{2}$$

where g_c , the standard gravitational acceleration, is 32.18 ft/s² or 9.807 m/s², to give V_{je} in units of ft/s and m/s, respectively.

Effective jet velocity for thermal rockets can be calculated from Eq. (3), where *R* is the universal gas con-

$$V_{je} = \sqrt{\frac{2g_c Rk}{k-1} \frac{T_m}{M}} \left[1 - \left(\frac{p_e}{p_c}\right)^{\frac{(k-1)}{k}} \right]$$
(3)

stant, *k* is the propellant specific heat ratio, T_m is the maximum propellant temperature, *M* is the molecular mass of the propellants, and p_e/p_c is the ratio of nozzle exit to chamber pressure. The importance of high propellant temperature and low molecular mass of the propellant is shown by this equation.

	Propellants	Specific impulse		Effective jet velocity			
Туре		lbf-s/lbm	N-s/kg	ft/s	m/s	Status*	Examples
Thermal rockets							
Cold gas rockets	Nitrogen	50	490	1,609	490	0	Manned Maneuvering Unit (MMU)
Chemical rockets							
Solid propellant	Ammonium perchlorate/organic polymers and powdered aluminum	290	2,844	9,332	2,844	0	Minuteman Space shuttle boosters Titan boosters
Liquid propellant Monopropellant	Hydrazine	230	2 256	7 401	2 256	0	Attitude control of
Monopropenant	Tyurazine	200	2,200	7,401	2,200	0	communication satellites
Bipropellant	Nitrogen tetroxide/monomethyl	310	3,040	9,976	3,040	0	Titan launch vehicle
	hydrazine (hydrazine)	(315)	(3,089)	(10,137)	(3,089)	-	
	Oxygen/kerosine (RP-1)	350	3,432	11,260	3,432	0	V-2*
	Oxygen/hydrogen	450	4,413	14,480	4,413	0	Space shuttle main engine
Nuclear rockets	Hydrogen	900	8,826	28,960	8,826	R&T	
Electric rockets							
Electrothermal rockets							
Resistojet	Hydrazine	300	2,942	9,654	2,942	0	Station keeping for Satcom
	Ammonia	500	4.904	16.090	4.904	R&T	
	Hvdrogen	700	6.865	22.530	6.865	R&T	
Arciet	Hydrazine	500	4.904	16.090	4,904	R&T	
,	Ammonia	830	8,140	26,710	8,140	B&T	
	Hydrogen	1.250	12.260	40.220	12,250	R&T	
Electromagnetic rockets	,	,	,	-, -	,		
Pulsed plasma	Poly(tetrafluoroethylene)	1,000	9,807	32,180	9,807	0	Orbit control for Nova navigation satellite
	Argon	1,500	14,710	48,270	14,710	R&T	······gatori catorito
Magnetoplasmadynamic Electrostatic rockets							
lon	Mercury, argon, xenon	2,000 to	19,610 to 58,840	64,360 to 193,100	19,610 to 58,840	ST&E	

 *O = operational; R&T = research and technology; ST&E = space test and evaluation. † No longer used.
Calculation of specific impulse for the various forms of electric rockets involves electrothermal, resistance or arc heating of the propellant or its ionization and acceleration to high jet velocity by electrostatic or electromagnetic body forces. Ions in the exhaust jets of these devices must be neutralized so the spacecraft will not suffer from space charging or other effects from the plumes of the devices' operation. Nominal values of performance of typical space propulsion systems with various types of rockets and their propellants are given in the **table**. *See* ELECTROTHERMAL PROPULSION; ION PROPULSION; PLASMA PROPULSION; ROCKET PROPUL-SION; SPACECRAFT PROPULSION. J. Preston Layton

Bibliography. G. P. Sutton, *Rocket Propulsion Elements*, 7th ed., Wiley, 2001.

Speckle

The generation of a random intensity distribution, called a speckle pattern, when light from a highly coherent source, such as a laser, is scattered by a rough surface or inhomogeneous medium. Although the speckle phenomenon has been known since the time of Isaac Newton, the development of the laser is responsible for the present-day interest in speckle. Speckle has proved to be a universal nuisance as far as most laser applications are concerned, and only in the mid-1970s did investigators turn from the unwanted aspects of speckle toward the uses of speckle patterns, in a wide variety of applications. *See* LASER.

Basic phenomenon. Objects viewed in coherent light acquire a granular appearance. The detailed irradiance distribution of this granularity appears to have no obvious relationship to the microscopic properties of the illuminated object, but rather it is an irregular pattern that is best described by the methods of probability theory and statistics. Although the mathematical description of the observed granularity is rather complex, the physical origin of the observed speckle pattern is easily described. The surfaces of most materials are extremely rough on the scale of an optical wavelength (approximately 5×10^{-7} m). When nearly monochromatic light is reflected from such a surface, the optical wave resulting at any moderately distant point consists of many coherent wavelets, each arising from a different microscopic element of the surface. Since the distances traveled by these various wavelets may differ by several wavelengths if the surface is truly rough, the interference of the wavelets of various phases results in the granular pattern of intensity called speckle. If a surface is imaged with a perfectly corrected optical system (Fig. 1), diffraction causes a spread of the light at an image point, so that the intensity at a given image point results from the coherent addition of contributions from many independent surface areas. As long as the diffraction-limited point-spread function of the imaging system is broad by comparison with the microscopic surface variations, many dephased coherent contributions add at each image point to give a speckle pattern.

The basic random interference phenomenon underlying laser speckle exists for sources other than lasers. For example, it explains radar "clutter," results for scattering of x-rays by liquids, and electron scattering by amorphous carbon films. Speckle theory also explains why twinkling may be observed for stars, but not for planets. *See* COHERENCE; DIFFRAC-TION; INTERFERENCE OF WAVES; TWINKLING STARS.

Applications. The principal applications for speckle patterns fall into two areas: metrology and astronomical speckle interferometry.

Metrology. In the metrology area, the most obvious application of speckle is to the measurement of surface roughness. If a speckle pattern is produced by coherent light incident on a rough surface, then surely the speckle pattern, or at least the statistics of the speckle pattern, must depend upon the detailed surface properties. If the surface root-mean-square roughness is small compared to one wavelength of the fully coherent radiation used to illuminate the surface, the roughness can be determined by measuring the speckle contrast. If the root-mean-square roughness is large compared to one wavelength, the radiation should be spatially coherent polychromatic, instead of monochromatic; the roughness is again determined by measuring the speckle contrast.

An application of growing importance in engineering is the use of speckle patterns in the study of object displacements, vibration, and distortion that arise in nondestructive testing of mechanical components. The key advantage of speckle methods in this case is that the speckle size can be adjusted to suit the resolution of the most convenient detector, whether it be film or a television camera, while still retaining information about displacements with an accuracy of a small fraction of a micrometer. Although several techniques for using laser speckle in metrology are available, the basic idea can be explained by the following example for measuring vibration. A vibrating surface illuminated with a laser beam is imaged (Fig. 1). A portion of the laser beam is also superimposed on the image of the vibrating surface. If the object surface is moving in and out with a travel of one-quarter of a wavelength or more, the speckle pattern will become blurred out. However, for areas of the surface that are not vibrating (that is, the nodal regions) the eye will be able to distinguish the fully developed high-contrast speckle pattern.

Electronic speckle interferometry (ESPI) is finding much use in industrial applications involving



Fig. 1. Physical origin of speckle for an imaging system.



Fig. 2. Electronic speckle interferogram of vibrating object.

measurement of the vibrational properties of structures. In this technique, a television camera is used to view a speckle pattern. The output of the television camera is processed electronically to emphasize the high spatial frequency structure. The result is viewed on a television monitor (Fig. 2). The bright fringes correspond to the region of the object which are at the nodes of the vibration pattern. By knowing the angles at which the object is illuminated and viewed, it is possible to calculate the amplitude of the vibration from the number of dark fringes present. Thus, simply by looking at the television monitor it is possible to determine the vibrational pattern over the entire object. The amplitude and frequency of the drive source can be changed and instantly the new vibrational pattern can be observed. See NON-DESTRUCTIVE EVALUATION.

Speckle patterns can also be used to examine the state of refraction of the eye. If a diffusing surface illuminated by laser light moves perpendicular to the line of sight of an observer, the speckles may appear to move with respect to the surface. For a normal eye, movement in a direction opposite to that of the surface indicates underaccommodation, and movement with the surface indicates overaccommodation. If the speckles do not move, but just seem to "boil," the observed surface is imaged on the retina. James C. Wyant

Astronomical speckle interferometry. Astronomical speckle interferometry is a technique for obtaining spatial information on astronomical objects at the diffraction-limited resolution of a telescope, despite the presence of atmospheric turbulence.

In the absence of turbulence, a telescope of aperture *D* operating at wavelength λ would produce a diffraction-limited image of a point source of an angular size $\alpha = 1.22 \lambda/D$ (the Rayleigh criterion). When observed through an Earth-based telescope, the image of a source in space is corrupted by distortions introduced during propagation of its light through the Earth's atmosphere. These distorting effects result from variations in the density, temperature, and index of refraction of the atmosphere, and give a blurred image, or seeing disk, which is independent of telescope aperture and seldom less than 1–2 arc-seconds in diameter. A sub-arc-second seeing disk is observed only under exceptionally good seeing conditions, at the most favorable highaltitude telescope locations, when the telescope environment is characterized by uniform temperature and nonturbulent conditions. Even in these optimal conditions, the seeing disk is dozens of times larger than the diffraction-limited angular size of the point source produced by a telescope of several meters' aperture. For example the diffraction-limited angular size of a point source produced at $\lambda = 500$ nanometers with a 157-in. (4-m) telescope is approximately 0.03 arc-second, and the seeing disk obtained at the very best observing conditions is at least 10 times larger.

At very short exposures (10-20 milliseconds), astronomical images consist of a fine-scale structure that fills the seeing disk, very similar to the speckle structures observed with laser-illuminated diffusers. These individual speckles contain spatial frequency information up to the diffraction limit of the telescope. Speckle interferometry techniques extract this information, and therefore reach spatial detail down to the diffraction-limited resolution of the telescope. In this process, long sequences of short-exposure images (each exposure a few milliseconds in duration) are recorded through magnifying optics and narrow-band filters by using detectors such as charge-coupled devices (CCDs) and photoncounting devices. These images are then combined by using various computational algorithms to reconstruct a single image with diffraction-limited resolution. The number of images required for convergence of speckle image reconstruction algorithms increases as the fourth power of the atmospheric seeing. See CHARGE-COUPLED DEVICES; IMAGE PROCESS-ING.

Speckle interferometry techniques have proven to be an invaluable tool for astronomical research, allowing studies of a wide range of scientifically interesting problems. They have been widely used to determine the separation and position angle of binary stars, and for accurate diameter measurements of a large number of stars, planets, and asteroids. Speckle imaging techniques have successfully uncovered details in the morphology of a range of astronomical objects, including the Sun, planets, asteroids, cool giants and supergiants, young stellar objects, the supernova SN1987A in the Large Magellanic Cloud, Seyfert galaxies, and quasars. *See* BINARY STAR; INTERFEROM-ETRY. Margarita Karovska

Bibliography. M. Francon, *Laser Speckle and Applications in Optics*, 1979; R. Jones and C. M. Wykes, *Holographic and Speckle Interferometry*, 2d ed., 1989; S. R. Robinson (ed.), *The Infrared and Electro-Optical Systems Handbook*, 1993; R. S. Sirohi, *Speckle Metrology*, 1993.

Spectral type

A label used to indicate the physical and chemical characteristics of a star, as indicated by study of the star's spectrum. Stars possess a remarkable variety of spectra, some simple, others extraordinarily



Fig. 1. Spectral sequence from B to M. (University of Michigan)

complex. To understand the natures of the stars, it was first necessary to bring order to the subject and to classify the spectra.

Standard sequence. Based in part on earlier work by A. Secchi, who had developed a system that correlated with stellar color, W. C. Pickering initiated the modern system starting about 1890. He ordered the spectra by letter, A through O, largely on the basis of the strengths of the hydrogen lines. He and Mrs. W. P. Fleming then found that several of the letters were unneeded or redundant. On the basis of continuity of lines other than hydrogen, A. Maury and A. J. Cannon found that B preceded A and O preceded B. The result is the classical spectral sequence, OBAFGKM (Fig. 1). Cannon also decimalized the classes, setting up the sequence O5, ..., O9, B0, ..., B9, A0, and so forth. (Not all the numbers are used, however.) The modern Harvard sequence, after the observatory where it was formulated, runs from O2 to M9. Over 350,000 stars, classified by Cannon, were included in the final result, the Henry Draper Catalog. Classes L and T were added to the end of the sequence in 1999. See ASTRONOMICAL CATALOGS.

Class A has the strongest hydrogen lines, B is characterized principally by neutral helium (with weaker hydrogen), and O by ionized helium. Hydrogen weakens notably through F and G, but the metal lines, particularly those of ionized calcium, strengthen. In K, hydrogen becomes quite weak, while the neutral metals grow stronger. The M stars effectively exhibit no hydrogen lines at all, but are dominated by molecules, particularly titanium oxide (TiO). L stars are dominated by metallic hydrides and neutral alkali metals, while T is defined by methane, ammonia, and water. At G, the sequence branches into R and N, whose stars are rich in carbon molecules; class S, in which the titanium oxide molecular bands of class M are replaced by zirconium oxide (ZrO), was introduced in 1923 (**Fig. 2**). Classes R and N are commonly combined into class C.

Spectral variation with temperature. At first appearance, the different spectral types seem to reflect differences in stellar composition. However, within the sequence OBAFGKMLT the elemental abundances are roughly similar. The dramatic variations seen in Fig. 1 are the result of changes in temperature (Fig. 3). The strength of an ion's absorption lines depends on the number of electrons capable of absorbing at those wavelengths, which in turn depends on the product of the atomic absorption probability, the number of electrons excited to the appropriate level of the ion, and the number of ions. In classes M and L (which divide near 2000 K), the temperature is too low to allow a significant number of electrons in the second level of hydrogen, from which the optical Balmer series arises. As a result, there are no hydrogen lines. As temperature climbs and collisional activity increases, more electrons are knocked into level 2 and the Balmer lines strengthen from K through A. Above class A, however, the temperature becomes so high that the hydrogen becomes ionized. That ionization and changes in the opacity of the stellar atmosphere cause the hydrogen lines to weaken

The other ions are subject to the same kind of variation. The second level of neutral helium lies at a very high energy. As a result, the helium lines do not become visible until class B. It also takes a great deal of energy to ionize helium, so the ionized helium lines are not observed until nearly class O. Ionized metal lines are produced at intermediate temperatures and neutral metal lines at low temperatures. Molecules, which are relatively fragile, cannot exist unless the temperature is relatively low.

Spectral variation with composition. The different spectra of the R, N, and S stars, however, are caused by true and dramatic variations in the chemical composition. The M stars have more oxygen than carbon. In the N stars, which have similar temperatures, the ratio is reversed and carbon dominates oxygen. The



Fig. 2. Spectral classes. The stars in the sequence O through T have common chemical compositions. Those in the side branches are rich in carbon.



Fig. 3. Variation in line strength for different ions as a function of temperature and spectral class. In class L (off the right end of the graph), TiO disappears and is replaced by metallic hydrides. (*After H. Voigt, ed., Landolt-Bornstein Numerical Data and Functional Relationships in Science and Technology, Group VI, vol. 1: Astronomy and Astrophysics, Springer-Verlag, 1965*)

R stars are carbon-rich versions of classes G and K. In the cool S stars, carbon and oxygen have about the same abundance. That condition and the increased abundance of zirconium (which has a greater affinity for oxygen than titanium) yield strong zirconium oxide lines. These composition variations are the result of internal thermonuclear processing and convection. *See* STELLAR EVOLUTION.

Morgan-Keenan-Kellman system. In the 1940s, W. W. Morgan, P. C. Keenan, and E. Kellman expanded the Harvard sequence to include luminosity. A system of roman numerals is appended to the Harvard class to indicate position on the Hertzsprung-Russell diagram: I for supergiant, II for bright giant, III for giant, IV for subgiant, and V for dwarf or main sequence. The Morgan-Keenan-Kellman class, which includes the refined Harvard type, is defined by a set of standard stars in the MKK atlas, to which a program star must be compared. On this system, the Sun is a G2 V star. Though some dwarfs are carbonrich, the R, N, and S stars are all giants. L stars are all dwarfs and brown dwarfs, and T stars are all brown dwarfs. Additional refinements have been added to indicate a variety of spectral peculiarities. See ASTRO-NOMICAL SPECTROSCOPY; HERTZSPRUNG-RUSSELL DI-AGRAM: STAR. James B. Kaler

Bibliography. J. B. Hearnshaw, *The Analysis of Starlight: One Hundred Fifty Years of Astronomical Spectroscopy*, Cambridge University Press, 1986; C. Jaschek and M. Jaschek, *The Classification of Stars*, 1990; J. B. Kaler, *Stars and Their Spectra*, Cambridge University Press, 1989, paperback 1997; J. B. Kaler, *The Cambridge Encyclopedia of Stars*, Cambridge University Press, 2006; W. W. Morgan, P. C. Keenan, and E. Kellman, *An Atlas of Stellar Spectra*, University of Chicago Press, 1943.

Spectrograph

An optical instrument that consists of an entrance slit, collimator, disperser, camera, and detector and that produces and records a spectrum. A spectrograph is used to extract a variety of information about the conditions that exist where light originates and along its paths. It does this by revealing the details that are stored in the light's spectral distribution, whether this light is from a source in the laboratory or a quasistellar object a billion light-years away.

Proper spectrograph design takes into account the type of light sources to be measured, and the circumstances under which these measurements will be made. Since observational astronomy presents unusual problems in these areas, the design of astronomical spectrographs may also be unique.

Astronomical versus laboratory design. Astronomical spectrographs have the same general features as laboratory spectrographs (Fig. 1). The width of the entrance slit influences both spectral resolution and the amount of light entering the spectrograph, two of the most important variables in spectroscopy. The collimator makes this light parallel so that the disperser (a grating or prism) may properly disperse it. The camera then focuses the dispersed spectrum onto a detector, which records it for further study.

Laboratory spectrographs usually function properly only in a fixed orientation under controlled environmental conditions. By contrast, most astronomical spectrographs are used on a moving telescope operating at local temperature. Thus, their structures must be mechanically and optically insensitive to orientation and temperature.

The brightness, spectral characteristics, and geometry of laboratory sources may be tailored to



Fig. 1. Basic optical components of a spectrograph.



Fig. 2. Typical laboratory spectrographs. (a) Classical concave grating spectrograph. (b) Czerny-Turner plane grating spectrograph.

experimental requirements and to the capabilities of a spectrograph. Astronomical sources, in the form of images at the focus of a telescope, cannot be manipulated, and their faintness and spectral diversity make unusual and difficult demands on spectrograph performance.

Typical laboratory spectrographs use either concave gratings, which effectively combine the functions of collimator, grating, and camera in one optical element, or plan reflection gratings with spherical reflectors for collimators and cameras. An example of each is shown in Fig. 2. These designs may severely constrain the optical characteristics of the collimator and camera, and sometimes the grating, to obtain acceptable imaging at the detector. As a result, they are generally too limited in the choice of important spectrograph parameters for use in astronomy. In particular, they tend to preclude the use of very short cameras, an essential feature of spectrographs intended for study of faint astronomical objects. The Czerny-Turner design (Fig. 2b) is used in some solar spectrographs, where it offers some advantages and ample light is available. See SPECTRO-HELIOGRAPH; SUN.

To provide maximum flexibility in its application, the collimators, gratings, and cameras of an astronomical spectrograph must be relatively free of aberrations when used in collimated light. Then the optical subassemblies can be interchanged without disturbing the spectrograph's optical performance. Grating interchange changes spectral resolution and wavelength coverage with minimal effect on efficiency. Camera and detector interchange achieves operating characteristics that may be crucial for certain applications. Examples of typical astronomical spectrographs are shown in **Fig. 3**.

Details of astronomical spectrographs. Astronomical spectrographs typically use a pair of adjustable polished slit jaws, which are tilted to the telescope beam so that unused light is reflected to guide optics for accurate image positioning. The collimator may be either an off-axis paraboloid or a two-mirror Cassegrain-type optical system. It has a relatively long focal length and an f/ ratio matching that of the telescope's for optimum slit transmission.

The disperser is usually a plane reflection grating. In laboratory applications, larger gratings are generally used for their increased spectral resolution. However, in astronomy the resolution of the grating (number of grooves times interference order) is seldom realized, and larger gratings are used to obtain longer collimator focal lengths in order to preserve or increase slit demagnification at the camera's focus.

The spectral resolution of many laboratory spectrographs can be increased (grating permitting) only by decreasing the slit width. However, this decrease reduces available light and taxes the imaging capabilities of the optical system and the resolution capabilities of the detector. A better approach, usually used in astronomical spectrographs, is to keep the slit width within a relatively small range of values, for a given camera focal length, and vary spectral resolution by using gratings having different angular dispersions.



Fig. 3. Typical astronomical spectrographs. (a) Spectrograph for use at the Cassegrain focus (often called a Cassegrain spectrograph) using a parabolic collimator and semisolid Schmidt camera. (b) Plan view of a coude spectrograph with four Schmidt cameras and four interchangeable gratings. Camera 1 is the shortest camera and camera 4 is the longest. The spectrograph is shown using camera 2.

The cameras used in astronomy are often quite short (fast) to allow the entrance slit to be as wide as possible for maximum light transmission and to maximize spectral coverage for a given detector size. The camera is often a variation of the classical Schmidt telescope, using a combination of mirrors and ultraviolet transmitting optics. Some spectrographs may have a selection of cameras to permit higher resolutions (with narrower slit openings) when there is sufficient light, or to improve imaging in some wavelength ranges. In particular, very large spectrographs, used at the coude or Nasmyth foci, may have an array of camera focal lengths (Fig. 3b). When this arrangement is combined with interchangeable gratings, it permits a large selection of spectral resolutions and wavelength coverages. See SCHMIDT CAMERA.

In the past, the detector is astronomical spectroscopy was almost always the photographic emulsion. However, image intensifiers and, more recently, the very efficient and accurate cooled chargecoupled devices (CCDs) have become the preferred detectors. *See* CHARGE-COUPLED DEVICES; IMAGE TUBE (ASTRONOMY).

Observing sites and telescopes. Spectrographs are particularly appropriate for use at poor observing sites because, although the atmosphere might reduce available light, it has little effect on its spectrum. Light pollution is also less of a problem, since dispersing the light both dilutes the intensity of the unwanted background and identifies the spectrum of this background so that it can be subtracted later.

Astronomical spectrographs can be used effectively on smaller telescopes. The reason is that star images are typically larger than the entrance-slit width, and the amount of light passing through the slit is thus proportional to the diameter, not area, of a telescope's aperture. Thus, the magnitude limit of a 50-in. (1.25-m) telescope will be only about 1 stellar magnitude fainter than that of a far less costly 20-in. (0.5-m) telescope using the same spectrograph. *See* ASTRONOMICAL SPECTROSCOPY.

Ron Hilliard

Bibliography. W. A. Hiltner (ed.), Astronomical Techniques, 2d ed., 1962; J. James, Spectrograph Design Fundamentals, Cambridge University Press, 2007; J. F. James and R. S. Sternberg, The Design of Optical Spectrometers, 1969; C. R. Kitchin, Astrophysical Techniques, 4th ed., 2003; J. A. Samson and D. L. Ederer (eds.), Vacuum Ultraviolet Spectroscopy, Academic Press, 2000; D. J. Schroeder, Astronomical Optics, 2d ed., Academic Press, 2000.

Spectroheliograph

A spectrographic instrument that produces monochromatic images of the Sun. The principle of operation was first suggested by J. Janssen in 1869, and a photographic version was constructed by G. E. Hale at Mount Wilson Observatory in 1892. Around the same time, H. A. Deslandres devised a closely related instrument for operation at l'Obervatoire de Paris, Meudon. In a simple form of the instrument, an image of the Sun from a solar telescope is focused on a plane containing the entrance slit of the spectroheliograph (see **illustration**). The light passing through the slit is collimated by a concave mirror



Optical configuration of a simple spectroheliograph.

that is tilted such that the light is incident on a plane diffraction grating. Part of the dispersed light from the grating is focused by a second concave mirror, identical to the first mirror, at an exit slit identical to the entrance slit. By symmetry of the optical system, the portion of the solar disk imaged on the entrance slit is reimaged in the plane of the exit slit with the same image scale but in dispersed wavelength. The light imaged along the exit slit then corresponds to the portion of the solar image falling on the entrance slit, but in the light of only a narrow region of the spectrum, as determined by the spectrographic dispersion. The particular wavelength sampled is set by the grating angle. By uniform transverse motion of the instrument such that the entrance slit is scanned across the solar image, the light passing through the exit slit maps out a corresponding monochromatic image of the Sun, which can be recorded photographically with a stationary camera. Alternatively, the solar image can be scanned across the entrance slit of a stationary spectroheliograph, the camera then synchronously moved in step with the image. See DIFFRACTION GRATING.

Narrow-bandpass tunable filters are also used to obtain monochromatic images of the Sun, with the advantage that the full image is recorded in one relatively short exposure. However, the transmittance of such filters is not as spectrally pure as that of a spectroheliograph of the same nominal bandwidth, and extremely narrow passband filters have restricted angular fields. Spectroheliographs also have the advantage that the slit widths can be adjusted to match lines of different spectral widths or other specific observing requirements, or can be set for different image-quality observing conditions. A disadvantage is that the time required to build up an image of the full Sun with a spetroheliograph that uses a photographic film-type camera can be many minutes, during which time critical solar features observed might change. See INTERFERENCE FILTER.

Digital recording has greatly expanded the performance capabilities, versatility, and range of applications of spectoheliographs, reducing the observing time and allowing flexibility in observing programs. Sequential readout of a linear array that is aligned with the exit slit, and that moves with the slit as the solar image is scanned at the entrance slit, provides a fast and convenient observing system. The maximum scan rate is set by the array readout rate. A chargecoupled-device (CCD) array, used directly at the focal plane without a slit, allows simultaneous recording of an image at the wavelength of a spectral line as well as the adjacent spectral region. Computer processing of the digital data can then provide velocity and magnetic field information, and extensive solar fields can be recorded in a few seconds. Digital video camera systems have been used for such applications. Without an exit slit, the optical configuration in this case is simply that of a spectrograph, except for an image scanning capability. *See* ASTRONOMICAL IMAG-ING; CHARGE-COUPLED DEVICES; SPECTROGRAPH.

Spectroheliographs have been designed that have more than one camera to record several spectral regions simultaneously, either in the same spectral order or in more than one order of the diffraction grating. Spectroheliographs have also been used at ultraviolet, extreme ultraviolet (EUV), and x-ray wavelengths, and a miniature version has been used successfully in spacecraft observations of the solar corona. Raymond N. Smartt

Bibliography. R. Beck et al. (eds.), *Solar Astronomy Handbook*, Willmann-Bell, Richmond, Virginia, 1995; P. Foukal, *Solar Astrophysics*, 2d ed., Wiley, New York, 2004; K. J. H. Phillips, *Guide to the Sun*, Cambridge University Press, 1992; M. Stix, *The Sun*, Springer-Verlag, Berlin, 1989.

Spectroscopy

An analytic technique concerned with the measurement of the interaction (usually the absorption or the emission) of radiant energy with matter, with the instruments necessary to make such measurements, and with the interpretation of the interaction both at the fundamental level and for practical analysis. Mass spectroscopy is not concerned with the interaction of light with matter, but was so named because the appearance of the data resembles that of the spectroscopic data as just defined.

A display of such data is known as a spectrum, that is, a plot of the intensity of emitted or transmitted radiant energy (or some function of the intensity) versus the energy of that light. Spectra due to the emission of radiant energy are produced as energy is emitted from matter, after some form of excitation, then collimated by passage through a slit, then separated into components of different energy by transmission through a prism (refraction) or by reflection from a ruled grating or a crystalline solid (diffraction), and finally detected. Spectra due to the absorption of radiant energy are produced when radiant energy from a stable source, collimated and separated into its components in a monochromator, passes through the sample whose absorption spectrum is to be measured, and is detected. Instruments which produce spectra are known as spectroscopes, spectrometers, spectrographs, and spectrophotometers. See SPECTRUM.

Interpretation of spectra provides fundamental information on atomic and molecular energy levels, the distribution of species within those levels, the nature of processes involving change from one level to another, molecular geometries, chemical bonding, and interaction of molecules in solution. At the practical level, comparisons of spectra provide a basis for the determination of qualitative chemical composition and chemical structure, and for quantitative chemical analysis.

Early history. In the historical development of spectroscopy, following the fundamental studies of crude spectra of sunlight by Isaac Newton in 1672, certain contributions and achievements are especially noteworthy. The significance of using a narrow slit instead of a pinhole or round aperture so as to produce spectra lines, each one an image of the slit and representing a different color or wave length, was demonstrated independently by W. H. Wollaston in 1802 and by Joseph Fraunhofer in 1814. Fraunhofer made many subsequent contributions to optics and spectroscopy, including first observation of stellar spectra, discovery and construction of transmission diffraction gratings, first accurate measurements of wavelengths of the dark lines in the solar spectrum, and invention of the achromatic telescope. The origin of the dark Fraunhofer lines in the solar spectrum was accounted for by G. R. Kirchhoff in 1859 on the basis of absorption by the elements in the cooler Sun's atmosphere of the continuous spectrum emitted by the hotter interior of the Sun. Further studies by Kirchhoff with R. Bunsen demonstrated the great utility of spectroscopy in chemical analysis. By systematically comparing the Sun's spectrum with flame or spark spectra of salts and metals, they made the first chemical analysis of the Sun's atmosphere. In 1861, while investigating alkali metal spectra, they discovered two new alkali metals, cesium and rubidium. These achievements by Kirchhoff and Bunsen provided tremendous stimulus to spectroscopic researches. The adoption in 1910 of the first international standards of wavelength gave further impetus. These and later standards made possible the measurement of wavelengths of any electromagnetic radiation with unprecedented accuracy. Since World War II, remarkable developments in spectroscopy have occurred in instrumentation, achieved largely through advances in electronics and in manufacturing technology. Direct reading, automatic recording, improved sensitivity with good stability, simplicity of operation, and extended capabilities are features provided by many commercial instruments, many of which now are microprocessor-controlled. Many newer instruments have dedicated data systems. Predictably, these developments, by facilitating widespread use of spectroscopic techniques, have had an enormous influence in promoting developments in both applied and theoretical spectroscopy.

The ultimate standard of wavelength is that of the meter, defined since 1983 as the length of the path traveled by light in vacuum during a time interval of 1/299,792,458 of a second. *See* WAVELENGTH STAN-DARDS.

Spectroscopic units. The change in energy of an ion, atom, or molecule associated with absorption and emission of radiant energy may be measured by the frequency of the radiant energy according to Max Planck, who described an equality E = hv, where E

is energy, v is the frequency of the radiant energy, and *b* is Planck's constant. The frequency is related to the wavelength λ by the relation $v\lambda = c/n$, where *c* is the velocity of radiant energy in a vacuum, and *n* is the refractive index of the medium through which it passes; *n* is a measure of the retardation of radiant energy passing through matter. The units most commonly employed to describe these characteristics of light are the following:

Wavelength: 1 micrometer $(\mu m) = 10^{-6} m$ 1 nanometer $(nm) = 10^{-9} m$ (= 10 angstroms)Frequency: 1 hertz (Hz) $= 1 \text{ s}^{-1}$

For convenience the wave number $\bar{\nu}$ (read nu bar), the reciprocal of the wavelength, may be used; for this, the common units are cm⁻¹, read as reciprocal centimeters or occasionally kaysers. This number equals the number of oscillations per centimeter.

Spectral regions. Visible light constitutes only a small part of the spectrum of radiant energy, or electromagnetic spectrum; the human eye responds to electromagnetic waves with wavelengths from about 380 to 780 nm, though there is individual variation in these limits. The eye cannot measure color and intensity quantitatively, even for the visible portion of the electromagnetic spectrum; therefore, instruments are used for measurements and recording.

There are broad regions of the electromagnetic spectrum with associated wavelengths greater or less than the visible region (see **table**).

Origin of spectra. Atoms, ions, and molecules emit or absorb characteristically; only certain energies of these species are possible; the energy of the photon (quantum of radiant energy) emitted or absorbed corresponds to the difference between two permitted values of the energy of the species, or energy levels. (If the flux of photons incident upon the species is great enough, simultaneous absorption of two or more photons may occur.) Thus the energy levels may be studied by observing the differences between them. The absorption of radiant energy is accompanied by the promotion of the species from a lower to a higher energy level; the emission of radiant energy is accompanied by falling from a higher to a lower state; and if both processes occur together, the condition is called resonance.

Transitions. Transitions between energy levels associated with electrons or electronic levels range from the near infrared, the visible, and ultraviolet for outermost, or highest-energy, electrons, that is, those which can be involved in bonding, to the x-ray region for the electrons nearest the nucleus. At low pressures such transitions in gaseous atoms produce sharply defined lines because the energy levels are sharply defined. Transitions between energy levels are sharply defined. Transitions between energy levels of the nucleus are observed in the gamma-ray region. In the absence of an applied electric or magnetic field, these electronic and nuclear transitions are the only ones which atoms can undergo. *See* ATOMIC SPEC-TROMETRY; ATOMIC STRUCTURE AND SPECTRA.

Electronic transitions in molecules are also observed. In addition, transitions can occur between

Spectral region	Approximate wavelength range	Typical source	Typical detector	Energy transitions studied in matter
Gamma	1–100 pm	Radioactive nuclei	Geiger counter; scintillation counter	Nuclear transitions and disintegrations
X-rays	6 pm-100 nm	X-ray tube (electron bombardment of metals)	Geiger counter	lonization by inner electron removal
Vacuum ultraviolet	10-200 nm	High-voltage discharge; high- vacuum spark	Photomultiplier	lonization by outer electron removal
Ultraviolet	200–400 nm	Hydrogen-discharge lamp	Photomultiplier	Excitation of valence electrons
Visible	400–800 nm	Tungsten lamp	Phototubes	Excitation of valence electrons
Near-infrared	0.8–2.5 μ m	Tungsten lamp	Photocells	Excitation of valence electrons; molecular vibrational overtones
Infrared	2.5–50 μm	Nernst glower; Globar lamp	Thermocouple; bolometer	Molecular vibrations: stretching bending, and rocking
Far-infrared	50–1000 μ m	Mercury lamp (high-pressure)	Thermocouple; bolometer	Molecular rotations
Microwave	0.1-30 cm	Klystrons; magnetrons	Silicon-tungsten crystal; bolometer	Molecular rotations; electron spin resonance
Radio-frequency	10 ⁻¹ –10 ³ m	Radio transmitter	Radio receiver	Molecular rotations; nuclear magnetic resonance

levels associated with the vibrations and rotations of molecules. Spacings between electronic levels are greater than between vibrational levels, and those between vibrational levels are greater than between rotational levels; each vibrational level has a set of rotational levels associated with it, and each electronic level a set of vibrational levels (Fig. 1). Transitions between vibrational levels of the same electronic level correspond to photons in the infrared region; transitions between rotational levels, to photons in the far-infrared and microwave region. Rotational spectra of gaseous molecules consist of sharp lines; vibrational spectra consist of bands, each of which arises from a given transition between vibrational levels altered in energy by increments due to changes in rotation occurring when the vibrational level changes. Likewise, molecular electronic spectra consist of bands due to transitions between electronic energy levels, altered by increments resulting from changes in vibration and rotation of the molecule on changing electronic state. See BAND SPECTRUM; INFRARED SPECTROSCOPY; LINE SPECTRUM; MICRO-WAVE SPECTROSCOPY; MOLECULAR STRUCTURE AND SPECTRA; SPECTRUM ANALYZER.



Fig. 1. Spacing of singlet electronic states S, vibrational states v, and rotational states J.

External fields. The application of a magnetic or electric field to the sample often separates normally indistinguishable, or degenerate, states in energy from each other. Thus, the orientation of the spin of an unpaired electron in an atom, ion, or molecule with respect to an applied magnetic field may have different values, and in the magnetic field the atom or molecule may have different energies. For typical field strengths, the difference in energy between these newly separated energy levels occurs in the microwave region. Similarly, for an atomic nucleus in an ion or molecule, differences in the orientation of the spin of the nucleus with respect to a magnetic field give rise to different energy levels of the nucleus in that field, so that energy differences will be found in the microwave region. The former phenomenon produces electron spin resonance or electron paramagnetic resonance spectra, and the latter produces nuclear magnetic resonance spectra (Fig. 2).

Nuclear magnetic resonance has been advanced by pulsed irradiation techniques which permit the identification of solid samples and complex biological molecules, and which are useful for the production of three-dimensional plots to permit analysis across the interior of objects without destruction. In medicine the imaging of internal organs is known as magnetic resonance imaging. Electron paramagnetic spectroscopy is used to establish structures of species containing unpaired electrons in solution or in a crystalline solid; some qualitative and quantitative analysis is also performed. *See* ELECTRON PARAMAGNETIC RESONANCE (EPR) SPECTROSCOPY; MEDICAL IMAGING; NUCLEAR MAGNETIC RESONANCE (NMR).

Electronic spectra are also altered by external fields; removal of degeneracy by an externally applied electric field is termed the Stark effect, and removal of degeneracy by an externally applied magnetic field is termed the Zeeman effect. *See* STARK EFFECT; ZEEMAN EFFECT.



Fig. 2. Splitting degenerate energy levels of an appropriate atom in a magnetic field. Transitions between the separated levels require interaction with a photon of energy ΔE .

Spontaneous emission. The spontaneous emission of light from a sample by the decay of an electron from a higher level to the lowest available level is known either as fluorescence or as phosphorescence. The fundamental difference between these two terms is associated with the orientation of the spin of this electron. If the spin of the electron is such that it falls to the lowest state without a change of spin (the atom or molecule having the same spin before and after the transition), fluorescence occurs, and the process is characterized as allowed and is relatively fast. If the spin is such that the electron can fall to the lower state only with a change in spin dictated by the Pauli exclusion principle, the process is phosphorescence, characterized as a forbidden process, which is relatively slow (Fig. 3). In practice, other factors also govern the time the electron spends in the higher level, and the time intervals associated with each process overlap; that for fluorescence is 10^{-4} to 10^{-8} s, and that for phosphorescence is 10^{-4} to 10 s. The spontaneous emission process thus cannot be completely distinguished merely on the basis of the time interval associated with the emission. Timeresolved emission studies may allow separation of the processes. Fluorescence and phosphorescence are measured at a right angle to the incident beam exciting the sample, in order to avoid background problems. The emission of light beyond blackbody radiation when a sample is heated (thermolumines-



Fig. 3. Spontaneous emission of light by the decay of an electron from a higher to the lowest available level. (a) Fluorescence. (b) Phosphorescence.

cence) or when it is strained or fractured (triboluminescence) has limited applicability. Similar emission during chemical reactions (chemiluminescence; in biochemical systems, bioluminescence) has broader analytical use if the reactions in which products emit light are, for example, catalyzed by a small amount of the substance to be analyzed. *See* BIOLUMINES-CENCE; CHEMILUMINESCENCE; EXCLUSION PRINCIPLE; FLUORESCENCE; PHOSPHORESCENCE.

Light rotation. Compounds whose molecules have a structure which cannot be superimposed on its reflection in a mirror (asymmetric molecules) rotate the plane of polarized light. If a device can be included in an instrument to plane-polarize the light from the source, that is, to produce light oscillating in only one plane, the amount of rotation of that plane by the sample can be studied as a function of the wavelength of light used in the experiment. A plot of this rotation versus wavelength is called an optical rotatory dispersion curve; such curves are useful in studying fundamentals of light rotation and in establishing the absolute structure of asymmetric molecules empirically. Asymmetric samples whose absorption of the components of plane-polarized light differs are said to exhibit circular dichroism; analysis of a sample in a magnetic field (magnetic circular dichroism) has a broader applicability in structural analysis. See COTTON EFFECT; OPTICAL ACTIVITY.

Instrumentation. Spectrometers require a source of radiation, a dispersion device, and a detector. For emission spectra, the source may be taken as the sample to be measured itself, although another source of radiation may be needed to excite the sample to an excited state. Absorption spectra require a source to generate appropriate radiation across the range of energies needed for the experiment. These radiant sources produce continuous spectra and can be distinguished from those producing discontinuous spectra.

Continuous spectra. Of the sources listed in the table, most of those from the vacuum ultraviolet region to the far-infrared region produce continuous spectra. Within this type, sources for the near ultraviolet, visible, and infrared regions consist of hot, incandescent solids; the maximum intensity of their emission occurs at a wavelength dependent on their temperature and emissivity. Electrical discharges in gases at higher pressures provide sources of continuous radiation for the vacuum ultraviolet and ultraviolet regions; hydrogen or deuterium gas discharges are especially used for the latter. X-ray continuous emission can be produced by collision of electrons or other particles accelerated through tens of kilovolts with a target. *See* BEAM-FOIL SPECTROSCOPY.

Discontinuous spectra. The sources of high-energy radiation which yield discontinuous spectra emit because of discrete processes in individual atoms, ions, or molecules. Such sources may include flames and furnaces to vaporize and atomize samples, and electrical discharges in gases at lower pressures. Other atomizing sources may be produced by high-frequency radio discharges and microwave discharges through a flowing gas. For sources of radiation connected with nuclear processes, solid samples may be used containing the element to be studied, either a radioactive isotope or one which can be made radioactive by particle bombardment. *See* ARC DISCHARGE.

Among these so-called line sources are hollowcathode lamps, in which a few layers of a cylindrical cathode containing the same element as is to be analyzed are vaporized by collisions with ions generated from rare gases and accelerated electrons. The gaseous atoms, also excited to higher energy states by this process, emit characteristic line spectra useful of analysis of that element. Lasers have become an especially useful source, and have opened different areas of spectroscopy, not only because of the intensity of the radiation supplied to the sample, which permits absorption of several photons in a single process, but also because of the coherent properties of the radiation, which permit the study of events occurring on a picosecond time scale. See LASER; ULTRAFAST MOLECULAR PROCESSES.

The sources for radio-frequency studies are tunable radio-frequency oscillators, designed with special regard for stability. The microwave sources may be spark discharges, magnetrons, or most commonly klystrons. An electron beam traveling through a closed cavity of defined geometry sets up electromagnetic oscillations, of which a certain value will be reinforced as a function of the dimension of the containing unit; the value can be varied somewhat by adjusting the size of the cavity. *See* KLYSTRON; MAGNETRON.

Dispersive elements. There are two kinds of dispersive elements, prisms and diffraction elements. The earlier and more commonly available prism has been supplanted in many cases by diffraction elements.

After collimation of the source radiation through a slit, diffraction is achieved in x-ray spectroscopy by the use of a pure crystal. The distances between atomic nuclei in the crystal are on the same order as the wavelength of the radiant energy; under these conditions an efficient dispersion of the radiation over an arc of many degrees is possible. For a distance d between repeating planes of atoms, the wavelength λ will be diffracted through an angle θ related by the Bragg law $n\lambda = 2d \sin \theta$, where *n* is an integer called the diffraction order. Materials used include gypsum (CaSO₄ · 2H₂O), ammonium dihydrogen phosphate (NH₄H₂PO₄), and alkali halides (lighter elements). Each of these gives a wide angular dispersion in a portion of the 0.03-1.4-nm range of the spectrum commonly used for x-ray studies; the appropriate one must be chosen according to the experiment to be performed. See X-RAY OPTICS.

For spectroscopic techniques using visible light and the regions adjacent to it, ultraviolet and infrared, ruled diffraction gratings are often employed. Because their production has become economical, they have largely replaced prisms, which may require the use of several materials to get adequate angular dispersion and wavelength resolution over the entire range of interest, for example, 2– 15 μ m for an infrared spectrum. The dispersion of a prism increases near the limit of its transparency, and there is always a compromise in efficiency between transmission and dispersion. The resolution *R*, defined as the quotient of the wavelength of a line and the wavelength difference between it and another line just separated from it, is given by $T/(dn/d\lambda)$, where *T* is the thickness of the prism base and $dn/d\lambda$ is the variation of the refractive index with respect to wavelength. Efficient prisms are made of quartz for the ultraviolet, glass for the visible, and various salts for the infrared. *See* DIFFRACTION GRATING; OPTICAL MATERIALS; OPTICAL PRISM.

Very short- or very long-wavelength spectra are usually produced without a dispersive element; gamma-ray detection uses a wavelength-sensitive detector (a pulse height discriminator, for example), and microwave and radio-frequency detection may use a variable tuning radio receiver; in the latter case the source is tuned to emit a highly resolved frequency as well.

Detectors. Detectors commonly used for the various spectral regions are listed in the table. The devices for the infrared and microwave regions are basically heat sensors. Except for the eye and photographic methods used sometimes in the regions for gamma ray to visible, each detector converts the signal into an electrical response which is amplified before transmittal to a recording device. Direct imaging of the response by vidicon detection has also been used. Otherwise an oscilloscope or oscillograph is used to record very rapidly obtained signals. *See* GAMMA-RAY DETECTORS; OSCILLOSCOPE; TELEVISION CAMERA TUBE.

Instruments. Spectroscopic methods involve a number of instruments designed for specialized applications.

Spectroscope. An optical instrument consisting of a slit, collimator lens, prism or grating, and a telescope or objective lens which produces a spectrum for visual observation is called a spectroscope. The first complete spectroscope was assembled by Bunsen in 1859 (**Fig. 4**).

Spectrograph. If a spectroscope is provided with a photographic camera or other device for recording the spectrum, the instrument is known as a spectrograph. For recording ultraviolet and visible spectra, two types of quartz spectrographs are in common use. One type utilizes a Cornu quartz prism constructed of two 30° prisms, left- and right-handed, as



Fig. 4. Prism spectroscope or spectrograph. S, slit; C, collimator lens, P, 60° prism; T, telescope; RV, red to violet spectrum. For producing ultraviolet spectra, the glass items in C, P, and T must be replaced by equal amounts of left- and right-handed crystal quartz.

well as left- and right-handed lenses, so that the rotation occurring in one-half the optical path is exactly compensated by the reverse rotation in the other. The other type of quartz spectrograph employs a Littrow 30° quartz prism with a rear reflecting surface that reverses the path of the light through prism and lens, thus compensating for rotation of polarization in one direction by equal rotation in the opposite direction. Thus, in either type the effect of optical activity and birefringence in crystal quartz which produces double images of the slit is eliminated.

Grating spectrographs cover a much broader range of wavelengths (vacuum ultraviolet to far infrared) than do prism instruments. Various-type mountings are employed. The most common mounting for a plane-reflection grating is in a Littrow mount, entirely analogous to that of the Littrow quartz spectrograph. Mountings for concave reflection gratings require that the grating, slit, and camera plate all lie on the Rowland circle (imaginary circle of radius equal to one-half the radius of curvature of the grating) in order to achieve proper focus of spectral lines (slit images) on the plate. Paschen, Rowland, Eagle, and Wadsworth mountings are common in grating spectrographs.

Spectrometers. A spectroscope that is provided with a calibrated scale either for measurement of wavelength or for measurement of refractive indices of transparent prism materials is called a spectrometer. Also, the term frequently is used to refer to spectrographs which incorporate photoelectric photometers instead of photographic means to measure radiant intensities.

Spectrophotometer. A spectrophotometer consists basically of a radiant-energy source, monochromator, sample holder, and detector (Fig. 5). It is used for measurement of radiant flux as a function of wavelength and for measurement of absorption spectra. The wavelength is scanned by the monochromator at a constant rate (geared to chart speed of recorder) while the detector responds to, and discriminates between, the two alternating beams, one from the sample and the other from the attenuator. The servomechanism drives the attenuator, as well as the recorder pen, so that the intensities of the alternating signals are kept identical. The displacement of the pen on the chart can be calibrated in terms of relative intensity, percent transmittance, or absorbance. Generally a linear attenuator is employed, and the pen displacement is directly proportional to the attenuator displacement and thus the transmittance.

Interferometer. This optical device divides a beam of radiant energy into two or more parts which travel different paths and then recombine to form interference fringes. Since an optical path is the product of the geometric path and the refractive index, an interferometer measures differences of geometric path when two beams travel in the same medium, or the difference of refractive index when the geometric paths are equal. Interferometers are employed for high-resolution measurements and for precise determination of relative wavelengths: they are capable of



Fig. 5. Schematic diagram of a double beam (in time) spectrophotometer.

distinguishing between two spectral lines that differ by less than 10^{-6} of a wave. *See* INTERFERENCE OF WAVES; INTERFEROMETRY.

Quantitative relationships. For practical analysis, the absorption of radiant energy is related to the concentration of the sample by the relationship $-\log P/P_0 = abc$, where *P* and P_0 are respectively the power of the light attenuated after passing through the sample and its unattenuated power before reaching the sample, *c* is the concentration of the sample, *b* is the path length of the light through the sample, and *a* is a proportionality constant called the absorptivity of the sample. If *c* is in moles per liter, the absorptivity is called the molar absorptivity and symbolized ϵ . This relationship is known as the Beer-Lambert-Bouguer law, or simply Beer's law. For x-ray spectroscopy, the right-hand variables are grouped differently.

The practical relation between emission of light and concentration is given by the equation $F = k\phi I_0 abc$, where *F* is the fluorescence intensity, *k* is an instrumental parameter, ϕ is the fluorescence efficiency (quanta emitted per quantum absorbed), I_0 is the intensity of the incident light upon the sample, and the other units are as defined previously. The choice between absorption and fluorescence techniques for determining concentration is sometimes made on the basis of the accuracy of the linear method (fluorescence) versus the versatility of the logarithmic method (absorption). *See* LUMINES-CENCE ANALYSIS.

Other methods and applications. Since the early methods of spectroscopy there has been a proliferation of techniques, often incorporating sophisticated technology.

Acoustic spectroscopy. When modulated radiant energy is absorbed by a sample, its internal energy increases. The loss of that excess produces a temperature increase that can be monitored as a periodic pressure change in the gas around the sample by using a microphone transducer. This is the optoacoustic effect. Its application provides a rapid method for study of some difficult samples. *See* PHO-TOACOUSTIC SPECTROSCOPY.

Astronomical spectroscopy. The radiant energy emitted by celestial objects can be studied by combined spectroscopic and telescopic techniques to obtain information about their chemical composition, temperature, pressure, density, magnetic fields, electric forces, and radial velocity. Radiation of wavelengths much shorter than 300 nm is absorbed by the Earth's atmosphere, and can only be studied by spectrographs transported into space by rockets. *See* ASTRONOMICAL SPECTROSCOPY; SPECTROHELIO-GRAPH.

Atomic absorption and fluorescence spectroscopy. This branch of electronic spectroscopy uses line spectra from atomized samples to give quantitative analysis for selected elements at levels down to parts per million, on the average. However, detection limits vary greatly from element to element. Plasmas, especially inductively coupled plasmas, provide a stable, efficient, novel method for atomization and ionization.

Attenuated total reflectance spectroscopy. Spectra of substances in thin films or on surfaces can be obtained by the technique of attenuated total reflectance or by a closely related technique called frustrated multiple internal reflection. In either method the sample is brought into contact with a total-reflecting trapezoid prism and the radiant-energy beam is directed in such a manner that it penetrates only a few micrometers of the sample one or more times.

The technique is employed primarily in infrared spectroscopy for qualitative analysis of coatings and of opaque liquids.

Electron spectroscopy. This area includes a number of subdivisions, all of which are associated with electronic energy levels. The outermost or valence levels are studied in photoelectron spectroscopy, which uses photons of the far-ultraviolet region to remove electrons from molecules and to infer the energy levels of the remaining ion from the kinetic energy of the expelled electron. This technique is used mostly for fundamental studies of bonding. Electron impact spectroscopy uses low-energy electrons (0–100 eV) to yield information similar to that observed by visible and ultraviolet spectroscopy, but governed by different selection rules.

X-ray photoelectron spectroscopy (XPS), also called electron spectroscopy for chemical analysis (ESCA), uses x-ray photons to remove inner-shell electrons in a similar way, and the kinetic energy of the removed electron is measured. In Auger spectroscopy, an x-ray photon removes an inner electron. Another electron falls from a higher electron to take its place, and a third is ejected to conserve energy, since the second electron gains energy when it falls to the lower level. The kinetic energy of the third electron is measured. The chemical elements contained within a sample can then be identified, except for hydrogen. The energies of inner-shell electrons are dependent (in the third or fourth significant figure) upon the oxidation state of the atom, and so the kinetic energy of an XPS or Auger electron often provides information on the oxidation states of the elements in a compound too. Both techniques are used to study surfaces, since x-rays penetrate many objects for only a few atomic or molecular layers. The x-rays may be directed at the sample at various

angles to the normal, and because they penetrate the sample for a more or less fixed distance, x-rays that are directed at the sample at a large angle from the normal (for example, 80°) encounter far fewer atomic or molecular layers than those that enter at smaller angles from the normal. This angle-resolved x-ray photoelectron spectroscopy permits more refinement in the analysis of the first few surface layers of a material. Ion neutralization spectroscopy uses protons or other charged particles instead of photons. *See* AUGER EFFECT; ELECTRON SPECTROSCOPY; SURFACE AND INTERFACIAL CHEMISTRY; SURFACE PHYSICS.

Fourier transform spectroscopy. This family of techniques consists of a fundamentally different method of irradiation of the sample, in which all pertinent wavelengths simultaneously irradiate it for a short period of time and the absorption spectrum is obtained by mathematical manipulation of the cyclical power pattern so obtained. It has been applied particularly to infrared spectrometry and nuclear magnetic resonance spectrometry, and allows the acquisition of spectra from smaller samples in less time, with high resolution and wavelength accuracy. The infrared technique is carried out with an interferometer of conventional instrumentation. *See* FOURIER SERIES AND TRANSFORMS.

Gamma-ray spectroscopy. Of special note in this category are the techniques of activation analysis, which are performed on a sample without destroying it by subjecting it to a beam of particles, often neutrons, which react with stable nuclei present to form new unstable nuclei. The emission of gamma rays for each element at different wavelengths with different half-lives as these radioactive nuclei decay is characteristic. *See* ACTIVATION ANALYSIS; NEUTRON SPECTRO-METRY.

Mössbauer spectroscopy results from resonant absorption of photons of extremely high resolution emitted by radioactive nuclei in a crystalline lattice. The linewidth is extremely narrow because there is no energy loss for the nucleus recoiling in the lattice, and differences in the chemical environment of the emitting and absorbing nuclei of the same isotope may be detected by the slight differences in wavelength of the energy levels in the two samples. The difference is made up by use of the Doppler effect as the sample to be studied is moved at different velocities relative to the source, and the spectrum consists of a plot of gamma-ray intensity as a function of the relative velocity of the carriage of the sample. It yields information on the chemical environment of the sample nuclei and has found application in inorganic and physical chemistry, solidstate physics, and metallurgy. See DOPPLER EFFECT; MÖSSBAUER EFFECT.

Laser spectroscopy. Laser radiation is nearly monochromatic, of high intensity, and coherent. Some or all of these properties may be used to acquire new information about molecular structure. In multiphoton absorption the absorption of more than one photon whose energy is an exact fraction of the distance between two energy levels allows study of transitions ordinarily forbidden when only a single photon is involved. Multiphoton ionization refers to the removal of an election after the absorption of several photons. Raman spectroscopy has also become more common because of the use of lasers as sources, especially with the development of resonance Raman spectroscopy.

Information on processes which occur on a picosecond time scale can be obtained by making use of the coherent properties of laser radiation, as in coherent anti-Stokes-Raman spectroscopy. Lasers may also be used to evaporate even refractory samples, permitting analysis of the atoms so obtained, and since the radiation may be focused, such a technique may be applied to one small area or a sample surface at a time, permitting profiling of the analysis across the sample surface. Laser fluorescence spectroscopy provides the lowest detection limits for many materials of interest in biochemistry and biotechnology, and intense work on development of rapid, sensitive methods proceeds. This will have application in the research that involves mapping of the human genome. See GENETIC MAPPING.

Ultrafast laser spectroscopy may be used to study some aspects of chemical reactions, such as transition states of elementary reactions and orientations in bimolecular reactions. *See* LASER SPECTROS-COPY.

Mass spectrometry. The source of a mass spectrometer produces ions, often from a gas, but also in some instruments from a liquid, a solid, or a material adsorbed on a surface. Ionization of a gas is most commonly performed by a current of electrons accelerated through about 70–80 V, but many other methods are used. The dispersive unit may consist of an electric field, or sequential electric and magnetic fields, or a number of parallel rods upon which a combination of dc and ac voltages are applied; these provide either temporal or spatial dispersion of ions according to their mass-to-charge ratio. *See* MASS SPECTROM-ETRY; SECONDARY ION MASS SPECTROMETRY (SIMS); TIME-OF-FLIGHT SPECTROMETERS.

Multiplex or frequency-modulated spectroscopy. The basis of this family of techniques is to encode or modulate each optical wavelength exiting the spectrometer output with an audio frequency that contains the optical wavelength information. Use of a wavelength analyzer then allows recovery of the original optical spectrum. The primary advantage of such a system is that it enables the detector to sense all optical wavelengths simultaneously, resulting in greatly decreased scan times and the possibility for signal averaging. Frequency modulation can be achieved, for example, by means of a special encoding mask as in Hadamard transform spectroscopy or by use of a Michelson interferometer as in Fourier transform spectroscopy.

Raman spectroscopy. Consider the passage of a beam of light, of a wavelength not corresponding to an absorption, through a sample. A small fraction of the light is scattered by the molecules, and so exits



Fig. 6. Interactions of a molecule with a photon of energy ΔE which is not absorbed. (a) Rayleigh scattering. (b) Raman scattering, producing a Stokes line. (c) Raman scattering, producing an anti-Stokes line.

the sample at a different angle. This is known as Rayleigh scattering if the wavelength of the scattered light is the same as the original wavelength; but if the wavelength is different, it is called Raman scattering (Fig. 6). The fraction of photons thus scattered by the Raman effect is even smaller than by Rayleigh scattering. Differences in wavelength correspond to the wavelengths of certain vibrational and rotational processes. If the scattered light has a longer wavelength, the new line in the spectrum is known as a Stokes line; if shorter, an anti-Stokes line. The information obtained is of use in structural chemistry, because the rules allowing and forbidding transitions are different from those in absorption spectroscopy. Use of lasers for sources has revived this technique. A related process, resonance Raman spectroscopy, makes use of the fact that Raman probabilities are greatly increased when the exciting radiation has an energy which approaches the energy of an allowed electronic absorption. Raman scattering is also enhanced by surfaces, providing a valuable tool for surface analysis. See RAMAN EFFECT.

X-ray spectroscopy. The excitation of inner electrons in atoms is manifested as x-ray absorption; emission of a photon as an electron falls from a higher level into the vacancy thus created is x-ray fluorescence. The techniques are used for chemical analysis. Extended x-ray absorption fine structure, obtained under conditions of high resolution, shows changes in the energy of inner electrons due to the presence of neighboring atoms, and can be used in structural chemistry for the determination of the nature and location of neighbor atoms nearest to that being studied. Generation of x-rays by bombarding a sample with a proton beam gives the microanalytical technique of proton-induced x-ray emission. See X-RAY FLUORESCENCE ANALYSIS; X-RAY SPECTROM-ETRY. Maurice M. Bursey

Bibliography. J. D. Ingle and S. R. Crouch, *Spectro-chemical Analysis*, 1988; S. P. Parker (ed.), *Spectroscopy Source Book*, 1988; F. A. Settle et al., *Instrumental Methods of Analysis*, 8th ed., 2007; *Spectroscopy* (trade journal), monthly; D. H. Williams and I. Flemming, *Spectroscopic Methods in Organic Chemistry*, 5th ed., 1995.

Spectrum

The term spectrum is applied to any class of similar entities or properties strictly arrayed in order of increasing or decreasing magnitude. In general, a spectrum is a display or plot of intensity of radiation (particles, photons, or acoustic radiation) as a function of mass, momentum, wavelength, frequency, or some other related quantity. For example, a beta-particle spectrum represents the distribution in energy or momentum of negative electrons emitted spontaneously by certain radioactive nuclides, and when radionuclides emit alpha particles, they produce an alpha-particle spectrum of one or more characteristic energies. A mass spectrum is produced when charged particles (ionized atoms or molecules) are passed through a mass spectrograph in which electric and magnetic fields deflect the particles according to their charge-to-mass ratios. The distribution of sound-wave energy over a given range of frequencies is also called a spectrum. See MASS SPECTROSCOPE; SOUND.

In the domain of electromagnetic radiation, a spectrum is a series of radiant energies arranged in order of wavelength or of frequency. The entire range of frequencies is subdivided into wide intervals in which the waves have some common characteristic of generation or detection, such as the radio-frequency spectrum, infrared spectrum, visible spectrum, ultraviolet spectrum, and x-ray spectrum. Spectra are also classified according to their origin or mechanism of excitation, as emission, absorption, continuous, line, and band spectra. *See* ELECTROMAGNETIC RADIATION.

radiations from an excited light source are dispersed. Excitation of emission spectra may be by thermal energy, by impacting electrons and ions. or by absorption of photons. Depending upon the nature of the light source, an emission spectrum may be a continuous or a discontinuous spectrum, and in the latter case, it may show a line spectrum, a band spectrum, or both.

An absorption spectrum is produced against a background of continuous radiation by interposing matter that reduces the intensity of radiation at certain wavelengths or spectral regions. The energies removed from the continuous spectrum by the interposed absorbing medium are precisely those that would be emitted by the medium if properly excited. This reciprocity of absorption and emission is known as Kirchhoff's principle; it explains, for example, the absorption spectrum of the Sun, in which thousands of lines of gaseous elements appear dark against the continuous-spectrum background.

A continuous spectrum contains an unbroken sequence of waves or frequencies over a long range (see **illus.**).

In illus. *a* the spectra for 1000 and 2000°C (1832 and 3632° F) were obtained from tungsten filament, that for 4000°C (7232°F) from a positive pole of carbon arc. The upper spectrum in illus. *b* is that of the source alone, extending roughly from 400 to 650 nanometers. The others show the effect on this spectrum of interposing three kinds of colored glass.

All incandescent solids, liquids, and compressed gases emit continuous spectra, for example, an incandescent lamp filament or a hot furnace. In general, continuous spectra are produced by high temperatures, and under specified conditions the



Spectra of a blackbody. (a) Continuous emission spectra of a solid, taken with a quartz spectrograph. (b) Continuous absorption spectra of the source alone and three kinds of colored glass interposed on it. (From F. A. Jenkins and H. E. White, Fundamentals of Optics, 4th ed., McGraw-Hill, 1976)

An emission spectrum is produced whenever the

distribution of energy as a function of temperature and wavelength is expressed by Planck's law. *See* HEAT RADIATION; PLANCK'S RADIATION LAW.

Line spectra are discontinuous spectra characteristic of excited atoms and ions, whereas band spectra are characteristic of molecular gases or chemical compounds. *See* ATOMIC STRUCTURE AND SPECTRA; BAND SPECTRUM; LINE SPECTRUM; MOLECULAR STRUC-TURE AND SPECTRA; SPECTROSCOPY.

W. F. Meggers; W. W. Watson

Spectrum analyzer

An instrument for the analysis and measurement of signals throughout the electromagnetic spectrum. Spectrum analyzers are available for subaudio, audio, and radio-frequency measurements, as well as for microwave and optical signal measurements.

Generally, a spectrum analyzer separates the signal into two components: amplitude (displayed vertically) and frequency (displayed horizontally). On some low-frequency analyzers, phase information can also be displayed. Low-frequency analyzers are sometimes grouped under the heading "harmonic analyzers," although this term is becoming less common.

Operation. On a conventional spectrum analyzer, a screen with a calibrated graticule displays the components of the input signal. The vertical scale displays the amplitude of each component, and the chosen frequency band is displayed horizontally. Components of the signal being analyzed are displayed as vertical lines whose height is proportional to amplitude and whose horizontal displacement equates to frequency. Originally, cathode-ray tubes were used for the display; solid-state displays such as liquid-crystal displays now are used. *See* CATHODE-RAY TUBE; ELECTRONIC DISPLAY.

Swept heterodyne technique. The traditional audiofrequency, radio-frequency, and microwave spectrum analyzer is essentially a swept heterodyne analyzer (Fig. 1). A ramp generator simultaneously increases the frequency of a local oscillator and moves the spot across the display. When the ramp voltage is zero, the spot will be on the left-hand side of the display and the local oscillator will be at a low frequency. Any signal at the input, at a frequency such that the difference between its frequency and the local oscillator is within the bandwidth of an intermediate-frequency filter, will be detected and will vertically deflect the spot on the display by an amount proportional to the amplitude of the input signal being analyzed. As the ramp voltage increases, the spot moves across the display and the analyzer tunes to increasingly higher-frequency signals. Each component of a complex waveform is thus isolated and displayed. See HETERODYNE PRINCIPLE.

Digital techniques. Traditionally, audio-frequency spectrum analyzers used a lower-frequency version of the swept heterodyne technique. Now, lower-frequency instruments generally use a digital



Fig. 1. Simplified block diagram of a spectrum analyzer with a tracking generator.

techniques known as the fast Fourier transform (FFT). The signal to be analyzed is converted to a digital signal by using an analog-to-digital converter, and the digital signal is processed by using the FFT algorithm. The input to this algorithm consists of blocks of digitized data. The algorithm analyzes the time-domain waveform, computes the frequency components present, and displays the results. All frequency components are analyzed simultaneously, in contrast to a swept instrument, which analyzes only a narrow portion of the spectrum at any one time. *See* ANALOG-TO-DIGITAL CONVERTER.

The FFT technique is much faster and the measurement is virtually real time, but it is restricted to lower frequencies. Real-time display is vital to analyze complex nonrepetitive signals such as speech or random vibrations. Digital processing techniques are used in radio-frequency and microwave instruments, but an analog input is required to down-convert the high-frequency signals to a signal low enough in frequency to be processed by currently available analogto-digital converters. *See* FOURIER SERIES AND TRANS-FORMS.

Controls. Modern spectrum analyzers incorporate sophisticated microprocessor control and can be remotely programmed (operated), so that a computer can control all functions and read the measured values. There are normally four basic groups of controls: frequency, horizontal scale, vertical scale, and markers. Both digital keypads and rotary controls are generally provided. The frequency controls determine the center or reference frequency of the display. The horizontal-scale controls determine the portion of the spectrum that can be viewed; this can be a scan from, say, 100 Hz up to 20 GHz or beyond. Separate controls determine the vertical scale and the sensitivity. Markers can be placed on spectral lines to measure accurately absolute and relative frequencies and levels. See MICROPROCESSOR.

Further controls allow the operator to optimize operation for different conditions; for example, the intermediate-frequency bandwidth can be changed to alter the resolution in order to resolve closely spaced signals. The input attenuation and gain of the intermediate-frequency section can be separately changed to optimize for either low-noise or lowintermodulation operation.

Digital storage. In order to obtain high resolution over a relatively wide band of frequencies, it is necessary to sweep slowly, especially in radio-frequency and microwave instruments. A single sweep may take as long as 100 seconds. Digital storage gives the effect of a constant display, even though a very slow sweep may have been used to acquire the displayed data.

Tracking generator. The tracking generator enhances the applications of spectrum analyzers (Fig. 1). Its output delivers a swept signal whose instantaneous frequency is always equal to the input tuned frequency of the analyzer. The swept signal is used to measure the frequency response of circuits such as filters and amplifiers over a wide dynamic range, which would not be possible with a broadband nonselective measurement method.

Harmonic analyzers. Spectrum analyzers covering up to typically 100 kHz can also be called harmonic analyzers. This term refers to instruments such as wave analyzers and distortion factor meters. Such instruments are still available, but the FFT spectrum analyzer has largely supplanted them, assisted with digital techniques using microprocessors. Faster and more accurate measurements are generally obtained using modern digital instrumentation.

Other harmonic analyzers that may be encountered include the tunable filter analyzer, stepped filter analyzer, parallel filter analyzer, and time compression analyzer.

Applications. Early radio-frequency and microwave analyzers were developed to measure the performance of microwave radar transmitters and to analyze signals from single-sideband transmitters. *See* RADAR; SINGLE SIDEBAND.



Fig. 2. Typical radio-frequency spectrum analyzer. (IFR Ltd.)

A typical use for radio-frequency and microwave spectrum analyzers is the measurement of spurious radiation (noise) from electrical machinery and circuits, known as radio-frequency interference (RFI). Other uses include monitoring and surveillance to detect unauthorized or unintended transmissions, such as the civil monitoring of broadcast and communication channels and the detection of electronic warfare signals. Another application is the analysis of radio communication transmitters and receivers, including those used in radio and television broadcasting, satellite systems, and mobile radio and cellular telephone communications (**Fig. 2**). *See* ELECTRICAL INTERFERENCE; ELECTRONIC WARFARE.

Spectrum analyzers also fulfill a variety of generalpurpose uses in research, development, manufacture, and field maintenance of electrical and electronic equipment. For certain applications they are used instead of oscilloscopes, since the spectrum analysis display of a complex waveform can provide more information than an oscilloscope display. Oscilloscopes are typically limited to frequencies up to 1 GHz, while spectrum analyzers can measure microwave signals over 300 GHz. *See* OSCILLOSCOPE.

Low-frequency spectrum analyzers are used in a variety of applications. The most obvious use is the measurement of distortion and unwanted signals in all types of audio equipment, from recording and broadcast studios to amplifiers used in the home. *See* SOUND RECORDING; SOUND-REPRODUCING SYSTEMS.

Further uses include the analysis of speech waveforms, measurement of vibration and resonances in mechanical equipment and structures, determination of echo delays in seismic signals, investigation of noise such as from aircraft engines or from machinery in factories, analysis of sonar signals used to detect objects underwater, and study of ultrasonic waves to determine the internal structure of objects such as human tissue and metal castings. *See* BIOMED-ICAL ULTRASONICS; MECHANICAL VIBRATION; NOISE MEASUREMENT; NONDESTRUCTIVE EVALUATION; SEIS-MOLOGY; SONAR; ULTRASONICS.

Optical spectrum analyzers use techniques such as a collimating mirror and a diffraction grating or a Michelson interferometer to separate out the lightwave components. They are used for a variety of applications, including measurements on lasers and light-emitting diodes, and for the analysis of opticalfiber equipment used to carry multichannel, digital telephony. *See* DIFFRACTION GRATING; INTERFEROM-ETRY; LASER; LIGHT-EMITTING DIODE; OPTICAL COM-MUNICATIONS; OPTICAL FIBERS. S. J. Gledhill

Bibliography. J. S. Bendat and A. G. Piersol, *Engineering Applications of Correlation and Spectrum Analysis*, 2d ed., 1993; M. Engelson, *Modern Spectrum Analyzer Theory and Applications*, 1984; S. Goldman, *Vibration Spectrum Analysis: A Practical Approach*, 2d ed., 1999; A. D. Helfrick, *Electrical Spectrum and Network Analyzers*, 1991; B. Kibble et al., *A Guide to Measuring Direct and Alternating Current and Voltage Below 1 MHz*, Institute of Measurement and Control, 2003; T. S. Laverghetta,

Modern Microwave Measurements and Techniques, 1988; R. A. Witte, Electronic Test Instruments: Analog and Digital Measurements, 2d ed., 2002; R. A. Witte, Spectrum and Network Measurements, 2001.

Speech

A set of audible sounds produced by disturbing the air through the integrated movements of certain groups of anatomical structures. Humans attach symbolic values to these sounds for communication. There are many approaches to the study of speech.

Speech Production

The physiology of speech production may be described in terms of respiration, phonation, and articulation. These interacting processes are activated, coordinated, and monitored by acoustical and kinesthetic feedback through the nervous system.

Respiration. Most of the speech sounds of the major languages of the world are formed during exhalation. Consequently, during speech the period of exhalation is generally much longer than that of inhalation. In providing a sublaryngeal air supply, the respiratory muscles coordinate with the laryngeal and supralaryngeal muscles to produce suitable driving pressures for the formation of many speech sounds (**Fig. 1**). The aerodynamics of the breath stream influence the rate and mode of the vibration of the vocal folds. This involves interactions between the pressures initiated by thoracic movements and the position and tension of the vocal folds. *See* RES-PIRATION.

The pressure pattern of the sublaryngeal air is closely related to the loudness of the voice and appears to be correlated with the perception of stress. For example, the word "permit" may be either a noun or a verb, depending on the placement of stress. Various attempts have been made to correlate units such as the syllable and the phrase to the contractions of specific respiratory muscles. However, the relation between the thoracic movements and the grouping of speech sounds is poorly understood.

Experimental studies on respiration for speech production employ techniques such as pressure and electromyographic recording at various points along the respiratory tract as well as x-ray photography to investigate anatomical movements.

Phonation. The phonatory and articulatory mechanisms of speech may be regarded as an acoustical system whose properties are comparable to those of a tube of varying cross-sectional dimensions. At the lower end of the tube, or the vocal tract, is the larynx. It is situated directly above the trachea and is composed of a group of cartilages, tissues, and muscles. The upper end of the vocal tract may terminate at the lips, at the nose, or both. The length of the vocal tract averages 6.5 in. (16 cm) in men and may be increased by either pursing the lips or lowering the larynx.



Fig. 1. Human head and thorax. (Courtesy of G. E. Peterson)

Mechanism of phonation. The larynx is the primary mechanism for phonation, that is, the generation of the glottal tone. The vocal folds consist of connective tissue and muscular fibers which attach anteriorly to the thyroid cartilage and posteriorly to the vocal processes of the arytenoid cartilages. The vibrating edge of the vocal folds measures about 0.92-1.08 in. (23-27 mm) in men and considerably less in women. The aperture between the vocal folds is known as the glottis (Fig. 2). The tension and position of the vocal folds are adjusted by the intrinsic laryngeal muscles, primarily through movement of the two arytenoid cartilages. By contraction of groups of muscles, the arytenoid cartilages may be pulled either anteriorposteriorly or laterally in opposite directions. They may also be rotated about the vertical axis by the cricoarytenoid muscles. During whispering and unvoiced sounds, the glottis assumes the shape of a triangle, with the apex directly behind the thyroid cartilage. See LARYNX.



Fig. 2. Action of the glottis during speech. (a) Six successive positions of the vocal folds in one vibration cycle at 120 Hz. (b) Relation between pitch and vocal fold tension. (Courtesy of D. W. Farnsworth)

Air pressure. When the vocal folds are brought together and there is a balanced air pressure to drive them, they vibrate laterally in opposite directions. During phonation, the vocal folds do not transmit the major portion of the energy to the air. They control the energy by regulating the frequency and amount of air passing through the glottis. Their rate and mode of opening and closing are dependent upon the position and tension of the folds and the pressure and velocity of airflow. The tones are produced by the recurrent puffs of air passing through the glottis and striking into the supralaryngeal cavities.

As the air passes through the glottis with increasing velocity, the pressure perpendicular to the direction of airflow is reduced. This allows the tension of the folds to draw them together, either partly or completely, until sufficient pressure is built up to drive them apart again. It is generally assumed that the volume flow is linearly related to the area of the glottis and that the glottal source has a waveform that is approximately triangular.

Voiced sound. Speech sounds produced during phonation are called voiced. Almost all of the vowel sounds of the major languages and some of the consonants are voiced. In English, voiced consonants may be illustrated by the initial and final sounds in the following words: "bathe," "dog," "man," "jail." The speech sounds produced when the vocal folds are apart and are not vibrating are called unvoiced; examples are the consonants in the words "hat," "cap," "sash," "faith." During whispering all the sounds are unvoiced.

Frequency and pitch. The rate of vibration of the vocal folds is the fundamental frequency of the voice (F0). It correlates well with the perception of pitch. The frequency increases when the vocal folds are made taut (Fig. 2b). Relative differences in the fundamental frequency of the voice are utilized in all languages to signal some aspects of linguistic information. In Fig. 3, the fundamental frequency of the voice for the phrase "speech production" is shown as a dotted line. Note that frequency is highest on the vowels "ee" and "u," indicating that these are stressed. Also, the fundamental frequency drops to the lowest point

at the end of the phrase, indicating the phrase is a statement rather than a question.

Many languages of the world are known as tone languages, because they use the fundamental frequency of the voice to distinguish between words. Chinese is a classic example of a tone language. There are four distinct tones in Chinese speech. Said with a falling fundamental frequency of the voice, ma means "to scold." Said with a rising fundamental frequency, it means "hemp." With a level fundamental frequency it means "mother," and with a dipping fundamental frequency it means "horse." In Chinese, changing a tone has the same kind of effect on the meaning of a word as changing a vowel or consonant in a language such as English.

The mode of vibration of the vocal folds can be varied to change the acoustical properties of the voice. For example, if the folds remain sufficiently apart during the vibration, the voice is perceived to be breathy in quality.

Study methods. The properties of the movements of the vocal folds have been studied by placing microphones at different points in relation to the larynx. Experiments have also been conducted with electrodes inserted directly into the laryngeal structures of hemilaryngectomized subjects and simultaneously recording their speech. Much information about the glottal vibration has been gathered through direct observation of the vocal folds by means of x-ray photography, the stroboscope, and high-speed motionpicture photography. The relation between the volume flow, the mode of glottal vibration, and the spectrum of the glottal tone has been studied by these techniques. Several electronic and mechanical models have been constructed to simulate phonation, based on the collated evidence from the acoustics and physiology of the larynx.

Articulation. The activity of the structures above and including the larynx in forming speech sound is known as articulation. It involves some muscles of the pharynx, palate, tongue, and face and of mastication (Fig. 1).

The primary types of speech sounds of the major languages may be classified as vowels, nasals,



Fig. 3. Sound spectrograms of the utterance "speech production." (a) Energy patterns: A, pulse; B, periodic; C, aperiodic; and D, silence. (b) The dotted lines represent the pitch pattern of the utterance.

plosives, and fricatives. They may be described in terms of degree and place of constriction along the vocal tract. *See* PHONETICS.

Vowels. The only source of excitation for vowels is at the glottis. During vowel production the vocal tract is relatively open and the air flows over the center of the tongue, causing a minimum of turbulence. The phonetic value of the vowel is determined by the resonances of the vocal tract, which are in turn determined by the shape and position of the tongue and lips.

The point of constriction of a vowel is where the cross-sectional area of the vocal tract is minimized by the humping of the tongue. At this point the vocal tract is divided approximately into an oral cavity in front and a pharyngeal cavity behind. Vowels are known grossly as back, central, and front as the point of constriction moves from the pharyngeal wall anteriorly along the palate, as in the words "boot" (back), "but" (central), "beet" (front). Vowels are referred to as high, mid, and low as the height of the tongue hump is increasingly lowered, as in the words "beet" (high), "bet" (mid), "bat" (low). As the two parameters, point of constriction and tongue height, are varied, the relations between the oral and pharyngeal cavities change to produce characteristic resonances for different vowels.

Vowels are grossly described as rounded when produced with the contraction of the muscles of the lips, as in "he." Since rounding lengthens the vocal tract, the resonances of the rounded vowel are generally lower than those of the corresponding unrounded vowel.

Nasals. The nasal cavities can be coupled onto the resonance system of the vocal tract by lowering the velum and permitting airflow through the nose. Vowels produced with the addition of nasal resonances are known as nasalized vowels. Nasalization may be used to distinguish meanings of words made up of otherwise identical sounds, such as *bas* and *banc* in French. If the oral passage is completely constricted and air flows only through the nose, the resulting sounds are nasal consonants. The three nasal consonants in "meaning" are formed with the constriction successively at the lips, the hard palate, and the soft palate.

Plosives. These are characterized by the complete interception of airflow at one or more places along the vocal tract. The pressure which is built up behind the intercepting mechanism may not be immediately released or may be released through the oral or nasal orifice. The places of constriction and the manner of the release are the primary determinants of the phonetic properties of the plosives. The words "par," "bar," "tar," and "car" begin with plosives.

When the interception is brief and the constriction is not necessarily complete, the sound is classified as a flap. By tensing the articulatory mechanism in proper relation to the airflow, it is possible to set the mechanism into vibrations which quasi-periodically



Fig. 4. Computer-generated three-dimensional spectral display of the spoken syllable "yah."

intercept the airflow. These sounds are called trills and can be executed with the velum, the tongue, and the lips. They usually are produced around 25 Hz, whereas the fundamental frequency of the male speaking voice ranges from 80 to 160 Hz.

Fricatives. These are produced by a partial constriction along the vocal tract which results in turbulence. Their properties are determined by the place or places of constriction and the shape of the modifying cavities. The fricatives in English may be illustrated by the initial and final consonants in the words "vase," "this," "faith," "hash."

Acoustical analysis. The muscular activities of speech production influence each other both simultaneously and with respect to time. It is frequently difficult to segment a sequence of sounds because the physiological activities which produce them form a continuum. Consequently the physical aspects of a sound type are determined to some extent by neighboring sounds. Each physiological parameter-for example, lip rounding and fundamental frequency of the glottal tone-may be varied continuously. Therefore, within its physiological limits the speech mechanism is able to produce a continuum of different sounds. The physiological and acoustical criteria in a classification of these sounds are therefore dependent on external conditions such as the threshold of the analyzing instruments.

There is a complex, many-to-one relation between the physiology of production and the resultant acoustical waves. Acoustically, speech sounds may be regarded as the simultaneous and sequential combinations of pulse, periodic, and aperiodic forms of energy interrupted by silence of varying duration. These energy patterns are labeled as A, B, C, and D, respectively, in the sound spectrogram of Fig. 3a. The horizontal bands in the periodic portions indicate the frequency positions of the vocal tract resonances. On the average, there is one resonance per kilohertz for vowels produced by a male vocal tract, though these resonances are differently spaced for various vowels. The pitch of the utterance is illustrated in the sound spectrograms of Fig. 3b. Each horizontal line indicates a harmonic of the glottal tone. The spectrum of the glottal tone is usually taken to have a slope of -6 to -12 decibels per octave when the effects of the vocal tract resonances are discounted. The sound spectrograph has been a valuable instrument for research on the acoustical aspect of speech. Essentially, it makes a Fourier type of analysis on the acoustical wave. The results are then translated into graphic form, usually with frequency along the ordinate and time along the abscissa.

Advances in computer technology have also been applied to the analysis of speech. **Figure 4** shows a three-dimensional display of the syllable "yah." In this display, the amplitude of each vocal tract resonance is indicated by the height of the peak. Thus it can be clearly seen that as the syllable progresses in time, the resonances change not only in frequency but in amplitude as well.

Novel methods have been introduced from signalprocessing technology, such as the so-called speech flakes shown in **Fig. 5**. The flakes are of the vowels as in "far," "food," and "fee." These three vowels are by far those most often used in the languages of the world. The redundant symmetry in such flakes may turn out to be quite helpful for visualizing speech sounds.

Neurology. The ability to produce meaningful speech is dependent in part upon the association areas of the brain. It is through them that the stimuli which enter the brain are interrelated. These areas are connected to motor areas of the brain which send fibers to the motor nuclei of the cranial nerves and hence to the muscles. Three neural pathways are directly concerned with speech production, the pyramidal tract, the extrapyramidal, and the cerebellar motor paths. It is the combined control of these pathways upon nerves arising in the medulla and ending



Fig. 5. Speech flakes of vowels in the words (a) "far," (b) "food," and (c) "fee." (Courtesy of Charles Wooters)



Fig. 6. Regions of the brain involved in speech. (a) Left hemisphere, usually more involved in speech functions than the right hemisphere. (b) A model for the neural processes involved in speaking a heard word. Arrows indicate the path of the relayed signal from the primary auditory area to the motor cortex.

in the muscles of the tongue, lips, and larynx which permits the production of speech. *See* NERVOUS SYSTEM (VERTEBRATE).

The part of the pyramidal tract which is most important for speech has its origin in the lower portion of the precentral gyrus of the cerebral cortex. In proximity to the precentral gyrus on the left side of the brain is Broca's area. This area is one of several which are believed to activate the fibers of the precentral gyrus concerned with movements necessary for speech production.

In most people, the left hemisphere of the brain is more involved in speech functions than the right hemisphere. **Figure** 6a shows the regions on the surface of the left hemisphere that are especially pertinent for speech. Figure 6b shows a possible model for speaking a heard word. First the signal is received at the primary auditory area. From there the signal is relayed to Wernicke's area, situated posteriorly in the temporal lobe, where it is recognized as a word. The heard word is then relayed to Broca's area, situated inferiorly in the frontal lobe, via bundles of nerve fibers collectively known as arcuate fasciculus. Finally it is relayed to the motor cortex, which ultimately structures the muscular activities that produce the sounds, that is, speak the word.

Areas of the extrapyramidal tract which contribute to speech are the caudate nucleus, globus pallidus, and the thalamus. Some of the fibers from these centers also go to the same motor nerves as those of the pyramidal tract. The chief function of these paths in regard to speech is their regulatory and refining actions. It is through the influence of the extrapyramidal system that the regulation and inhibition of opposing sets of muscles are controlled. The cerebellar motor paths also contribute to the coordination and tonus of muscle structures requiring movements of paired muscles. In addition, the cerebellar motor paths are important in coordinating breathing for speech production.

Six of the 12 cranial nerves send motor fibers to the muscles that are involved in the production of speech.

These nerves are the trigeminal, facial, glossopha-

ryngeal, vagus, spinal accessory, and the hypoglossal. They represent the link between the neural activity which begins in the cerebral cortex and the coordinated muscular movements which produce speech. The relations between the neurology, physiology, and acoustics of speech are extremely complex and poorly understood. Within normal anatomical variation, the skill in producing particular speech sounds is almost entirely determined by the early linguistic environment of the speaker. Through the interactions between the linguistic environment and the speech of the individual, the sounds and their symbolic values change from community to community and through time. *See* PSYCHOACOUSTICS; PSY-CHOLINGUISTICS.

Development

In the early stages of speech development the child's vocalizations are quite random. The control and voluntary production of speech are dependent upon physical maturation and learning.

It is possible to describe the development of speech in five stages. In the first stage the child makes cries in response to stimuli. These responses are not voluntary but are part of the total bodily expression. The second stage begins between the sixth and seventh week. The child is now aware of the sounds he or she is making and appears to enjoy this activity. During the third stage the child begins to repeat sounds heard coming from himself or herself. This is the first time that the child begins to link speech production to hearing. During the ninth or tenth month the child enters the fourth stage and begins to imitate without comprehension the sounds that others make. The last stage begins between the twelfth and eighteenth month, with the child intentionally employing conventional sound patterns in a meaningful way. The exact time at which each stage may occur varies greatly from child to child.

Although most children begin to use speech meaningfully by the eighteenth month, they are not able to articulate all the necessary sounds. The first sounds the child uses are mostly front vowels. The back vowels become more frequent as the child develops. The ability to produce vowels correctly seems to be almost completely learned by 30 months. The earliest consonants the child uses are those formed with both lips. By 54 months, the child can produce all the necessary consonants. The approximate age level at which each of 23 consonant sounds of English is mastered by American children follows.

Age,	Sounds		
monuns			
$3^{1}/_{2}$	b(baby), p(papa), m(mama), w(wet), h(be)		
$4^{1}/_{2}$	d(<i>dada</i>), t(<i>t</i> wo), n(<i>n</i> ose), g(<i>g</i> 0), k(<i>cat</i>), ng(sing), y(yet)		
$5^{1}/_{2}$	f(<i>f</i> un)		
6 ¹ / ₂	v(very), th(<i>th</i> at), z(azure), sh(<i>sh</i> oe), l(<i>l</i> ie)		
$7^{1}/_{2}$	s(see), $z(zoo)$, $r(rain)$, $th(thin)$, wh(who)		

The close coupling of age and sound acquisition, as exhibited above, together with other observations, has led some researchers to post innate, genetic mechanisms which serve exclusively speech. This innate hypothesis is a focal point for discussion on how speech emerged, in the child as well as in the human species.

Speech Technology

Since about 1990, technologies have arisen that are taking over an ever-expanding share of the marketplace. These technologies are based on the knowledge that has accumulated from many decades of research on how speech is produced and perceived. Speech technology has been developing within three areas.

One area has to do with identifying a speaker by analyzing a speech sample. Since the idea is analogous to that of identifying an individual by fingerprint analysis, the technique has been called voice print. However, fingerprints have two important advantages over voice prints: (1) they are based on extensive data that have accumulated over several decades of use internationally, whereas no comparable reference exists for voice prints; and (2) it is much easier to alter the characteristics of speech than of finger prints. Consequently, this area has remained largely dormant. Most courts in the United States, for instance, do not admit voice prints as legal evidence. Some progress may be made in voice print identification after a better understanding of the process of phonation is achieved, since this is more specific to individuals.

In contrast, the two other areas of speech technology, synthesis and recognition, have seen explosive growth. Humans prefer to interact with speech, which they have mastered since early childhood, rather than via keyboard and screen. This preference becomes a necessity when visibility is poor, or when the hands are otherwise occupied. With the rapid pace in computer advancement, speech technology is expected to extend into more and more sectors of human life in the years to come.

In many applications where a limited repertoire of speech is required, computer-synthesized speech is used instead of human speakers. These applications began with automobile devices which remind the driver to fasten his seat belt, and telephone directory services which tell customers the numbers they seek. The area has advanced greatly. Coupled with a machine that can read the printed page, the conversion of text to speech is almost a reality. A common technology currently used in speech synthesis involves an inventory of pitch-synchronized, prestored human speech. These prestored patterns are selected according to the particular requirements of the application and recombined with some overlap into the desired sentence by computer, almost in real time.

The quality of synthesized speech for English is remarkably good, though it is limited at present to neutral, emotionless speech. Many other languages are being synthesized with varying degrees of success. Much research will be required to endow synthesized speech with the rich panorama of emotion that human speech can convey.

The recognition of speech by computer is much more difficult than synthesis. Instead of just reproducing the acoustic wave, the computer must understand something of the semantic message that the speech wave contains, in order to recognize pieces of the wave as words in the language. Humans do this easily because they have a great deal of background knowledge about the world, because they are helped by contextual clues not in the speech wave, and because they are extensively trained in the use of speech. Although there are various systems in the marketplace at present which purport to perform dictation, their levels of performance are typically quite low. This is especially true when the speech takes place in noisy conditions.

Nonetheless, given various constraints, some of the existing systems do remarkably well—even to the extent of helping surgeons in the operating room, where reliability is of paramount concern. These constraints include (1) stable acoustic conditions in which speech is produced, (2) a speaker trained by the system, (3) limited inventory of utterances, and (4) short utterances. The research here is strongly driven by the marketplace, since all sorts of applications can be imagined where spoken commands are required or highly useful. It is expected that speech technology will continue to grow at a fast pace in the coming decades. *See* SPEECH DISORDERS.

William S.-Y. Wang

Bibliography. G. Fant, Speech Sounds and Features, 1973; O. Fujimura (ed.), Vocal Fold Physiology, vol. 2 of Vocal Physiology: Voice Production, Mechanisms and Functions, 1988; J.-C. Jungva and J.-P. Haton, Robustness in Automatic Speech Recognition: Fundamentals and Applications, 1996; P. Lieberman, Speech Physiology and Acoustic Phonetics, 1977; W. S.-Y. Wang, The Basis of Speech, 1989; W. S.-Y. Wang (ed.), The Emergence of Language: Development and Evolution, 1991.

Speech disorders

Typically, problems of speech production, articulation, or voice, or disruptions of the ability to use language. These disorders can be categorized in a variety of ways, but most belong to one of the following types: articulation, voice, fluency, and language.

Another way to categorize these disorders is on the basis of the cause, organic or nonorganic. Organic speech and language disorders typically result from disease or injury to the structures, muscles, and nerves necessary to move and control the mechanisms for speech and voice production. Damage to the brain, particularly the left hemisphere, as the result of stroke, dementia, or traumatic head injury can result in disruptions of normal language function. In children, failure of the brain or systems subserving language and speech to develop adequately can result in significant developmental delays in the acquisition of appropriate language and speech functions. Disorders of language and speech for which there is no demonstrable organic etiology are often the result of poor learning or inadequate speech and language models. In some instances, there may be a psychogenic basis for the disorder.

Articulation. Articulation for speech production involves the appropriate sequential movement of structures known as the articulators: mandible, tongue, lips, velum, and so forth. There is a wide range of acceptable productions of a given sound (phoneme) in the language. The variations, sometimes known as dialect, reflect regional, ethnic, and socioeconomic characteristics; they should not be considered as deviant productions. In most of these instances, production is precise and falls within the accepted bounds of the individual's speech and language community. However, when the production of one or more sounds in the language is imprecise and cannot be understood, then an articulation disorder is present. Disorders can range from mild problems such as the distortions associated with a lisp, or can be so severe as to render the speaker unintelligible. Most articulation problems can be categorized in terms of how the sounds are produced. Speech sounds can be misarticulated by distortion, omission, substitution, or addition. These problems can arise from organic causes, such as impairment to motor control of the speech musculature following a stroke or cerebral palsy (dysarthria), cleft palate or lip, or hearing loss. Many articulation problems have no discernible organic basis; these are labeled as functional, since the etiology of the problem is unclear. Many of these problems begin in childhood and can persist into adult life if not treated appropriately. See CLEFT LIP AND CLEFT PALATE; HEARING IMPAIRMENT; PHONETICS.

There are a number of approaches for treating these disorders, both organic and functional. For example, discrimination training is based on the concept that the basis of the problem is poor sound discrimination and the inability to match the auditory feedback from production with the patterns that others are able to produce. Treatment focuses on training auditory discrimination to discover the differences and then modifying production to match the pattern. Other approaches include behavior modification; production training, which requires relearning specific motor patterns; linguistic approaches based on theories of phonological structure; and psychological approaches, including the psychoanalytic model. The psychological approach has not been widely accepted; however, psychological counseling can sometimes be of great help to children who have not been accepted because of their poor articulatory behaviors. Finally, in cases of very severely impaired persons, such as those with cerebral palsy and accompanying dysarthria, communication boards and alternative communicative methods, including computers using synthetic speech, have become increasingly common. In persons with orofacial abnormalities, such as cleft lip or palate, that interfere with articulation, surgical and dental repair or protheses may be necessary before speech therapy can be initiated. See PSYCHOLINGUISTICS.

Production of speech sounds is a complex process that requires precise programming and direction by the brain. When there is damage to those brain regions controlling articulatory programming because of stroke or trauma, a disorder known as apraxia may ensue. This is in contrast to dysarthria, where articulation is distorted because of paralysis or paresis of the musculature necessary for speech production. *See* APRAXIA; BRAIN.

Voice. Voice disorders occur when the phonatory mechanism is not functioning properly. Some impairments of function can be the result of vocal abuse or misuse. These voice disorders are said to be functional, although prolonged abuse or misuse can result in temporary or permanent damage to the phonatory mechanism. There are, however, voice disorders that have a true organic etiology. These disorders result from disease or some other condition that affects the laryngeal, oral, or respiratory structures, which are involved in the production of voice. Phonation problems associated with cleft palate, which is a structural defect in the palate resulting in hypernasality, are not included in functional or organic disorders. Aphonia is a complete absence of voice. More commonly, individuals with voice disorder are dysphonic, retaining some voice quality. Whether functional or organic in origin, voice disorders are described in terms of pitch, quality (resonance), or loudness. Instruments have become available that permit sensitive and objective measurement of vocal characteristics. These instruments are of value in both diagnosis and treatment of voice disorders.

Pitch. Disorders of voice in which pitch is the predominant factor occur when the speaker's voice pitch is inappropriate for age and gender. In these instances, the pitch is described as too high, too low, monotonous, or demonstrating repeated inappropriate pitch patterns. A voice that is pitched too high is often the result of psychological tension. The muscles of the larynx are contracted in such a way as to raise pitch significantly beyond its normal range. It may also result when the larynx is small for age and gender and the vocal folds are thin and short. In some instances, the speaker's inability to perceive pitch and pitch patterns may result in a high-pitched voice, monotonous voice, or inappropriate use of repeated pitch patterns. *See* PITCH.

Quality. Disorders of voice quality (resonance) can be divided into two major groups. In one group the problem is at the point of origin of the voice, the vocal folds and larynx; in the other, at the supralaryngeal resonating cavities. An example of the first type is functional dysphonia, related to emotional stress, in which breathy voice indicates vocal cord vibration without a closed phase. Vocal abuse may also cause vocal fold nodules, which do not permit adequate closure of the folds, yielding a breathy and harsh vocal quality. In contrast, when the vocal cords are squeezed too tightly, they cannot vibrate properly, and the voice sounds hoarse and low-pitched. Hoarseness associated with excessive vocal cord closure is known as hyperfunctional dysphonia. Hoarseness or breathiness can also result from organic disease affecting the vocal cords, such as paralysis of one vocal cord, trauma, myasthenia, or cancer of the larvnx.

Disorders of quality related to the resonating cavities are typically of two types, hypo- or hypernasal. Hyponasal voice results when there is inadequate nasal resonance. This may be due to a physiologic or anatomic blockage of the vibrating air that ordinarily passes through the nasal cavities. In this case, the vibrating airstream does not reach the nasal passages. Hypernasality is the opposite condition, where too much air comes through the nasal cavities during phonation. This is a common problem associated with cleft palate. Speakers who are imprecise in their articulation and speak rapidly also demonstrate a mild form of this problem. Another type of hypernasality occurs when there is obstruction in the anterior nasal cavities (pinching the nostrils), which produces a quality similar to that of a nasal twang. In this case, the vibrating airstream reaches the nasal cavities but is unable to pass through them unimpeded.

Loudness. Disorders of loudness relate to the intensity of voice production. Here the voice is either not loud enough to be audible or too loud. The latter problem may occur in persons who work in very noisy environments and must shout in order to be heard. These persons carry this loud production over to the nonwork environment. Persons whose voice intensity is significantly below that which can be heard are typically those individuals whose feel inadequate and insecure. Loudness changes can often by symptomatic of hearing loss. When there is a conductive hearing impairment, the voice is often not loud enough, while in the case of sensoryneural impairment the voice is too loud. *See* LOUD-NESS.

Alaryngeal speech. One other disorder of vocal production that should be considered is alaryngeal or esophageal speech. This results following surgical removal of the larynx, usually because of cancer. The individual who has undergone such surgical removal can, by means of esophageal speech or an artificial larnyx, produce adequate but limited voice. *See* LAR-YNX.

Treatment. Treatment of the voice disorders begins with diagnosis. It is essential that any organic basis or the disorder be determined and treated before beginning voice therapy with a certified speech-language pathologist. Beyond medical treatment, therapy may include modification of the environment, psychological counseling, and direct therapy. The last will usually include the development of listening skills, relaxation exercises, and specific vocal exercises to eliminate vocal abuse and determine the optimal pitch and resonance. Listening exercises are particularly useful in treating disorders where the voice is too loud.

Fluency. Among the most difficult disorders of speech to treat are those of fluency, which are characterized by a disruption of the normal rate of speech production. There are three major types of fluency disorders: stuttering, cluttering, and those related to motor speech disorders of neurological origin (dysarthria and apraxia). Of these, stuttering is the most commonly observed.

Stuttering. No one definition of stuttering encompasses the views of scholars devoted to its study. However, the following features appear in most definitions: a disruption or blockage of the normal flow of speech because of repetitions, prolongations of sounds or syllables, and periods of silence. All of these features are involuntary and appear to be associated with effortful struggles to avoid the behavior.

There is no universal agreement as to the cause of the disorder. Three prominent theories attempt to explain the disorder. One assumes that it is a learned behavior, another that it represents a psychological disorder, and the third that it has an organic basis in disordered neurologic function.

There is no cure for stuttering. However, many persons with this disorder can be treated so that the degree of dysfluency and the associated distracting secondary behaviors (for example, facial twitches or body movement) can be significantly reduced. No single accepted mode for treatment of stuttering is generally accepted. Almost everything from surgery of the tongue to psychoanalysis and hypnotherapy has been used to try to cure the disorder. Interestingly, of the children who reportedly stutter prior to puberty, only 20% continue to do so after the onset of puberty. This is the case even in the absence of therapy.

The approach to treatment is different for children and adults; however, most therapeutic approaches fall into one of three types: psychological approaches, modification of the stuttering, and modification of speech. In the first, the emphasis is on counseling and psychotherapy to alter the stutterer's self-image and attitude toward the dysfluency and to reduce avoidance behaviors. In treatment directed at modifying the stuttering itself, the focus is on enabling the individual to stutter more efficiently and fluently. When treatment is directed at modifying speech, the goal is to achieve "stutter-free" production by altering the rhythm, rate, and voice. This approach draws heavily on the principles of behavior modification. Most speech pathologists use a combination of these general approaches when working with the dysfluent individual.

Cluttering. This is distinguished from stuttering by the fact that it is produced without awareness or concern on the part of the speaker. Cluttering is characterized by very rapid production accompanied by alterations in articulatory patterns, particularly involving substitutions and omissions. This combination, especially in the case of children, frequently renders speech unintelligible. It is often the classroom teacher who is the first to detect the problem. In adults who have been unrecognized clutterers, it may be an employer, friend, or colleague who recognizes the problem. Treatment focuses on recognition of the cluttering by the speaker, followed by specific techniques to alter rate and articulatory patterns.

Other disorders. Speech rate and rhythm can be significantly impeded by motor speech disorders. Diseases of the nervous system such as multiple sclerosis or parkinsonism and conditions such as cerebral palsy interfere with the normal sensory-motor control over the articulatory and rhythmic production of speech. Because the etiology of this group of fluency disorders varies so much, there are no specific patterns that can be described. Treatment therefore focuses on having the individual attempt to gain as much control over production as possible, given the limits imposed by the cause. In some cases, such as Parkinson's disease, the use of drugs such as 1-dopa can have a marked effect on speech production. Muscle relaxants are sometimes used along with speech therapy for patients with multiple sclerosis. However, in cases of progressive neurologic disease, treatment can provide only temporary respite from the disorder. See MULTIPLE SCLEROSIS; PARKINSON'S DISEASE.

Language. Language disorders involve impairment of the ability to voluntarily use or comprehend the language of the individual's linguistic community. Such disorders need not involve a concomitant impairment of speech or voice production. These disorders can be divided into two groups. In delayed language acquisition, normal language development fails to occur because of neurologic or intellectual problems. Aphasia typically occurs in the adult following a stroke or traumatic injury to the brain after the individual had acquired and was using language appropriate to age, gender, and linguistic community. In addition, there has been increasing recognition of language disorders associated with the dementias. The language disorders associated with delayed acquisition in children and those seen in individuals who had previously acquired adequate language share few similarities, and the approaches to treatment are dissimilar.

Aphasia. The aphasias can be categorized in a variety of ways. Nonfluent aphasics show good comprehension but have difficulty with confrontation naming and produce cryptic "telegraphic" speech. Fluent aphasics show great language output, often meaningless, with very poor comprehension. However, there are some aphasics whose only apparent difficulty is in the use of names, and while they have reactively good comprehension, their expressive language depends on circumlocutions and nonspecific words. Finally, there are global aphasics who are nonfluent, with poor comprehension and repetition, and often able only to utter a meaningless syllable.

Because the event that produces the aphasia usually involves more than one region of the brain, the aphasic often exhibits other problems such as hemiplegia (paralysis of one side of the body); motor speech problems (dysarthria); visual problems such as blindness of one visual field (hemianopsia); or apraxia (difficulty in programming, planning, initiating, and sequencing motor behaviors, especially those required for articulation). In addition, because the onset of the neurologic event producing the aphasia is frequently sudden, the aphasic is often frightened and devastated. The impact of the disorder is felt not only by the patient but also by family members who will have to play a significant role in rehabilitation.

Some spontaneous reduction of the aphasia often occurs in the first few days and weeks following the onset of the disorder. The amount of improvement is governed by the extent of the brain damage, its locus, and the patient's overall condition and motivation. Most of these changes take place during the first 3 months following onset, with some continuation at a slower pace through the sixth month. However, language therapy during this period will enhance the spontaneous recovery process. Continued improvement, with the goal of having the aphasic reach the best level of communicative function possible, requires language therapy by a qualified speechlanguage pathologist. The first step in this process is a detailed evaluation of the aphasic's current language status. From this evaluation a treatment plan can be developed. Both individual and group treatment approaches are helpful. There are no uniform treatment prescriptions for aphasia. Techniques such as programmed learning, behavior modification, and cognitive and language stimulation have all been used with varying degrees of success. Some have recommended the use of Amerind (American Indian sign language) for severely impaired aphasics. Another approach attempts through a series of specific steps to achieve improved communication by a program known as Promoting Aphasic Communicative Effectiveness (PACE). There have also been efforts to incorporate personal computers into aphasia rehabilitation. Most of these techniques, as well as others, do function to reduce the communicative impairment associated with aphasia. However, it is unusual for aphasics to make a complete recovery, and most must learn to function with limited language abilities. See APHASIA.

Childhood disorders. Language disorders of childhood are often associated with other disorders, such as mental retardation, hearing loss, and autism. However, there are some children who show specific language impairments. They are considered to be normal in other aspects of development, in particular, hearing and nonverbal intelligence. Terms uesd to categorize the childhood disorders include language delay, developmental aphasia, or language impairment. In the mentally retarded child, most of the linguistic features of the language are slow to emerge. However, when they do appear, they are similar to those of the normal child. The more abstract the features, the greater the delay in acquisition. Some features may never emerge. This is dependent in part on the severity of the retardation. Autistic children, on the other hand, suffer from marked failure to acquire and use language in the ordinary sense. Some use echolalia, unsolicited repetitive or imitative verbal patterns echoing a previously heard utterance. Since language is dependent upon hearing, children with hearing loss or deafness will show mild to profound delays in language acquisition, with the degree of impairment relative to the degree of hearing loss. See AUTISM; MENTAL RETARDATION.

Therapy for children with language delay often focuses on production of specific linguistic structures. This may involve modeling the structure for the child. Some speech-language pathologists advocate a behavioral-learning approach using aspects of behavior modification. Others focus on an interpersonal-interactive methodology. Because of the variety of problems presented by these children, treatment must be tailored to their specific needs and environments.

Dementia. Persons with dementia, such as Alzheimer's type or multi-infarct dementia, often show a disturbance of language. These disturbances can take a variety of forms, such as the inability to recall and use names or significantly impaired comprehension and language output. Linguistic deficits associated with dementia, except in rare instances, are related to the deterioration of intellect and memory rather than a primary deficit in language function. Treatment for these deficits is often indirect, working to create a language environment that is carefully structured to the present level of linguistic and cognitive function. Often therapy is directed at helping the family and care givers appreciate the problem and cope with it. *See* ALZHEIMER'S DISEASE; SPEECH.

R. S. Tikofsky Bibliography. D. R. Boone, *Human Communication and Its Disorders*, 1987; G. H. Shames, *Human Communication Disorder*, 5th ed., 1998.

Speech perception

A term broadly used to refer to how an individual understands what others are saying. More narrowly, speech perception is viewed as the way a listener can interpret the sound that a speaker produces as a sequence of discrete linguistic categories such as phonemes, syllables, or words. *See* PHONETICS; PSY-CHOLINGUISTICS.

Classical work in the 1950s and 1960s concentrated on uncovering the basic acoustic cues that listeners use to hear the different consonants and vowels of a language. It revealed a surprisingly complex relationship between sound and percept. The same physical sound (such as a noise burst at a particular frequency) can be heard as different speech categories depending on its context (as "k" before "ah," but as "p" before "ee" or "oo"), and the same category can be cued by different sounds in different contexts. Spoken language is thus quite unlike typed or written language, where there is a relatively invariant relationship between the physical stimulus and the perceived category.

The reasons for the complex relationship lie in the way that speech is produced: the sound produced by the mouth is influenced by a number of continuously moving and largely independent articulators. This complex relationship has caused great difficulties in programming computers to recognize speech, and it raises a paradox. Computers readily recognize the printed word but have great difficulty recognizing speech. Human listeners, on the other hand, find speech naturally easy to understand but have to be taught to read (often with difficulty). It is possible that humans are genetically predisposed to acquire the ability to understand speech, using special perceptual mechanisms usually located in the left cerebral hemisphere. See HEMISPHERIC LATERALITY.

Building on the classical research, the more recent work has drawn attention to the important contribution that vision makes to normal speech perception; has explored the changing ability of infants to perceive speech and contrasted it with that of animals; and has studied the way that speech sounds are coded by the auditory system and how speech perception breaks down in those with hearing impairment. There has also been substantial research on the perception of words in continuous speech.

Hearing and vision. Traditionally, speech perception has concentrated on the problem of turning sound into linguistic categories; but an important influence on the field has been the need to integrate visual information about articulation with sound.

Although lip-reading is a commonly recognized skill of the hearing-impaired, the contribution that is made by visible face movements to normal speech perception was not fully realized until H. McGurk serendipitously discovered the illusion that has become known as the McGurk effect. The sound of someone saying "bah" was dubbed onto a video recording of the same person saying either "dah" or "gah." Listeners who simply heard the sound track reported correctly "bah," but those who also saw the face reported that the person had said "dah." Vision seems to give the listener clear information that can overrule or combine with the information that is being received by the ear. In this example, vision clearly indicates that the speaker's lips did not close. Since the lips should have closed if a "b" had been produced, perception produces a conciliatory percept that is both compatible with the seen information and auditorially similar to the actual sound.

Vision gives very clear information about certain speech sounds (primarily those involving the lips and jaw); however, it provides very little about other sounds and none about the presence or the value of voice pitch. The presence of voicing conveys important information about consonants, distinguishing "b" from "p" for instance; and the pitch contour indicates the position of stressed syllables, and varies the semantic and pragmatic value of an utterance. Prostheses for the profoundly deaf that simply convey information about voice pitch can substantially improve their ability to understand speech.

The McGurk effect is much weaker in young children than in adults, but very young infants are surprisingly sensitive to the relationship between speech sounds and the corresponding face movements. Even at a few months old, they will prefer to look at faces that match the sounds they are hearing than at those that do not. Attention to the visible as well as the audible consequences of articulation is undoubtedly helpful to the child in learning language. A visually simple gesture, such as closing the lips, can influence many different auditory cues; seeing the single gesture may help the child to grasp the underlying integrity of the sound. Visually handicapped children are slower than normal children in learning to produce sounds that have a visible articulation.

Perception and development of speech categories. Adult listeners are exquisitely sensitive to the differences between sounds that are distinctive in their language. The voicing distinction in English (between "b" and "p") is cued by the relative timing of two different events (stop release and voice onset). At a difference of around 30 milliseconds, listeners hear an abrupt change from one category to another, so that a shift of only 5 ms can change the percept. On the other hand, a similar change around a different absolute value, where both sounds are heard as the same category, would be imperceptible. The term categorical perception refers to this inability to discriminate two sounds that are heard as the same speech category.

Categorical perception can arise for two reasons: it can have a cause that is independent of the listener's language—for instance, the auditory system may be more sensitive to some changes than to others; or it can be acquired as part of the process of learning a particular language. The example described above appears to be language-independent, since similar results have been found in animals such as chinchillas whose auditory systems resemble those of humans. But other examples have a language-specific component. The ability to hear a difference between "r" and "l" is trivially easy for English listeners, but Japanese perform almost at chance unless they are given extensive training. How such languagespecific skills are developed has become clearer following intensive research on speech perception in infants.

Newborn infants are able to distinguish many of the sounds that are contrasted by the world's languages. Their pattern of sucking on a blind nipple signals a perceived change in a repeated sound. They are also able to hear the similarities between sounds such as those that are the same vowel but have different pitches. The ability to respond to such a wide range of distinctions changes dramatically in the first year of life. By 12 months, infants no longer respond to some of the distinctions that are outside their native language, while infants from language communities that do make those same distinctions retain the ability. Future experience could reinstate the ability, so it is unlikely that low-level auditory changes have taken place; the distinctions, although still coded by the sensory system, do not readily control the infant's behavior.

Listening in the speech mode. Individual speech categories (such as "p" or "b") are cued by a wide variety of different aspects of sound. For example, "b" has a longer preceding vowel and a shorter silence than does "p." In addition, when the lips open, the initial burst of noise energy has a lower amplitude; and when voicing starts the peak of the lowfrequency energy is at a lower frequency for "b" than for "p." Experiments with synthetic speech show that these different cues can be traded off against each other: a longer silence can be compensated for by a lower first format onset frequency. Such so-called trading relations between these different cues may sometimes be explained by basic sensory mechanisms such as adaptation in the auditory nerve, but others require more specialized mechanisms specific to speech. A particularly useful way of demonstrating this is to synthesize a caricature of speech known as sine-wave speech; it consists of sine waves at the formant frequencies. Short utterances of sine-wave speech can be heard either as nonspeech whistles or as (peculiar) speech. Some of the trading relations between cues in normal speech are also found in sine-wave speech, but can be shown to arise only when the listener hears the sound as speech. They are therefore a result of special speechprocessing mechanisms, rather than purely auditory ones

Words and continuous speech. Continuous speech is not simply a concatenation of isolated syllables; syllables change their pitch, their duration, and their phonetic character in continuous speech. In extreme circumstances, they can vanish or be completely transformed—for example, "Do you want to?" becoming "Djawona?" Although such changes make it more difficult to understand how the lexicon can be accessed from the speech waveform, the changes in duration and pitch convey essential information about words, grammar, and intention.

Under good listening conditions and with clearly articulated speech, it is often possible to identify words before they have been completed verbally. A word such as "penguin" becomes identifiable in principle before the final vowel, since there is no other word in English that begins the same way (except for derived forms such as "penguins"). In practice, listeners can identify isolated words at their uniqueness points, but may not do so in normal speech where the necessary phonetic information can be absent or distorted. Subsequent context can then be used to limit the possibilities before the word is "heard." Speakers adjust the clarity of their speech depending on the context, relying on listeners' ability to use that context to make up for the decreased information in the signal. Stressed syllables are most resistant to such adulteration; they may serve as starting points for the perceptual process and, in English at least, as candidates for the beginnings of words.

The hearing-impaired. Although conductive hearing losses can generally be treated adequately by appropriate amplification of sound, sensorineural hearing loss involves a failure of the frequency-analyzing mechanism in the inner ear that humans cannot yet compensate for. Not only do sounds need to be louder before they can be heard, but they are not so well separated by the ear into different frequencies. Also, the sensorineurally deaf patient tolerates only a limited range of intensities of sound; amplified sounds soon become unbearable (loudness recruitment).

These three consequences of sensorineural hearing loss lead to severe problems in perceiving a complex signal such as speech. Speech consists of many rapidly changing frequency components that normally can be perceptually resolved. The **illustration** shows a spectrogram of speech that has been put through a filter bank similar to that found in the normal human ear. In the low-frequency region the individual harmonics of the voice are separated; at higher frequencies the ear cannot resolve individual harmonics, but it still separates the different formants (or resonances of the vocal tract).

The lack of frequency resolution in the sensorineural patient makes it harder for the listener to identify the peaks in the spectrum that distinguish the simplest speech sounds from each other; and the use of frequency-selective automatic gain controls to alleviate the recruitment problem reduces the distinctiveness of different sounds further. These patients may also be less sensitive than people with normal hearing to sounds that change over time, a disability that further impairs speech perception.

Although normal listeners can still recognize speech against a background of other noises (including speech, as in the cocktail-party effect), sensorineural patients are disproportionately affected by competing sounds. Their impaired frequency and temporal selectivities reduce their ability to separate



An auditory spectrogram of the world "telephone." (a) The vertical striations in the bands represent amplitude modulation at the frequency of the voice pitch. Such striations are absent from the initial noise portion corresponding to the voiceless initial "t." The medial "f" is too weak to appear on the spectrogram. (b) The first eight or so harmonics of the voice pitch are resolved and appear as dark, roughly horizontal bands. Higher harmonics are not resolved (here above about 1.5 kHz). The dark horizontal bands now represent the formant frequencies.

the required signal from its background. Hearing aids that can listen in to sound from a particular direction can help to filter out unwanted sounds, but they do not perform as well when echoes and reverberation are present. A practical solution to this problem has not been developed; however, it is possible in artificial situations to separate the speech of one speaker from another on the basis of voice pitch, so that a deaf listener can hear the selected voice more clearly. The same technique is also used by normal listeners as one way of solving the problem of competing background sounds. Extension of this method to normal situations is complicated by the fact that not all of speech has a pitch. Other, more complex information must be used to separate voiceless sounds and to link the voiced segments appropriately.

Some profoundly deaf patients can identify some isolated words by using multichannel cochlear implants. Sound is filtered into different frequency channels, or different parameters of the speech are automatically extracted, and electrical pulses are then conveyed to different locations in the cochlea by implanted electrodes. The electrical pulses stimulate the auditory nerve directly, bypassing the inactive hair cells of the damaged ear. Such devices cannot reconstruct the rich information that the normal cochlear feeds to the auditory nerve, but they promise to put profoundly deaf patients back into contact with the world of sound and to help them to communicate more effectively by speech. See HEAR-ING (HUMAN); HEARING AID; HEARING IMPAIRMENT; PERCEPTION; PSYCHOACOUSTICS; SPEECH.

C. J. Darwin

Bibliography. W. D. Brooks and R. W. Heath, *Speech Communication*, 7th ed., 1993; B. Dodd and R. Campbell, (eds.), *Hearing by Eye: The Psychology of Lip-Reading*, 1987; U. H. Frauenfelder and L. K. Tyler (eds.), *Spoken Word Recognition*, 1978; I. G. Mattingly and M. G. Studdert-Kennedy (eds.), *Modularity and the Motor Theory of Speech Perception*, 1990; M. E. H. Schouten (ed.), *The Psychophysics of Speech Perception*, 1987.

Speech recognition

Speech recognition, in a strict sense, is the process of electronically converting a speech waveform (as the acoustic realization of a linguistic expression) into words (as a best-decoded sequence of linguistic units). At times it can be generalized to the process of extracting a linguistic notion from a sequence of sounds, that is, an acoustic event, which may encompass linguistically relevant components, such as words or phrases, as well as irrelevant components, such as ambient noise, extraneous or partial words in an utterance, and so on. Applications of speech recognition include an automatic typewriter that responds to voice, voice-controlled access to information services (such as news and messages), and automated commercial transactions (for example, price inquiry or merchandise order by telephone), to name a few. Sometimes, the concept of speech recognition may include "speech understanding," because the use of a speech recognizer often involves understanding the intended message expressed in the spoken words. Currently, such an understanding process can be performed only in an extremely limited sense, often for the purpose of initiating a particular service action among a few choices. For example, a caller's input utterance "I'd like to borrow money to buy a car" to an automatic call-routing system of a bank would connect the caller to the bank's loan department.

Fundamental concept. Converting a speech waveform into a sequence of words involves several essential steps (Fig. 1). First, a microphone picks up the acoustic signal of the speech to be recognized and converts it into an electrical signal. A modern speech recognition system also requires that the electrical signal be represented digitally by means of an analogto-digital (A/D) conversion process, so that it can be processed with a digital computer or a microprocessor. This speech signal is then analyzed (in the analysis block) to produce a representation consisting of salient features of the speech. The most prevalent feature of speech is derived from its short-time spectrum, measured successively over short-time windows of length 20-30 milliseconds overlapping at intervals of 10-20 ms. Each short-time spectrum is transformed into a feature vector, and the temporal sequence of such feature vectors thus forms a speech pattern. Figure 2 illustrates the speech analysis process.

The speech pattern is then compared to a store of phoneme patterns or models through a dynamic programming process in order to generate a hypothesis (or a number of hypotheses) of the phonemic unit sequence. (A phoneme is a basic unit of speech and a phoneme model is a succinet representation of the signal that corresponds to a phoneme, usually embedded in an utterance.) A speech signal inherently has substantial variations along many dimensions. First is the speaking rate variation-a speaker cannot produce a word of identical duration at will. Second, articulation variation is also abundant, in terms both of talker-specific characteristics and of the manner in which a phoneme is produced. Third, pronunciation variation occurs among different speakers and in various speaking contexts (for example, some phonemes may be dropped in casual conversation). Dynamic programming is performed to generate the best match while taking these variations into consideration by compressing or stretching the temporal pattern (a process often called dynamic time warping) and by probabilistically conjecturing how a phoneme may have been produced. The latter includes the probability that a phoneme may have been omitted or inserted in the utterance. The knowledge of probability (often called a probabilistic model of speech) is obtained via "training," which computes the statistics of the speech features from a large



Fig. 1. Basic block diagram of a speech recognition system.

collection of spoken utterances (of known identity) according to a mathematical formalism. At the present time, the most widely adopted formalism is the hidden Markov model.

The hypothesized phoneme sequence is then matched to a stored lexicon to reach a tentative decision on the word identity. The decoded word sequence is further subject to verification according to syntactic constraints and grammatical rules, which in turn define the range of word hypotheses for lexical matching. This process of forming hypotheses about words and matching them to the observed speech pattern, and vice versa, in order to reach a decision according to a certain criterion is generally referred to as the "search." In limited domain applications in which the number of legitimate expressions is manageably finite, these constraints can be embedded in an integrated dynamic programming process to reduce search errors.

Categorization of tasks. Most current speech recognition systems are essentially designed to convert a speech waveform into one or more words in

order to perform a specific task. Linguistic expressions via speech can take the form of word utterances spoken individually in isolation, or phrases and sentences spoken naturally without intentionally controlled word boundaries. A task involving the former case is called isolated (or discrete) word recognition, while the latter case is generally referred to as continuous speech recognition. However, the speaker's articulation effect can be substantially different when reading a prepared text and when conversing spontaneously with another person. Also at times, a task may involve connected words without the usual syntactic constraints, such as a digit sequence (for example, one's credit card number). Thus, a speech recognition task can be categorized, in terms of the way that speech is rendered, into four essential types: isolated word, connected word, continuously read speech, and spontaneous speech.

The degree of sophistication of a speech recognition task is largely a function of the size of the vocabulary it has to deal with. A task involving, say, less



Fig. 2. Short-time spectral analysis of speech.

than 100 words is called small vocabulary recognition and is mostly for command-and-control applications with isolated word utterances as input. There are usually very few grammatical constraints associated with these types of limited tasks. When the vocabulary size grows to 1000 words, it is possible to construct meaningful sentences, although the associated grammatical rules are usually fairly rigid. For dictation, report writing, or other tasks such as newspaper transcription, a speech recognition system with a large vocabulary (on the order of tens of thousands of word entries) is needed.

A speaker-independent recognition system is designed to serve nonspecific users and is suitable for public service applications where the user population is large. A voice dialing system embedded in a telephone central office switch, a telephone callrouting system, and a customer service unit that can respond to a caller's voice without having to know the user's identity are all in this category. It has to be able to handle a broad range of accents and voice characteristics from various talkers. Another type of speech recognizer, called a speakerdependent recognition system, may be designed to serve only a specific user; for example, a personal word-processor that responds to its authorized user's voice. A speaker-dependent system requires speakerspecific training and is thus also referred to as a speaker-trained system. Such a system can also be designed by adapting a speaker-independent system to the voice patterns of the designated user via an enrollment procedure; it is often called a speakeradaptive system. For a limited number of frequent users of a particular speech recognizer, such an enrollment procedure generally leads to good recognizer performance.

Understanding of speech. Another type of speech recognition system has to do with the fundamental idea of task-oriented speech understanding. Using the popular call-routing system that sorts operatorassisted calls into five types (credit-card, collect, third-party billing, person-to-person, and operator) as an example, a recognition system design engineer can use a large-vocabulary continuous speech recognition system to convert the request utterance (for example, "I'd like to make a collect call") into a word sequence as best it can and then employ a natural language processing system to parse and understand the intent of the caller. Alternatively, the system designer can make the assumption that in such a limited task the intent of the caller will most likely be expressed in a "keyword" (for example, the word "collect" in the above sentence) embedded in the otherwise naturally spoken sentence. A system that can detect or spot the keyword would thus suffice for the task. This alternative approach is referred to as keyword spotting.

The keyword-spotting system has proved to be very useful, because it makes the interaction between the user and the machine natural and robust. Experience shows that many users of a speech recognition system in telecommunication applications often do not speak words or phrases according to the prescribed vocabulary, thus creating out-ofvocabulary (OOV) or out-of-grammar (OOG) errors. Rather than attempting to recognize every word in the utterance, a keyword-spotting system hypothesizes the presence of a keyword in appropriate portions of the speech utterance and verifies the hypothesis by a likelihood ratio test. By avoiding forced recognition decisions on the unrecognizable and inconsequential regions of the speech signal, the system can cope with natural sentences, obtaining satisfactory results in constrained tasks. A more sophisticated keyword spotting system may involve "gisting," that is, trying to compile the detected keywords into a possible message. Similarly, it is possible to perform a primitive analysis on the decoded words based on their correlation with a particular semantic notion in order to derive an understanding for limited domain applications.

Implementation. Realization of a speech recognition system involves two essential steps: training and decoding. The objectives of training include reference models of the phonetic units, which characterize (probabilistically) how linguistic units are acoustically realized (acoustic modeling), and probabilistic relationships among words that form grammati-cally legitimate sentences of a language (language modeling). These acoustic unit models and language models are the basis of reference for inferring a decision when given an unknown input utterance. Training is usually performed off-line in nonreal time.

The hidden Markov model is the prevalent choice for the reference structure in acoustic modeling. A hidden Markov model is a doubly stochastic process involving (at least) two levels of uncertainty. It is a finite state (say N-state) machine in which each state is associated with a random process. The random process in each state accounts for the variation in the realization of a speech code (such as a phoneme), including both the spectral (pertaining to the vocal tract) and the temporal (pertaining to the rate of articulation) behavior. The finite state machine is a Markov chain, which characterizes the probabilistic relationship among the states in terms of how likely one state is to follow another state. Figure 3 illustrates a three-state hidden Markov model. See STOCHASTIC PROCESS.

Training of a hidden Markov model requires a large collection of sample utterances, each with a known label. In the case of isolated-word recognition, each utterance would consist of a word. The identity or label of the utterance is usually manually obtained by a human listener. For continuous speech recognition, each utterance in the training set would be a sentence, represented by a manually annotated word sequence. Model training involves a procedure in which locating the proper unit boundaries in the utterance (segmentation) and adjusting the model parameter values for each encountered models (model estimation) interleave and iterate until a certain criterion is optimized.



Fig. 3. Three-state hidden Markov model.

A similar principle applies in the case of language modeling, which is imperative in large-vocabulary continuous speech recognition. Language modeling is usually performed on a text database consisting of tens or hundreds of millions of words, sufficient to allow derivation of the probabilistic characteristics of a language (for example, the likelihood of a particular word appearing after another).

Decoding is the process of evaluating the probabilities that the unknown utterance manifests the same characteristics as each of the stored references and inferring the identity of the utterance based on the evaluated probabilities. For large-vocabulary, continuous speech recognition, the reference models (such as sentence hypotheses) are constructed from unit models and can be prohibitively large in number. Figure 4 shows an example of representing a hypothesized sentence ("Show all alerts") by a sequence of context-dependent phoneme units. Efficient search algorithms need to be invoked such that the decoding result can be obtained within a reasonable computational bound. For a recognition system to be useful, it needs to be able to produce the recognition result without obvious latency, preferably in real time.

Except for some extremely simple tasks, a speech recognition system is implemented as a stored program for execution with a computer or a digital signal processor. The reference models are stored in read-only devices, with easy data access for decoding purposes. For telecommunication applications in which the system often handles multichannel traffic or sessions, implementation efficiency is achieved by sharing the reference models and proper division of the computational task among multiple channels.

Difficulties and remedies. Speech recognition research has made great progress since 1970. However, the current technology is still far from achieving its ultimate goal: accurate recognition of unconstrained speech from any speaker in any environment or operating condition. There are several key reasons why this remains a challenging problem.

One primary difficulty in speech recognition is that the variation in articulation is difficult to characterize. While phonemes are considered the fundamental speech units demonstrating regular acousticphonetic behavior (as determined by human ears), their acoustic realization usually contains ambiguity and uncertainty. A phoneme recognition system that attempts to convert a speech waveform into a phoneme sequence without incorporating additional knowledge, such as vocabulary entries and syntax, cannot achieve high accuracy. Grouping these error-prone phoneme or subword sequences into a sequence of words from a standard lexicon is a nontrivial task. Units as large as words and phrases have been suggested for limited task environments with good success. Large speech units, however, are unsuitable for large-vocabulary, continuous speech recognition. The longer the basic unit, the more the number of unit-classes there will be for the system to deal with. (Consider the extreme case of using the sentence as the unit in continuous speech recognition. The number of possible sentences becomes astronomical even with a moderate vocabulary size.) Having more unit classes also means higher complexity in training and decoding. For practical



Fig. 4. Sentence representation of "show all alerts" with a finite state network using context-dependent units; "sil" denotes silence.

reasons, subword units such as phonemes, diphones (pairs of phonemes), or syllables, which are limited in number almost independent of the vocabulary size, are preferred. To circumvent the articulation problem, modern large-vocabulary, continuous speech recognition systems almost universally use context-dependent unit models, an example of which is shown in Fig. 4. A context-dependent phoneme model is a statistical characterization of a phoneme when uttered in a particular context; for example, a phoneme unit /a/ that follows a /n/ and precedes a /p/, as in "snap," is a member of the unit classes with model denoted by 'n-a-p.' This class is treated as a class different from, say, 'm-a-d' in "mad." Context dependency helps reduce the degree of acoustic-phonetic variation within the class and often leads to higher recognition accuracy.

Another major difficulty is the variation in the characteristics of the transmission medium over which recognition is performed (for example, over a telephone line, in an airplane cockpit, or using a handsfree speakerphone). This kind of variation includes the acoustic background noise, the various distortions inherently caused by the microphones that are used to pick up the acoustic signal, reverberation that varies substantially from one room to another, and the transmission distortion existing in the telephone network. To cope with this type of variation, assumptions, such as that the noise is additive and that the distortion is linear, are usually made. Another implicit assumption about the characteristics of the interference is that they do not change as rapidly as speech. For example, in telephone applications, most of the transmission properties are considered virtually constant over a reasonable period of time once the phone circuit connection is made. Signal processing methods to remedy these adverse effects include spectral subtraction to reduce the impact of additive noise, and cepstral bias or mean removal to suppress the influence of spectral distortions. The principle behind spectral subtraction and cepstral bias removal is quite similar. Spectral subtraction operates in the power spectrum domain and assumes that the power spectrum of the received signal is the sum of the speech spectrum and the noise spectrum. Cepstral bias removal operates in the log power spectrum domain and hence assumes that the log power spectrum of the received signal is the sum of the log power spectrum of the speech and the log frequency response of the transmission channel. In both cases, the system estimates the additive noise/bias component from an average of the signal (in the corresponding domain taken over a relatively long period, perhaps the whole utterance, based on the aforementioned assumption on the slow-changing nature of these adverse effects) and subtracts the bias estimate from each of the short-time speech representations that form the speech pattern of the utterance.

Talker variation is yet another source of difficulty. Variations due to the talker include regional accents, articulation and pronunciation, and mode and mood of speaking. Talker adaptation provides means for the system to adjust its operating parameters, such as the stored reference models, to suit the particular user. However, very limited knowledge exists at present on the effect of psycholinguistic factors, such as the contextual or acoustic environment (for example, a noisy café) in which the speech conversation takes place, upon the manifestation of acoustic-phonetic features in the speech signal. Some studies exist on the Lombard effect, which is used to describe the observation of change in the talker's articulation effort when he or she speaks in a noisy environment (for example, trying to raise the voice or to make his or her voice better understood by the listener).

Similarly, studies to establish systematic understanding of the difference between read and spontaneous speech are just beginning. The difference is intuitively obvious, but the science behind it is not. Unconstrained automatic speech recognition by a machine that converses with people like a human is still a challenging undertaking.

State of the art. Speech recognition by the end of the twentieth century achieved modest success in limited applications, particularly with systems designed to recognize small- to moderate-size vocabularies for well-defined tasks (such as word processing and document editing on a personal computer using voice, dictating a readout in a radiologist's office, and telephone dialing by speaking the name or the number).

A summary of typical speech recognition performance in terms of word error rate is provided in **Tables 1** and **2**. Table 1 shows the average word error rate for recognition of isolated words without contextual constraints for both speaker-dependent (SD) and speaker-independent (SI) systems. In general, the performance is more sensitive to vocabulary complexity than to vocabulary size. The 39-word alpha-digits vocabulary (letters A–Z, digits 0–9, three command words) contains highly confusable word subsets, such as the E-set (B, C, D, E, G, P, T, V, Z) and the digit 3, and has an average accuracy of 93– 96%. In comparison, a task involving an 1109-word basic English vocabulary has an average accuracy of 96%.

For connected-word recognizers applied to tasks with various types of syntactic or grammatical constraints as in natural language, an additional parameter called perplexity is needed to indicate roughly the degree of difficulty of the task. Perplexity is defined as the average number of words that can follow an arbitrary word. It is usually substantially less than the size of the vocabulary because of the constraints of the task-specific language. Table 2 shows the typical performance, again in terms of word accuracy, of connected and continuous speech recognition systems for various tasks.

The resource management task involves a highly stylized language used in military duties with a perplexity of no more than 60. Sentences in the air travel information system are usually simple query utterances aiming at flight information such as

TABLE 1. Performance of isolated word recognition systems						
Recognition task	Vocabulary size	Mode	Word accuracy			
Digits (0–9)	10	SI	~100%			
Alpha-digits and command words	37 39	SD SD	96%			
Computer commands	54	SI	93% 96%			
Air travel words	129	SD SI	99% 97%			
City names (Japanese) Basic english	200 1109	SD SD	97% 96%			

airfare, flight time, and the origin and destination of the trip. The vocabulary includes a number of city or airport names in the task database. While the application language appears to be natural, the grammatical structure in the query sentence is actually rather limited with a perplexity of less than 25. In the business newspaper transcription task, the input is "read" speech; speakers read the text of news articles in the paper. Read speech has less variability in its acoustic realization, and is thus easier to recognize than spontaneous, conversational speech. The broadcast news transcription task involves "found" speech, such as the news announcements from radio or television. The signal may be degraded by noise (such as background music or conversation) and distortion (for example, due to radio transmission). Obviously, this is a more difficult task than transcribing a news article in a quiet office.

Despite progress, the performance of current systems is barely acceptable for large vocabulary systems. Tasks listed in the tables are all considered substantially easier than transcribing and understanding truly spontaneous conversational speech.

Speaker recognition. The technology of speech recognition often finds applications in speaker recognition tasks as well. Speaker recognition applications can be classified into two essential modes, speaker identification and speaker verification. The goal of speaker identification is to use a machine to find the identity of a talker, in a known population of talkers, using the speech input. Speaker verification aims to authenticate a claimed identity from the voice signal. These two modes call for different methodologies. Applications of speaker recognition include controlled entry to restricted premises, and authorization of access to privileged information or other service transactions. For these applications, speaker verification is used in most practical systems.

There are also two types of talker authentication with regard to the underlying technology. Whereas speaker verification (SV) verifies a speaker's identity based on speaker-specific characteristics reflected in spoken words or sentences (sometimes referred to as a talker's voice print), another method, known as verbal information verification (VIV), checks the content of the spoken password or pass-phrase [such as a personal identification number (PIN), a social security number, or a mother's maiden name]. In the deployment of a VIV system, the user assumes the responsibility in protecting the confidentiality of the password as in many electronic transactions applications.

Research aimed at finding acoustic and linguistic features in the speech signal that carry the characteristics of the speaker has a long history. The advent of a statistical pattern matching approach (such as the hidden Markov model) and its dominance in the 1990s has caused a convergence in the technological approach to this age-old problem. Statistical modeling methods are data-driven and are able to automatically "learn" speaker-specific (for SV) or passwordspecific (for VIV) characteristics from the provided spoken utterances with a known identity. During verification, the talker first makes an identity claim, such as a name or an account number, and then, in response to the system prompt or question, speaks a test phrase to the system. The system evaluates the input speech against the stored model for the person with the claimed identity (in SV) or the stored model for the password (in VIV), and decides to accept or reject the talker.

Key factors affecting the performance of a speaker verification system are the nature of input (whether the utterance is based on a specific text and how long the test utterance is) and the operating environment in which the verification system is used. The performance of a verification system is usually

TABLE 2. Performance benchmark of state-of-the-art automatic speech recognition systems for various connected and continuous speech recognition tasks

Recognition task	Vocabulary size	Perplexity	Word accuracy
Connected digit strings	10	10	~99%
Resource management	991	<60	97%
Air travel information system	1,800	<25	97%
Business newspaper transcription	64.000	<140	94%
Broadcast news transcription	64,000	<140	86%

measured by the equal error rate (EER), which is defined as the error rate of the system when the operating threshold for the accept/reject decision is adjusted such that the probability of false acceptance and that of false rejection become equal. A state-ofthe-art system using a text-dependent sentence-long test utterance is capable of achieving an EER of 1– 2%. With short, unconstrained spoken utterances in a noisy environment, current systems would have an EER of 4–8%.

The performance of a VIV system depends on how many passwords or pass phrases the user is asked to give as input to the system in a verification session. Experiments showed that a VIV system could achieve perfect (0 error) verification after three rounds of question and answer.

The deployment of speaker verification often involves nontechnical issues, such as the trade-off between false rejection and false acceptance. In business applications, such as an automated banking machine where a false rejection may mean unjustified inconvenience causing the loss of a legitimate longterm customer, the operating threshold may be set to allow a high false-acceptance rate. Determination of the operating point is thus a matter of business decision. *See* LINGUISTICS; PSYCHOACOUSTICS; SPEECH. Biing Hwang Juang

Bibliography. J. L. Flanagan, Speech Analysis, Synthesis, and Perception, Springer-Verlag, New York, 1972; S. Furui, Digital Speech Processing, Synthesis, and Recognition, Marcel Dekker, New York, 2001; B. H. Juang and S. Furui (eds.), Spoken Language Processing, special issue of *IEEE Proc.*, 2000; L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ, 1993; L. R. Rabiner and B. H. Juang, An introduction to hidden Markov models, *IEEE Signal Proc. Mag.*, 3(1):4–16, 1986.

Speed

The time rate of change of position of a body without regard to direction. It is the numerical magnitude of a velocity and hence is a scalar quantity. Linear speed is commonly measured in such units as meters per second, miles per hour, or feet per second. It is the most frequently mentioned attribute of motion.

Average linear speed is the ratio of the length of the path Δs traversed by a body to the elapsed time Δt during which the body moved through that path, as in Eq. (1), where s_0 and t_0 are the initial position

Speed (average) =
$$\frac{s_f - s_0}{t_f - t_0} = \frac{\Delta s}{\Delta t}$$
 (1)

and time, respectively, s_f and t_f are the final position and time, and Δ stands for "the change in."

Instantaneous speed, defined by Eq. (2), is the lim-

Speed (instantaneous) =
$$\lim_{\Delta t \to 0} \frac{\Delta s}{\Delta t} = \frac{ds}{dt}$$
 (2)

iting value of the foregoing ratio as the elapsed time approaches zero. *See* VELOCITY. Rogers D. Rusk

Speed regulation

The change in steady-state speed of a machine, expressed in percent of rated speed, when the load on the machine is reduced from rated value to zero. The definition of regulation usually means the net change in a steady-state characteristic, and does not include any transient deviation or oscillation that may occur prior to reaching the new operation point. This same definition is used for stating the speed regulation of electric motors and for certain prime movers, such as steam turbines.

In specifying the speed regulation of dc motors, for example, one would compute the steady-state regulation R by Eq. (1), where N_R is the speed (in any

$$R = \frac{N_0 - N_R}{N_R} (100) \qquad \text{percent} \qquad (1)$$

convenient units) at rated load and N_0 is the speed (in the same units used to specify N_R) after the load is removed and a new steady-state value is reached. Regulation is usually stated as a positive quantity, even though the regulation characteristic of speed versus load, when plotted as a straight line, would have a negative slope.

For small load changes about an operating point, an incremental steady-state regulation is also defined by Eq. (2), where both the speed *N* and power *P*

$$R = \frac{\Delta N}{\Delta P} (100) \tag{2}$$

are normalized by using rated values as the base quantities. Equation (2) recognizes the fact that the speed-versus-power regulation characteristic is not necessarily a straight line. In computing the regulation it should also be specified that all adjustments to the speed-changing mechanism should remain unchanged throughout the test. *See* MOTOR.

The speed regulation characteristic of steam turbines driving synchronous generators is an important parameter in the control of frequency for large power systems. Here the speed regulation, often called the droop characteristic, determines not only the overall systems frequency response characteristic to load changes but also how any load change is shared by the connected turbine-generator units. The droop characteristic is important in estimating steady-state response, but this parameter enters prominently into the determination of transient behavior as well. *See* GENERATOR.

The definition of steady-state regulation (1) is also used in the specification of the performance of automatic control systems, where it is used as the measure of goodness of a controlled variable, such as speed. *See* CONTROL SYSTEMS. Paul M. Anderson

Bibliography. N. Cohn, Control of Generation and Power Flow on Interconnected Power Systems, 1966; IEEE Power Engineering Society, Power Generation Committee, Recommended Practice for Functional and Performance Characteristics of Control Systems for Steam Turbine-Generator Units, IEEE Std. 122-1991; IEEE Standard Dictionary of Electrical and Electronics Terms, 7th ed., IEEE Std.
100-2001; L. K. Kirchmayer, *Economic Control of Interconnected Systems*, 1959.

Speedometer

A device for indicating the speed of a vehicle. The mechanical analog speedometer (**Fig. 1**) was introduced around 1906, when automobiles were beginning to increase in numbers. While this type of speedometer continues in wide use, changing technology has resulted in increasing numbers of quartz electric analog speedometers and digital microprocessor speedometers.

Mechanical speedometer. The mechanical speedometer is driven by a cable housed in a casing and connected to a gear at the transmission. This gear is designed for the particular vehicle model, considering the vehicle's tire size and rear axle ratio.

In most cases, the speedometer is designed to convert 1001 revolutions of the drive cable into registering 1 mi on the odometer, which records distance traveled by the vehicle. Thus, 1001 revolutions of the cable per minute will indicate a speed of 60 mi/h (96 km/h). Odometers with six wheels record 99,999 mi (or kilometers), while those with seven wheels record 999,999 mi (or kilometers).

The speed-indicating portion of the speedometer operates on the magnetic principle. In the speedometer head, the drive cable attaches to a revolving permanent magnet that rotates at the same speed as the cable. Floating on bearings between the upper frame and the revolving permanent magnet is a nonmagnetic movable speed cup. (There is almost no mechanical connection between the revolving magnet and the speed cup.) The magnet revolves within the speed cup, producing a rotating magnetic field. The magnetic field is constant, and the amount of speed cup movement is at all times in proportion to the speed of the magnet rotation. The magnetic field exerts drag on the speed cup, causing it to revolve in the same direction and in direct ratio to the speed of the revolving magnet and the vehicle itself. A pointer, attached to the speed cup spindle, indicates the speed on the



Fig. 1. Components of a mechanical speedometer system.

speedometer dial. The speed cup movement is opposed by a hair spring, so that the speed cup is held steady and the pointer indicates the true speed. The hair spring also pulls the pointer back to a point less than 5 mi/h (8 km/h) when the magnet stops rotating.

Quartz electric speedometer. The quartz electric speedometer and odometer assembly (**Fig. 2**) displays the speed of the vehicle, total vehicle mileage, and trip mileage. A conventional (analog) dial and pointer assembly, along with odometer wheels, are used to display this information. However, due to improved drive methods, the need for the conventional speedometer (flexible shaft) cable is eliminated.

To provide accurate vehicle information, the quartz speedometer utilizes an accurate clock signal supplied by a quartz crystal, along with integrated electronic circuitry to process an electrical speed signal. This signal, used by the circuitry to drive the air-core gage and odometer stepper motor, is generated by a permanent-magnet generator mounted in the transmission. This permanent-magnet generator, designed to be used with both quartz and digital speedometers, provides a sinusoidal speed signal that is proportional to vehicle speed at the rate of 4004 pulses per mile (2503 per kilometer). This speed signal is transmitted to a buffer amplifier which filters and squares the speed signal. After conditioning, the signal is transmitted to the speedometer circuitry at a rate of 1.111 Hz/(mi)(h) [0.694 Hz/(km)(h)].

The odometer assembly for the quartz electric speedometer consists of conventional odometer wheels driven by an electric stepper motor. The stepper motor receives a frequency-modulated square-wave signal of 0.556 Hz/(mi)(h) [0.348 Hz/(km)(h)] from the speedometer circuitry. This signal controls the rate at which the motor turns the odometer gears.

The air-core gage mechanism in the electric speedometer consists of a coil, bobbin, and case assembly that utilizes electrical signals supplied by the circuit assembly to display vehicle speed. Current flow through the coils of the air-core gage causes a flux field to be induced. As the flux field varies, the armature magnet inside of the coils rotates to align itself with the induced field. A spindle and pointer is attached to the magnet and indicates information on the dial. By controlling the flow of current through the coils (via the electronic circuit), the air-core gage can be used to display accurately the desired analog information. Once the current potential across the coils is removed, the pointer returns to zero.

Digital microprocessor speedometer. In this type of instrument (**Fig. 3**), the vehicle speed is monitored by the permanent speed sensor mounted in the transmission. Like the quartz speedometer system, the speed signal is filtered and squared by the buffer amplifier. After conditioning, the signal is transmitted to the microprocessor where the counter converts the speed signal to a digital signal and stores it in memory. The timing circuit has the capacity to handle



Fig. 2. Components of a quartz electric speedometer system.



Fig. 3. Schematic diagram of the operation of a digital electronic speedometer. The permanent-magnet speed sensor generates four pulses for every revolution; at 60 mi/h (97 km/h), 4004 pulses are generated.

the counter and memory storage in less than 0.25 s. Memory circuit signals are sent to the display circuit, which selects the display numerals representing the vehicle's speed, according to the number of pulses received from the speed sensor. The electronic displays are light-emitting diodes, fluorescent tubes, liquid crystals, plasma gas discharge tubes, or electroluminescent panels. Odometers used in conjunction with digital speedometers are stepper motor-driven drum registers or digital displays employing nonvolatile memory. *See* AUTOMOTIVE TRANSMISSION; ELECTRONIC DISPLAY; TACHOMETER.

Robert A. Grimm

Bibliography. D. Maschino, D. Taylor, and G. Thoman, *Practical Electric Speedometer System for Passenger Car Application*, SAE Tech. Pap. 830326, February 1983; Society of Automotive Engineers, *SAE Handbook Nineteen Eighty-Eight: Parts and Components*, vol. 2., 1988; W. K. Toboldt, *Automotive Encyclopedia*, rev. ed., 2000.

Spelaeogriphacea

A crustacean order within the class Malacostraca, superorder Peracarida. Few spelaeogriphacean species are known, four extant species and two supposed fossil species. *Spelaeogriphus lepidops* is from a stream in a cave in Table Mountain, South Africa. *Potiicoara brasiliensis* inhabits a pool in a cave in the Mato Grosso, Brazil. *Mangkurtu mityula* (see **illustration**) and *M. kutjarra* have been reported from aquifers in the Pilbara region of Western Australia. *Acadiocaris novascotica* is a fossil from Carboniferous marine sediments in Canada, and *Liaoningogripbus quadripartitus* is a fossil from lacustrine (lake-associated) deposits of Jurassic age in China. It is probable that *Acadiocaris* is truly a spelaeogriphacean, but it is less certain that *Liaoningogriphus* is; on the basis of characters visible in fossils, its classification in this order or any other is impossible to justify.

Spelaeogriphus lepidops is the largest of the living species, reaching 9 mm (0.36 in.) in length; the others attain half that length at most. All have a slender flexible body of similar-sized segments. The head, incorporating the first thoracic segment, supports a short carapace that extends backward over the top and sides of the second thoracic segment. The carapace encloses a branchial space in which sit a pair of maxillipedal gills, less developed in Potiicoara than in the other genera. A pair of articulating disc-like eyestalks sit on top near the front of the head but without pigmented eyes. The limbs on the head are a pair of antennules with a prominent scaphocerite (scale), a pair of long biramous (having two branches) antennae, complex mandibles with grinding molar surfaces, incisor and palp (sensory appendage near the mouth), setose maxillae and maxillules and maxillipeds each with a



Lateral (top) and dorsal views of the Western Australian spelaeogriphacean species Mangkurtu mityula.

prominent basal endite. The patterns of setae (bristles) on the mouthparts distinguish speleaogriphaceans from the most closely related peracaridan order, Mictacea. The seven pairs of walking legs on the remaining thoracic segments are thin and biramous, with the long exopods on the more anterior ones generating a respiratory current under the carapace and past the reduced posterior exopods. Each of the first five abdominal segments bears a pair of biramous flattened pleopods (abdominal appendages used for swimming) except in Spelaeogriphus where the last pair is rudimentary. The last abdominal segment carries a tail-fan of paired uropods and terminal telson. Females incubate 10-16 eggs in four or five pairs of oostegites, ventral plates between the thoracic legs (as in most Peracarida). Males and females differ only in features of the antennae and pleopods, which are not the same in all genera. The food consists of detritus manipulated by the mouthparts. The animals creep over the bottom with their walking legs and swim by undulations of the whole body. All extant species are inhabitants of fresh water in subterranean environments, caves, or aquifers. It is hypothesized that the group arose in the sea but dispersed to Gondwanan continents (the continents that formed from the fragmentation of the ancient southern supercontinent Gondwana, including South America, Africa, Antarctica, and Australia), where they were stranded in localized fresh-water refugia (areas that escaped the great changes that occurred in the regions as a whole, providing conditions for relic colonies to survive). Because all species are found in such small areas and are dependent on the persistence of ground water, it could be advocated that they are critically endangered. See PERACARIDA. Gary C. B. Poore

Bibliography. I. Gordon, On Spelaeogriphus, a new cavernicolous crustacean from South Africa, Bull. Brit. Mus. Nat. Hist. (Zool.), 5:31-47, 1957; J. R. Grindley and R. R. Hessler, The respiratory mechanism of Spelaeogriphus and its phylogenetic significance (Spelaeogriphacea), Crustaceana, 20:141-144, 1971; A. M. S. Pires, Potiicoara brasiliensis: A new genus and species of Spelaeogriphacea (Crustacea: Peracarida) from Brazil with a phylogenetic analysis of the Peracarida, J. Nat. Hist., 21:225-238, 1987; G. C. B. Poore and W. F. Humphreys, First record of Spelaeogriphacea from Australasia: A new genus and species from an aquifer in the arid Pilbara of Western Australia, Crustaceana, 71:721-742, 1998; Y.-b. Shen et al., A new spelaeogriphacean (Crustacea: Peracarida) from the Upper Jurassic of China, Contrib. Zool., 68:19-35, 1998.

Sperm cell

The male gamete. The typical sperm of most animals has a head containing the nucleus and acrosome, a middle piece with the mitochondria, and a tail with the 9 + 2 microtubule pattern (**Fig. 1**). Sperm, as well as the acrosome shape, varies with the species



Fig. 1. Diagram of a human spermatozoon, based on electron micrographs.

(**Fig. 2**). The nucleus consists of condensed chromatin (deoxyribonucleic acid, DNA) and histone proteins. The acrosome, which is derived from the Golgi complex, contains hydrolytic enzymes, that is, hyaluronidase capable of lysing the egg coats at fertilization. Actin molecules which aid in the interaction between sperm and egg are found in the area between the acrosome and nucleus. The mitochondria in the middle piece apparently provide the energy necessary for the motility created by the tail. The tail has a central core, or axial filament, made up of nine double tubules and two central tubules. *See* CILIA AND FLAGELLA.

Many groups, including nematodes, myriapods, and crustaceans, have atypical sperm which lack a flagellum and are presumably nonmotile. The sperm of *Ascaris* is round and moves by ameboid means. The crustacean sperm have a large acrosome of several components. In the anomurans a middle with mitochondria and arms filled with microtubules precedes the nucleus. In the true crabs the nucleus forms arms which possess microtubules and also surrounds the many component acrosomes. The nucleus of crustacean sperm does not condense with



Fig. 2. Forms of spermatozoa. (After C. R. Austin, Fertilization, Prentice-Hall, 1965)

maturation as it does in typical sperm mentioned above. *See* SPERMATOGENESIS. Gertrude W. Hinsch

Spermatogenesis

The differentiation of spermatogonial cells into spermatozoa. This process occurs within the testes of the male (see **illus.**). The primordial germ cells are designated spermatogonia when the gonad is identifiable as a testis.

Spermatogonial divisions occur continuously throughout the life of mammals; these divisions both maintain the stem cell population (spermatogonial cells) and supply cells which develop into sperm. Clusters of spermatogonia maintain communication through cytoplasmic bridges, and these groups become primary spermatocytes when they synchronously enter the first meiotic prophase. The first meiotic prophase is characterized by a series of remarkable changes in chromosome morphology, which are identical to those seen in the corresponding stage of oogenesis. The secondary spermatocyte produced by this division then undergoes a division in which the chromosomes are not replicated; the resulting spermatids contain half the somatic number of chromosomes. *See* MEIOSIS.

The spermatids become embedded in the cytoplasm of Sertoli cells, and there undergo the distinctive changes which result in formation of spermatozoa. These morphological transformations include the conversion of the Golgi apparatus into the acrosome and progressive condensation of the chromatin in the nucleus. A centriole migrates to a position



Cellular events in human spermatogenesis. Cell types are labeled.

distal to the nucleus and begins organizing the axial filament which will form the motile tail of the sperm. Mitochondria may fuse to form a nebenkern as is the case for many vertebrates, or there may be less extensive fusion as in mammals. In all cases the resulting structures become located around the axial filament in the midpiece. The cytoplasm of the spermatid is reflected distally away from the nucleus during spermatid maturation; eventually, most of the cytoplasm is sloughed off and discarded (see illus.).

The Sertoli cells are thought to provide nutrition for the developing sperm, because their cytoplasm contains large stores of glycogen which diminish as spermatids mature. There is no direct evidence for this nutritive function, but some forms of male sterility are associated with the failure to produce normal Sertoli cells. Electron microscopy has revealed distinct plasma membranes surrounding the two cell types at the points of contact, and thus the Sertoli cell-spermatid relationship is not syncytial as once thought.

Spermatogenesis is cyclical to a varying extent depending on the species, and under endocrine control. The periodic function of the human testis is not nearly as clearly definable as the menstrual cycle in the female, and can only be detected by examining deoxyribonucleic acid (DNA) synthesis and the changes in proportions of the various stages present. In seasonal-breeding animals, the testis regresses, and spermatogenesis ceases during the quiescent season. Spermatogenesis is maintained and regulated by male steroid hormones such as testosterone, which is produced by the interstitial or Leydig cells found in the connective tissue of the testis. Interstitial cells, in turn, are stimulated by luteinizing hormone (LH) which is produced by the pituitary gland. The male testis-regulating hormone was formerly known as interstital cell-stimulating hormone (ICSH), but it is now known to be identical to LH. *See* ENDOCRINE MECHANISMS; GAMETOGENESIS; MEN-STRUATION; TESTIS.

Sphaeractinoidea

An extinct group of fossil organisms generally referred to as the Mesozoic stromatoporoids. As originally proposed by Othmar Kühn in 1927, the order included the Mesozoic stromatoporoids (Sphaeractiniidae), the heterastridiids (a group of obscure spherical Mesozoic fossils commonly regarded as hydrozoans), and the spongiomorphs. Kühn placed all of these in the class Hydrozoa. The discovery of spicules (spikelike supporting structures) in some genera of the sphaeractinids has allowed some genera formerly considered Mesozoic stromatoporoids to be classified in the Porifera as representatives of the class Demospongiae and the subclasses Ceractinomorpha and Tetractinomorpha. This discovery has opened the probability that all the sphaeractinids will eventually be referable to living sponge families and orders, if traces of spicules can be found in them. Both Mesozoic and Paleozoic stromatoporoids are now considered to be poriferans; the Paleozoic representatives are best considered a class of the phylum separated from the Mesozoic forms by time, lack of spicules, and microstructure. Those genera of the Mesozoic group that, at present, cannot be assigned to a living poriferan taxon owing to the lack of (or our failure as yet to discover) spicules are a disparate orphan group that can best be informally referred to as sphaeractinids or Mesozoic stromatoporoids.

These Mesozoic stromatoporoids appear in the late Permian with the genus Circopora, but did not reach abundance until the Jurassic and early Cretaceous. They declined rapidly in late Cretaceous time, and their presence in younger beds is in doubt. They grew large, calcareous, internally laminate skeletons in encrusting, dome-shaped, bulbous, fingerlike, and branched forms. Internally, the skeleton is composed of an irregularly repetitive three-dimensional network of longitudinal pillars or walls, tangential laminae with beams and cyst plates (see illus.) much like that of their Paleozoic look-alikes. In some genera, longitudinal canals crossed by tabulae are prominent. The microstructure of the structural elements is fibrous in well-preserved fossils and has suggested to some that they secreted aragonite rather than calcite. However, some living sponges secrete calcite in a similar microstructure, and it may be that the Mesozoic stromatoporoids did the same.



Sphaeractinoid Actinostromaria rhodaclada, (a) tangential and (b) longitudinal section, \times 6. (From Steinor, 1932, Cretaceous, Jura Mountains)

The Mesozoic stromatoporoids are important reefbuilding organisms in the Middle East, particularly in areas that surrounded the Mesozoic Tethys seaway. They are known from the Permian of Pakistan (Salt Range); the Triassic of the Caucasus, Bulgaria, and Indonesia; the Jurassic of Israel, Arabia, the Middle East-Mediterranean region, and North America (the eastern continental shelf); and the Lower Cretaceous of the Mediterranean region. *See* DEMOSPONGIAE; PORIFERA; REEF; SPONGIOMOR-PHIDA; STROMATOPOROIDEA. C. W. Stearn

Bibliography. C. W. Stock et al., Annotated checklist of post-Devonian "Stromatoporoid genera" and their type species, *Fossil Cnidaria Porifera Newsl.*, 21(1.2):1-22, 1992; D. Turnšek and A. Barbulescu, Upper Jurassic hydrozoa in central Dubrogea (Romania), *Geologija*, 12:73-84, 1969; R. A. Wood, Biology and revised systematics of some late Mesozoic stromatoporoids, *Spec. Pap. Palaeontol.*, no. 37, pp. 1-89, 1987; R. A. Wood and J. Reitner, Poriferan affinities of Mesozoic stromatoporoids, *Palaeontology*, 29:469-478, 1986.

Sphaerocarpales

An order of liverworts in the subclass Marchantiidae, consisting of two families, the terrestrial Sphaerocarpaceae (*Sphaerocarpos* and *Geothallus*) and the aquatic Riellaceae (*Riella*). The plants are characterized by envelopes surrounding each antheridium and archegonium, absence of elaters, poor development of seta, and absence of thickenings in the unilayered wall of an indehiscent capsule.

The gametophytes consist of a simple, lobed thallus of rosette form, or a stemlike axis with lateral lobes or appendages. The rhizoids are smooth, and air chambers and pores are lacking. Oil bodies may be present. The antheridia are separately contained within receptacles on the dorsal surface or at the margins, while the archegonia are separately produced in receptacles dorsally or behind the growing point. The sporophyte consists of a bulbous foot, short or obsolescent seta, and globose indehiscent capsule with a unilayered wall. There are no elaters, but minute green cells are mingled with the spores. The haploid chromosome number is 8 or 9. *See* BRYOPHYTA; HEPATICOPSIDA; MARCHANTIIDAE.

Howard Crum

Sphagnopsida

A class of the plant division Bryophyta containing plants commonly called peatmosses. The spongelike plants (Fig. 1) grow as perennials in soft cushions or lawns in wet habitats (rarely they grow submerged). The class consists of a single genus, Sphagnum, of some 200 species. The thallose protonema, fascicled branches, dimorphous leaf cells, and spores developed from amphithecial tissue are characters unique to the class. The plants are ecologically important, owing to the spongelike construction of leaves and outer cells of stems and branches, and the ability, whether living or dead, to create acid conditions by exchanging hydrogen ions for cations in solution. The dried plant parts are used as a mulch in horticultural practice, and as fuel where the plant is abundant.

Peatmoss stems are simple or forked, with branches usually of two kinds, some spreading and others pendant, in fascicles spirally arranged around the stem and crowded in a headlike apex. Rhizoids (present only in early stages) are multicellular with slanted crosswalls. The stems and branches are enveloped in a cortex of large, empty, thin-walled hyaline cells, which often have pores to the outer



Fig. 1. Sphagnobrya. (a) Sphagnum capillaceum. (b) Sphagnum palustre. (After E. L. Palmer, Fieldbook of Natural History, McGraw-Hill, 1949)

surface. Internal to the cortex is a wood cylinder of long, thick-walled cells surrounding an inner tissue of more or less thin-walled and undifferentiated cells.

The stem and branch leaves are generally differentiated. The branch leaves consist of a network of linear green cells (chlorophyllose cells) enclosing large, empty hyaline cells that are nearly always reinforced internally by annular fibrils and perforated by pores on the free surface (**Fig. 2**). The globose, stalked antheridia occur among the leaves of somewhat differentiated spreading branches, and are often catkinlike and highly colored; one antheridium is produced at the side of each leaf. The archegonia are few and flask-shaped on very short branches and enclosed by large perichaetial leaves.



Fig. 2. Cellular structure of a leaf. (After F. W. Emerson, Basic Botany, 2d ed., McGraw-Hill, 1954)



Fig. 3. Sporophyte of Sphagnum. (After F. W. Emerson, Basic Botany, 2d ed., McGraw-Hill, 1954)



Fig. 4. Stages in the development of a protonema. (a) Germination of spore; (b) division of terminal cell to form an apical cell; (c) formation of multicellular rhizoids (*after A. Engler and K. Prantl, Die natürlichen Pflanzenfamilien, Bd. 10, 2 Aufl., 1924*). (d) Thalloid protonema with young, upright leafy branch forming (*after F. W. Emerson, Basic Botany, 2d ed., McGraw-Hill, 1954*).

The sporophytes are anchored by a foot, and are raised beyond the perichaetial leaves at maturity by an extension of the gametophyte called a pseudopodium (Fig. 3). The globose capsules dehisce by a lidlike operculum. The capsule wall consists of several layers of cells forming a solid tissue, with large numbers of nonfunctional pseudostomata (having two guard cells but no pore) scattered over the surface. The spores are derived from the amphithecium, or outer portion of the capsule. The spores are tetrahedral, with triradiate scars marking a tetrad origin. A membranous calyptra invests the capsule until its maturity. The protonema is normally a small, lobed thallus producing one or rarely two leafy plants but sometimes proliferating marginal filaments that produce at their tips secondary protonemata (Fig. 4). The haploid chromosome number is 19 and 38. See BRYOPHYTA. Howard Crum

Bibliography. H. Crum, Sphagnopsida, Sphagnaceae, *No. Amer. Flora* 2, 11:1–180, 1984; B. Lange, Key to the northern boreal and arctic species of *Sphagnum*, based on characteristics of the stem leaves, *Lindbergia*, 8:1–29, 1982; D. H. Vitt and R. E. Andrus, The genus *Sphagnum* in Alberta, *Can. J. Bot.*, 55:331–357, 1977.

Sphalerite

A mineral, β -ZnS, also called blende. It is the lowtemperature form and more common polymorph of ZnS. Pure β -ZnS on heating inverts to wurtzite, α -ZnS, at 1020°C (1868°F), but this temperature can be lowered substantially by impurity-atom solid solution (especially Cd²⁺ and Fe²⁺) and sulfur fugacity. Sphalerite crystallizes in the hextetrahedral class of the isometric system with a structure similar to that of diamond. The space group is F43m, and the cubic unit cell has an edge a = 0.543 nanometer, which contains four ZnS molecules. Zinc atoms occupy the positions of half the carbon atoms of diamond, and sulfur atoms occupy the other half. Each zinc atom is bonded to four sulfur atoms, and each sulfur atom is bonded to four zinc atoms. The



Sphalerite. (a) Crystals in limestone from Joplin, Missouri (specimen from Department of Geology, Bryn Mawr College). (b) Crystal habit (after C. Klein and C. S. Hurlbut, Jr., Manual of Mineralogy, 21st ed., John Wiley and Sons, 1993).

common crystal forms of sphalerite are the tetrahedron, dodecahedron, and cube, but crystals are frequently complex and twinned (see **illus.**). The mineral is most commonly in coarse to fine, granular, cleavable masses. The luster is resinous to submetallic; the color is white when pure, but is commonly yellow, brown, or black, darkening with increased percentage of iron. It has been shown that excess sulfur can also contribute to the darkening of the color. There is perfect dodecahedral cleavage; the hardness is $3^{1}/_{2}$ on Mohs scale; specific gravity is 4.1 for pure sphalerite but decreases with increasing iron content.

Pure sphalerite contains 67% zinc and 33% sulfur, but iron is usually present, substituting in the structure for zinc and may amount to 36%. The amount of iron in sphalerite varies directly with temperature at the time of crystallization. Cadmium and manganese may be present in small amounts.

Sphalerite is a common and widely distributed mineral associated with galena, pyrite, marcasite, chalcopyrite, calcite, and dolomite. It occurs both in veins and in replacement deposits in limestones. As the chief ore mineral of zinc, sphalerite is mined on every continent. The United States is the largest producer, followed by Canada, Mexico, Russia, Australia, Peru, the Congo, and Poland. *See* ZINC. Cornelius S. Hurlbut, Jr.; Paul B. Moore

Sphenisciformes

The penguins, a small monotypic order of flightless marine swimming birds found in the colder southern oceans (see illustration). They have most likely evolved from the order Procellariiformes, perhaps from a diving petrel-like ancestor. Classification schemes that hypothesize a link between penguins and loons have no factual support. At one time, the Sphenisciformes had been placed in a separate superorder, Impennes, but such a designation suggested too great a divergence of the penguins from their ancestral group of tube-nosed swimmers. Some taxonomists have enlarged the Ciconiiformes to include the Sphenisciformes penguins, as in the Sibley-Ahlquist taxonomy (a radical new approach to taxonomy, based on DNA hybridization studies). See CI-CONIIFORMES; GAVIIFORMES; PROCELLARIIFORMES.

Fossil record. The fossil record of the penguins begins in the late Eocene and continues throughout the Tertiary. Fossils have been found in Antarctica, Australia, and South Africa, which approximates the range of present-day penguins. The earliest fossil penguins were already well specialized as flightless, wing-propelled swimming birds, and their fossil history provides little new information about the evolutionary history and distribution of the group. Some fossil penguins were much larger than the emperor penguin, the largest living species. Some Eocene species were gigantic and had a long spearlike bill.

Characteristics. Penguins are medium-sized to large birds. They are completely flightless, their wings having been modified into stiff flattened flippers. They stand upright on legs that are placed far posterior and that terminate in four toes, the anterior three of which are webbed. Penguins swim and dive well, using only their wings for propulsion: their feet are used only for steering during swimming. Terrestrial locomotion is by walking, hopping, or sliding on the belly while pushing with the wings. The plumage consists of dense scalelike feathers that are black dorsally and white ventrally. A distinctive pattern or crest, often yellow, occurs on the head.

Penguins are gregarious, breeding in large colonies along the coast. The males and females, which are identical, form strong pair bonds and share in the incubation and care of the downy nestlings. The older young of some species are kept in large groups, or creches. The emperor penguin (*Aptenodytes forsteri*) breeds on the ice pack along the Antarctic coasts during the fall. Incubation of the one egg is the responsibility of the male, which remains on the nest for over 2 months in the winter without eating. The females leave the breeding colony during this period to feed, and return fat and laden with



Aptenodytes patagonicus, the king penguin. (Photo by Gerald and Buff Corsi, © California Academy of Sciences).

food for the young. Males then leave for the ocean and return with food to relieve the females. This alternation continues until the young are old enough to enter the water. By this time, the sea ice has melted back close to site of the nesting colony.

Habitat. Penguins are found only in the cold southern oceans, on the Antarctic continent and its surrounding islands and northward to Australia, New Zealand, South America, and Africa. One species, the Galápagos penguins (*Spheniscus mendiculus*), is found in the Galápagos Islands, which are on the Equator but are surrounded by the cold Humboldt Current. *See* AVES. Walter J. Bock

Bibliography. L. S. David and M. Renner, *Penguins*, Yale University Press, 2004; L. Davis and J. Darby (eds.), *Penguin Biology*, Academic Press, 1990; T. D. Williams, *The Penguins*, Oxford University Press, 1995.

Sphenophyllales

An extinct order of articulate land plants, common during Late Pennsylvanian and Early Permian times. They are typified by Sphenophyllum (see illus.), a small, branching plant, probably of trailing habit. The long, jointed stems rarely exceeded 0.4 in. (1 cm) in diameter and had superposed, longitudinal, surficial ribs between nodes. The vascular system contained a solid xylem core with triangular primary wood. The leaves were wedge-shaped, usually shorter than 0.8 in. (2 cm), and had toothed, notched, or rounded distal margins. They were attached at the nodes by their narrow ends, in whorls of usually 6 or 9 leaves each, rarely 18. Long, terminal cones, usually called Bowmanites when found detached, contained sporangia and spores. The sporangia terminated slender stalks, forming concentric whorls that alternated



Reconstruction of Sphenophyllum cuneifolium, showing whorled leaves and two cones. (After G. M. Smith, Cryptogamic Botany, vol. 2, 2d ed., McGraw-Hill, 1955)

with whorls of sterile bracts. Most species were homosporous (produced spores of a single type). *See* PALEOBOTANY; PLANT KINGDOM; SPHENOPHYTA. Sergius H. Mamay

Sphenophyta

One of the major divisions (formerly known as Equisetophyta) of vascular plants that includes both living and fossil representatives. The three principal orders are Pseudoborniales (Devonian), Sphenophyllales (Devonian-Triassic), and Equisetales (Devonian-Recent); the Hyeniales (Devonian) may also be sphenophytes. *See* EQUISETALES; HYENIALES; PSEU-DOBORNIALES; SPHENOPHYLLALES.

All sphenophytes are characterized by axes with distinct nodes that produce whorls of small leaves or branches; branches often contain longitudinal ribs and furrows. Internally the stems of sphenophytes are characterized by longitudinally oriented canals, some of which functioned in gaseous exchange. Secondary tissues were produced in a few fossil forms. The reproductive organs of this group are loosely arranged strobili or cones consisting of a central axis bearing whorls of modified branches, each terminating in a recurved, thick-walled sporangium. Most sphenophytes produced one type of spore (homospory), although a few fossil forms were heterosporous.

Pseudoborniales are known from a single species collected from a few Devonian localities south of Spitsbergen and in Alaska. Impression fossils suggest that the plant was a tree about 66 ft (20 m) tall, constructed of whorls of branches and leaves. The reproductive parts of *Pseudobornia* are not well known, but appear to contain recurved sporangia with one type of spore.

Sphenophyllales, less than 3 ft (1 m) tall, formed a portion of the Carboniferous forest understory. The leaves were often wedge shaped, and many were produced in a whorl. Some forms produced secondary tissues in the stems. All were homosporous, with the sporangia produced in cones.

During the Carboniferous, the sphenophytes were represented by numerous species in the order Equisetales, ranging from herbaceous forms to trees. Both homosporous and heterosporous forms were present, but the spores of all species contained extra exinous structures termed elaters. During the Mesozoic Period, the group declined, with fewer, smaller plants.

Today all that remains of the sphenophytes is the genus *Equisetum*, commonly called the horsetail or scouring rush. Except in Australia and New Zealand, the members are worldwide in distribution and typically grow in damp habitats along the edges of streams, although some species are adapted to mesic conditions. A few species attain considerable size, but none of these produce secondary tissues. All species are characterized by one type of spore (homospory), each bearing two pairs of elaters on the surface. *See* EMBRYOBIONTA.

Bibliography. T. N. Taylor and E. L. Taylor, *The Biology and Evolution of Fossil Plants*, 1993.

Sphere

Both in euclidean solid geometry and in common usage the word sphere denotes a solid of revolution obtained by revolving a semicircle of radius *r* about its diameter. Its total volume is $V = 4/3\pi r^3$.

However, in analytic geometry, and more generally in modern mathematics, the word sphere denotes a spherical surface that bounds a solid sphere. In this sense a sphere is the locus of all points P in threedimensional space whose distance from a fixed point O (called the center) is equal to a given number. The word radius may refer either to one of the segments OP, or to their common length r. A plane that intersects a sphere in just one point is called a tangent plane and is perpendicular to the radius drawn from the center of the sphere to that point. A plane that intersects a sphere in more than one point intersects it in a circle. The circle is called a great circle or a small circle of the sphere according to whether the plane does or does not pass through the center of the sphere. If two parallel planes intersect a sphere (see illus.), the spherical surface between them is called a zone, and its area is $2\pi rb$, where the height b of the zone is the distance between the parallel planes. The volume bounded by the zone and the two base planes is called a spherical segment, and is measured by the formula $V = (\pi b/6)(3r_1^2 + 3r_2^2 + b^2)$, where r_1 and r_2 are the radii of the two base circles. It is also measured by the prismoidal formula $V = (b/6)(B_1 +$ $4M + B_2$), where B_1 , B_2 denote the areas of the two bases, and M the area of the midsection halfway between the bases.

Any great circle of a sphere divides it into two hemispheres. A second great circle cuts a hemisphere into two lunes, each having vertices at the two points where the great circles intersect and each having an area proportional to the angles at the vertices between the tangents to the great circle boundaries. A third great circle cuts each lune into two spherical triangles. In general a spherical triangle is the figure formed by connecting any three points *A*, *B*, *C* on the surface of a sphere (but not on the same great circle) by great circle arcs called sides. The sides *BC*, *CA*, *AB* are measured by the angles *a*, *b*, *c* that they subtend at the center of the sphere. If the angles between the pairs of sides at the vertices *A*, *B*, *C* are denoted by α , β , and γ , respectively, and



Segment of a sphere formed by intersection of two parallel planes with sphere.

if $2\sigma = \alpha + \beta + \gamma - 180^{\circ}$ denotes the excess of the sum of these angles over two right angles, then the area of a spherical triangle of excess 2σ on a sphere of radius *r* is $(\sigma/90^{\circ})\pi r^2$ (or $2\sigma^2$ if σ is measured in radians). Important relations between the sides and angles of a spherical triangle are as follows:

Law of sines:
$$\frac{\sin \alpha}{\sin a} = \frac{\sin \beta}{\sin b} = \frac{\sin \gamma}{\sin c}$$

Law of cosines:

 $\cos a = \cos b \cos c + \sin b \sin c \cos \alpha$

J. Sutherland Frame

Spherical harmonics

A spherical harmonic or solid spherical harmonic of degree n is a homogeneous function, $R_n(x,y,z)$, of degree n which satisfies Laplace's equation (1).

$$\Delta R \equiv \frac{\partial^2 R}{\partial x^2} + \frac{\partial^2 R}{\partial y^2} + \frac{\partial^2 R}{\partial z^2} = 0 \tag{1}$$

Here *n* is any number and

$$(x^{2} + y^{2} + z^{2})^{(-n-1)/2} R_{n}(x, y, z)$$

is a spherical harmonic of degree -n - 1. There are analogous definitions for spaces of any number of dimensions. In the present article, n will be a nonnegative integer and R_n a polynomial in x_iy_iz (polynomial spherical harmonic). In terms of spherical coordinates r, θ , ϕ , $R_n(x_iy_iz) = r^nS_n(\theta,\phi)$ where S_n , a polynomial in $\cos \theta$, $\sin \theta$, $\cos \phi$, $\sin \phi$ is a spherical surface harmonic of degree n. There are 2n + 1linearly independent spherical surface harmonics of degree n; any spherical surface harmonic of degree n is a linear combination of these, and conversely any linear combination of spherical surface harmonic of degree n is again a spherical surface harmonic

Applications. Spherical harmonics occur in potential theory. They occur in connection with Laplace's equation not only in spherical coordinates but also in spheroidal coordinates (spheroidal harmonics) and confocal coordinates (ellipsoidal surface harmonics). In the latter case their occurrence is due to the circumstance that the natural affine mapping of an ellipsoid onto a sphere carries the partial differential equation of ellipsoidal surface harmonics into the partial differential equation satisfied by spherical surface harmonics. In spherical coordinates, spherical surface harmonics occur in connection with Laplace's and Poisson's equations, the wave equation, the Schrödinger equation, and generally in connection with partial differential equations of the form $\Delta U + f(r)U = 0$. In the latter case one has special solutions of the form $F(r)S_n(\theta,\phi)$, where F satisfies the ordinary differential equation (2). In geometry,

$$\frac{d^2F}{dr^2} + \frac{2}{r}\frac{dF}{dr} + \left[f(r) - \frac{n(n+1)}{r^2}\right]F = 0$$
(2)

spherical surface harmonics are used in the theory of

surfaces. In mathematical physics, spherical harmonics appear in the theories of gravitation, electricity and magnetism, hydrodynamics, and in other fields.

Spherical harmonics of degree n. Notation (3), with

$$\int_{-\pi}^{\pi} (x\cos u + y\sin u + iz)^n f(u) \, du \qquad (3)$$

f(u) an integrable function, is a polynomial spherical harmonic of degree n, and every such spherical harmonic can be so represented. The representation is not unique. If c_n is a constant, b_1, b_2, \ldots, b_n are n directions (not necessarily distinct), and $\partial/\partial b$ denotes directional differentiation in the direction b, then notation (4) is a polynomial spherical harmonic of de-

$$c_n r^{2n+1} \frac{\partial^n}{\partial b_1 \cdots \partial b_n} \frac{1}{r} \tag{4}$$

gree *n*, and every such spherical harmonic can be so represented. The representation is unique. In a zonal spherical harmonic the *n* directions coincide, and in a sectorial spherical harmonic they are in a plane at angles of π/n . If n - m directions coincide in the axis and the remaining directions are in the plane perpendicular to the axis at angles of π/m , one has a tesseral spherical harmonic of degree *n* and order *m*.

Explicit forms. For spherical harmonics whose axis is the *z* axis (m = 0, 1, 2, ..., n), Eq. (5) represents

$$S_n^{\pm m}(\theta,\phi) \equiv \frac{(-1)^{n-m}r^{n+1}}{(n-m)!} \frac{\partial^{n-m}}{\partial z^{n-m}} \left(\frac{\partial}{\partial x} \pm i\frac{\partial}{\partial y}\right)^m \frac{1}{r}$$
$$= P_n^m(\cos\theta)e^{\pm im\phi} \tag{5}$$

a linearly independent system of spherical surface harmonics of degree n. $S^m_n \pm S^{-m}_n$ is a zonal, sectorial, tesseral spherical surface harmonic of degree nand order m according as m = 0, m = n, $1 \le m \le$ n - 1. The $P_n^m(w)$ are associated Legendre functions which satisfy the associated Legendre equation (6).

$$(1 - w^{2})\frac{d^{2}P}{dw^{2}} - 2w\frac{dP}{dw} + \left[n(n+1) - \frac{m^{2}}{1 - w^{2}}\right]P = 0 \quad (6)$$

 $P_n^0 = P_n$ is the Legendre polynomial of degree *n*.

Properties. A function is said to be harmonic in a region if it is a twice continuously differentiable solution of Laplace's equation there, and if, in addition, it vanishes at infinity in case the point at infinity is an interior point of the region. Every function harmonic inside a sphere about the origin can be expanded in a series

$$\sum_{n=0}^{\infty} r^n S_n(\theta, \phi)$$

convergent inside that sphere. Every function harmonic outside a sphere about the origin can be expanded in a series

$$\sum_{n=0}^{\infty} r^n S_n(\theta, \phi)$$

convergent outside that sphere. The reciprocal distance of the points (x,y,z) and (0, 0,a) is harmonic in the regions r < a and r > a and possesses in these regions the expansions of Eqs. (7).

$$\frac{1}{\sqrt{a^2 - 2ar\cos\theta + r^2}} = \begin{cases} \sum_{n=0}^{\infty} \frac{r^n}{a^{n+1}} P_n(\cos\theta) & r < a \\ \sum_{n=0}^{\infty} \frac{a^n}{r^{n+1}} P_n(\cos\theta) & r > a \end{cases}$$
(7)

Here θ , ϕ determine a point on the unit sphere. The scalar product of two functions, *f* and *g*, on the unit sphere is suitably defined as Eq. (8) where \overline{g}

$$(f,g) = \int_0^{\pi} \int_{-\pi}^{\pi} f(\theta,\phi) \overline{g(\theta,\phi)} \sin\theta \ d\theta \ d\phi \qquad (8)$$

is the complex conjugate of g. If (f,g) = 0, f and g are orthogonal. Spherical surface harmonics are functions on the unit sphere. Any two spherical surface harmonics of different degrees are orthogonal. The spherical surface harmonics of Eq. (9) form

$$S_n^m(\theta,\phi)m = -n, -n+1, \dots, n \tag{9}$$

$$n=0,1,\ldots$$

an orthogonal system; that is, $(S_n^m, S_n^{m'}) = 0$ unless m = m' and n = n'. This orthogonal system is complete; that is, a continuous function which is orthogonal to all the S_n^m vanishes identically. With an integrable function *f* is associated the Laplace expansion (10)

$$\sum_{n=0}^{\infty} \sum_{m=-n}^{\infty} c_{mn} S_n^m \tag{10}$$

where

Under suitable conditions, the Laplace expansion will converge to f. For instance, if f is continuous and continuously differentiable on the unit sphere, then the Laplace expansion converges to f uniformly.

 $C_{mn} = \frac{\left(f, S_n^m\right)}{\left(S_n^m, S_n^m\right)}$

Let (θ_0, ϕ_0) be a fixed point, and let $\cos \gamma = \cos \theta \cos \theta_0 + \sin \theta \sin \theta_0 \cos (\phi - \phi_0)$ be the spherical distance of (θ, ϕ) and (θ_0, ϕ_0) . The Laplace expansion of Eq. (11) is the addition theorem of Legendre

$$P_n(\cos\gamma) = P_n(\cos\theta)P_n(\cos\theta_0) + 2\sum_{m=-n}^n \frac{(n-m)!}{(n+m)!}$$

 $\cdot P_n^m(\cos\theta)P_n^m(\cos\theta_0)\cos m(\phi-\phi_0) \quad (11)$

polynomials and expresses the change to a new axis through the point (θ_0, ϕ_0) . Other spherical surface harmonics have corresponding addition theorems.

Let $\cos \gamma$ be the spherical distance of (θ, ϕ) and (θ_0, ϕ_0) , and let k(w) be a continuous function for $-1 \le w \le 1$. Then for any spherical surface harmonic

 S_n of degree n, Eq. (12) holds.

$$\int_{0}^{\pi} \int_{-\pi}^{\pi} K(\cos \gamma) S_{n}(\theta, \phi) \sin \theta \ d\theta \ d\phi$$
$$= \lambda_{n} S_{n}(\theta_{0}, \phi_{0}) \quad (12)$$

 $\lambda_n = 2\pi \int_{-1}^{1} K(w) P_n(w) \, dw$

where

See DIFFERENTIAL EQUATION.

A. Erdélyi

Bibliography. A. Erdélyi et al., *Higher Transcendental Functions*, 3 vols., 1953-1955, reprint 1981; E. W. Hobson, *Spherical and Ellipsoidal Harmonics*, 1955; M. N. Jones, *Spherical Harmonics and Tensors for Classical Field Theory*, 1985; E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, 4th ed., 1927.

Sphingolipid

Any lipid containing the long-chain amino alcohol sphingosine (structure 1) or a variation of it, such as dihydrosphingosine, phytosphingosine (structure 2), or dehydrophytosphingosine. Sphingo-

$$CH_{3}(CH_{2})_{12}CH = CH - CH - CH - CH - CH_{2}OH | | | OH NH_{2} (1) CH_{3}(CH_{2})_{13}CH - CH - CH - CH_{2}OH | | | | OH OH NH_{2} (2)$$

sine itself is synthesized by condensing a long-chain fatty acid with the amino acid serine.

Sphingosine is converted into a variety of derivatives to form the family of sphingolipids. The simplest form is a ceramide which contains a sphingosine and a fatty acid residue joined by an amide linkage. The structure of the ceramide *N*-acylsphingosine is shown in structure **3**.



Ceramide is the basic building block of practically all of the naturally occurring sphingolipids. Ceramide can be further modified by the addition of a phosphorylcholine at the primary alcohol group to form sphingomyelin (structure 4), a ubiquitous phospholipid in the plasma membranes of virtually all cells. Modification of a ceramide by addition of one or more sugars at the primary alcohol group converts it to a glycosphingolipid, which occurs widely in both the plant and animal kingdoms. The simplest form of a glycosphingolipid is a cerebroside which usually contains a glucose or a galactose residue joined to the ceramide moiety through a glycosidic linkage. The carbohydrate unit can be further extended to form more complex structures. A notable example of modification is the addition of one or more sialic acid residues to an existing carbohydrate chain to form the family of gangliosides which are particularly abundant in the nervous system. The structure of a typical ganglioside, GM1, is shown (structure 5). See GLYCOSIDE; LIPID.



In addition to being important structural components of cellular membranes, sphingolipids participate in diverse cellular functions. A number of inheritable diseases that can cause severe mental retardation and early death occur as the result of a deficiency in one or more of the degradative enzymes, resulting in the accumulation of a particular sphingolipid in tissues. These diseases are collectively called sphingolipidoses and include Niemann-Pick disease, Gaucher disease, Krabbe disease, metachromatic leukodystrophy, and several forms of gangliosidoses, such as Tay-Sachs disease. Functionally, glycosphingolipids are known to serve as important cell-surface molecules for mediating cell-to-cell recognition, interaction, and adhesion. They also serve as receptors for a variety of bacterial and viral toxins. Many glycosphingolipids can modulate immune responses as well as the function of hormones and growth factors by transmitting signals from the exterior to the interior of the cell. A number of glycolipids are also found to participate in a variety of immunological disorders by serving as autoantigens. Other sphingolipids and their metabolites, notably ceramide, sphingosine-1phosphate, sphingosylphosphorylcholine, and NNand N,N,N-methylsphingosine, may serve as second messengers in several signaling pathways that are important to cell survival or programmed cell death (apoptosis). See AUTOIMMUNITY; METABOLIC DISOR-Robert K. Yu DERS.

Spica

A 1st-magnitude star in the constellation Virgo. Spica (α Virginis) is a hot main-sequence star of spectral type B1, effective temperature about 25,000 K (45,000°F), at a distance of 80 parsecs from the Sun (2.59 × 10¹⁵ km or 1.61 × 10¹⁵ mi). *See* SPECTRAL TYPE.

It is a double-line spectroscopic binary with an orbital period of almost exactly 4 days. The pair has been resolved with an intensity interferometer, which allows an accurate determination of the fundamental parameters of the system. The primary has a mass that is 10.8 times that of the Sun, and a radius of 8.1 solar radii; the mass and radius of the secondary are 6.8 and 4.4, respectively, in solar units. The system constitutes an example of an ellipsoidal variable, where the separation between the two stars is so small that powerful tidal forces distort the stars. The changing aspect of the primary as seen from Earth causes light variations with a period equal to the orbital period. The separation between the two stars is only about 3.5 times the radius of the primary. Both components are rotating around their axes very rapidly at projected equatorial velocities of 160 km/s (100 mi/s) and 100 km/s (60 mi/s), respectively. See BINARY STAR; ECLIPSING VARIABLE STARS; STELLAR RO-TATION.

The primary itself is an intrinsic variable of the β Cephei type and is pulsating with a period of about 4 h. As a result, the observed light curve is a superposition of two different effects. There are also changes in the profiles of the spectral lines, which are related to the same mechanisms. *See* OCCULTATION; STAR; VARIABLE STAR. David W. Latham

Spice and flavoring

Ingredients added to food to provide all or a part of the flavor. Spices, a unique category of flavorings which are given preferred legal status because of the long history of their use in foods, are pungent or aromatic substances of vegetable origin used in foods at levels that yield no significant nutritive value. Flavor is the perception of those characteristics of a substance taken orally that affect the senses of taste and olfaction. The term flavoring refers to a substance which may be a single chemical species or a blend of natural or synthetic chemicals whose primary purpose is to provide all or part of the particular flavor effect to any food or other product taken orally. Flavorings are categorized by source: animal, vegetable, mineral, and synthetic.

Chemical nature of flavor. The sensation of flavor has a chemical basis that involves both taste and smell.

Taste. The elements of taste consist of sourness, caused by hydrogen ions found in organic and inorganic acids; sweetness, caused by organic chemicals containing hydroxyl, carbonyl and amino groups, for example glycerol, sugars, amino acids, saccharin, hexamic acid salts; saltiness, a characteristic of sodium chloride, but also exhibited by other inorganic salts; and bitterness, which is caused by inorganic salts and some relatively high-molecular-weight organic compounds, such as alkaloids, for example, quinine, tannins, and sucrose octacetate. The taste-producing substances are usually nonvolatile.

Sensations of pain, heat, and coolness are also caused by chemicals. Several fatty acid amides of vanillylamine are responsible for the heat or bite in red peppers, while another amide, piperine, and its relatives provide the pungency of black pepper. Menthol produces a coolness at low concentrations in the mouth. The astringency that is produced by alum is also produced by tannins from tea and other botanicals.

Smell. The smell (odor) of a food is judged by the nose as volatile components of the food enter either through the nostrils or through the mouth. The chemical constituents which cause the odor component of flavor therefore are volatile. No simple characterization of basic odors has yet been found such as is available for tastes. Attempts to reduce the thousands of complex identifiable odors to a simple classification have been made, but have not been widely accepted.

Types of flavors. The application of the physical processes of distillation and extraction developed during the Middle Ages led to a wide range of essential oils, extracts, and oleoresins. Research involving the chemistry of natural products in the second half of the nineteenth century provided a nucleus

Туре	Hydrocarbons	Alcohols	Aldehydes and ketones	Acids	Esters	Bases	Sulfur compounds	Miscellaneous	Total
Apple	8	40	39	39	80	4	_	11	221
Cheddar cheese	32	11	25	28	25	14	6	12	153
Cinnamon	24	9	10	15	8	_	_	15	81
Cocoa	45	27	52	37	38	103	18	43	383
Onions and garlic	1	5	15	2	5	4	59	8	99
Potato	24	24	31	13	1	9	23	9	134
Tomato	17	31	90	12	28	4	8	9	217
Cooked beef	74	32	76	22	25	25	45	30	329
Vanilla	37	24	21	15	39	2	1	23	162
Roasted peanuts	37	25	72	29	11	63	16	25	278

of chemicals for use as flavoring ingredients. The reactions studied at that time (Knoevenagel condensation, Knorr pyrrole synthesis, Perkins reactions and rearrangements, Kolbe-Schmitt reaction) produce the characterizing ingredients of many spices and fruits. Spices contain the largest concentration of flavor volatiles and were the first to be studied. This early work confirmed the presence of, and the structure of, vanillin in vanilla beans at 2% concentration, eugenol in clove buds at 12%, and cinnamic aldehyde in cinnamon at 1–2% concentration.

The first half of the twentieth century brought a continuous effort to identify the flavorful volatiles in food by classical organic chemical methods. The many ingredients in citrus peel oils and noncitrus fruits began to yield their flavor secrets. Creative flavorists utilized this knowledge and filled in the gaps with information supplied by their noses and their palates. Reasonably representative flavors of the fruits became possible to produce. These fruits contained volatiles which effected their flavor at concentrations of 0.01–0.1% (100–1000 ppm).

The second half of the twentieth century, especially since the application of gas-liquid chromatography, has seen an explosion of information concerning the identity of the volatile components of all flavorful foods. The sensitivity of equipment for detecting ingredients and the speed with which equipment can identify these ingredients has confirmed the complex nature of flavor. Many of the volatiles in the less highly flavored foods have been examined, isolated, and identified. As many as several hundred volatile organic chemicals have been identified in common flavorful foods (**Table 1**). These components range from simple hydrocarbons to complex chemicals with sulfur or nitrogen (**Table 2**). While many of these chemicals are not present in sufficient quantities to affect the flavor and some key ingredients have not yet been identified, flavorists have used this information in attempts to prepare artificial flavors.

A so-called nature-identical chemical is selfdefining. If the flavor of foods were due to stable chemicals, flavors could be made from natural-identical chemicals alone. But the natural flavors of strawberries, grilled hamburger, or roast beef are not stable. Research involving development of synthetic shelf-stable flavors continues. Many of the types of flavors available are listed in **Table 3**.

Classification. There are several ways of dividing flavorings into subgroups other than source. They can be divided into two groups; one group affects primarily the sense of taste, and the other affects primarily the sense of olfaction. The members of the first group are called seasonings, the members of the second group are called flavors.

The same terms can be used to divide flavorings in another way. The term flavors can be applied to those products which provide a characterizing flavor to a food or beverage. The term seasoning can then be applied to those products which modify or enhance the flavor of a food or beverage—spices added to meats, blends added to potato chips, lemon added to apple pie.

Flavorings can also be classified by their physical form, as solids, liquids, or pastes. The form will be determined by the nature of the flavoring, and the form of food to be flavored.

TABLE 2. Flavor com	TABLE 2. Flavor components of various types of flavorful foods			
Туре	Component	Species	Total flavor components, ppm	
Cooked and				
processed meats	Amino acids, sugars	Sulfides, pyrazines, thiazoles, furanones	50	
Fish	Amino acids	Amines, substituted phenols	50	
Nuts	Texture	Pyrazines, lactones, substituted phenols	50	
Vegetables	Starch, sugars	Unsaturated aldehydes, sulfur compounds	10–100	
Dairy	Fats, acids	Acids, ketones, esters	100-1000	
Fermented	Alcohols, acids	Alcohols, aliphatic esters	100-1000	
Fruits	Sucrose, citric acid	Esters of aliphatic acids, aliphatic aldehydes, aromatic aldehydes, terpenes	100–1000	
Spices	Amides	Terpenes, phenols, aldehydes, ketones, ethers	2000–200,000	

Туре	Example	Use
Cooked and	Beef	Gravy mix
processed meat	Bacon	Bacon bit or analog, snack
Seafood	Shrimp	Snack cracker
Nut	Walnut	Toppings, cake mixes
	Peanut	Cookies, candies
Vegetable	Tomato	Sauces
	Cucumber	Dressings
Dairy	Butter	Margarine, baked goods
	Cheese	Snacks, analogs
Fermented	Vinegar	Salad dressings
	Bread	Chemically raised bread
	Beer and wine	Sauces
	Rum	Baked goods
Fruits	Lemon	Candy, baked goods
	Apple	Baked goods, chewing gum
	Strawberry	Candy, brandy
	Papaya	Noncarbonated beverages
Spices	Anise	Baked goods, candy
	Dill	Pickles, sauces
	Nutmeg	Sauces, condiments, processed meats

In the United States most flavorings are processed, since undried herbs and fresh fruits for desired flavor effects are available only on a small scale. Therefore another classification is by method of manufacture. There is general agreement on the several commercial designations of flavorings, indicating the method of manufacture of each and the type of flavors which are available by using ingredients from each category.

Chemicals. These are produced by chemical synthesis or physical isolation and include such substances as vanillin, salt, monosodium glutamate, citric acid, and menthol. Artificial or natural, they are only occasionally used alone, and usually are compounded or blended with other ingredients or diluted to achieve a specific flavor character at a workable concentration. *See* SALT (FOOD).

Concentrates. Powders are manufactured by dehydrating vegetables, such as onions and garlic, or concentrated fruit juices. These expensive natural products are available only in limited quantities.

Condiments. These are single ingredients or blends of flavorful (sometimes exotic) foods, spices, and seasonings, some of which may have been derived by fermentation, enzyme action, roasting, or heating. They are usually designed to be added to prepared food at the table (for example, chutney, vinegar, soy sauce, prepared mustard). Some are completely natural, while others are a blend of natural and synthetic ingredients.

Spices. These are natural substances that have been dehydrated and consist of whole or ground aromatic and pungent parts of plants, for example, anise, cinnamon, dill, nutmeg, and pepper. Such spices are

also classified as herbs if grown in temperate climates (Table 4).

Extracts. These are natural substances produced by extraction from solutions of the sapid constituents of spices and other botanicals in food-grade solvents. They are also available as synthetics. Extracts are used in home food preparation, and industrially by beverage manufacturers. Examples include vanilla, lemon, kola nut, and coffee.

Oleoresins. These are the natural volatile and nonvolatile flavor constituents of spices and other botanicals produced by volatile solvent extraction and subsequent removal of the solvent by distillation. They are used industrially in processed meat, canning, and baking industries.

Essential oils. These are the volatile oils obtained from spices and other plant materials which have the characteristic odors of their sources. They are usually obtained by distillation, although the citrus peel oils are obtained by expression. Essential oils are used in confectionery and beverage industries, and in compounded flavors for many applications. *See* ESSENTIAL OILS.

Hydrolyzed vegetable proteins. These flavors are produced by optimizing the development of a basic meaty flavor through hydrolysis of various vegetable proteins. They are used in soup formulations and in compounding flavors and seasonings.

Process flavors. These flavors are natural or artificial products of roasting, heating, fermentation, or enzymolysis of foods or normal food ingredients.

Compounded or blended flavorings. These are a group of substances produced by mixing.

Seasonings. Available in dry or liquid form, these are natural or artificial flavorings produced by the blending of various categories of flavoring agents with flavorful foods and flavoring adjuncts, for example, freeflow agents and antioxidants. They are suitable for altering the flavor of foods to which they are added, with the emphasis on the taste of the food; for example, barbecue seasoning for potato chips or chili seasoning for beans.

Flavors. These are natural or artificial flavorings produced by the compounding of various categories of flavoring ingredients and flavoring adjuncts (for example, solvents, antioxidants, and preservatives) suitable for imparting, modifying, or augmenting the flavor of foods to which they are added, with the emphasis on smell (odor, aroma), for example, artificial strawberry flavor for a gelatin dessert, natural orange flavor for a breakfast drink mix, artificial bacon flavor for a flavored dairy dip, or artificial meat flavor for a hamburger extender. Further processing before or after compounding can provide these flavors in dry or liquid form. They may be water- or oil-soluble, and are usually designed to be suitable for specific food products and processes. See FOOD ENGINEER-ING; FOOD MANUFACTURING.

Regulations. Perhaps because spices were at one time used to mask deterioration of foods, and perhaps because other flavorings were at one time used to hide economic adulterations, a complex, not

Spice	Plant and source	Plant part used	Principal use
Allspice* (Pimenta)	Evergreen of myrtle family from Jamaica and Guatemala	Dried fruit	Pickles, roast meat, ketchup
Anise* Basil	Parsley family from Mediterranean Mint family grown in U.S. and other parts of	Seed Leaves and tender	Baked goods, anisette (a liqueur) Sauces and soups, especially
Bay	Evergreen of laurel family from Mediterranean, Turkey, Greece, and Portugal	Leaf	Pickles, stews, soups, sauces
Caraway*	Biennial of parsley family from north-central Europe and southern England	Seed	Bread and baked goods, cheese, sauces kummel (a liqueur)
Cardamon*	Ginger family from Guatemala and India	Seed	Baked goods, coffee blends, curry powders
Cayenne	Capsicum family grown over most of the world (name used for those high in pungency)	Pod or fruit	Many foods, pickles, sauces, meats, curry powder
Celery*	Parsley family principally from France and India	Seed	Sauces, salads, pickles, soups
Cinnamon*	Evergreen member of laurel family from Ceylon	Bark	Baked goods, pickles, candy
Clove*	Evergreen tree of the myrtle family from Madagascar and Zanzibar	Unopened bud flower	Pork products, pickles, stews, meats, and gravies
Coriander*	Plant of parsley family from former Yugoslavia and Morocco	Dried fruit	Frankfurters, bologna, baked goods
Cumin*	Annual plant of parsley family from Iran and Morocco	Dried fruit	Chili powder
Dill*	Member of parsley family from India and domestic sources	Seed and leaves	Pickles, soups, sauces
Fennel*	Perennial of parsley family from India and Rumania	Seed	Baked goods, salad dressings, meat products
Garlic*	Member of lily family from U.S. and Mediterranean	Bulb	Baked meats, sauces, dressings, soups
Ginger*	Tuberous perennial from Jamaica, India, and Africa	Rhizome	Soft drinks, baked goods, pickles,
Mace	Tropical tree similar to rhododendron from Indonesia and West Indies	Aril covering nutmeg	Baked goods, processed meats
Marjoram	Perennial of mint family from France, Peru, and Chile	Leaves	Soups, stews, sauces
Mint	Perennial herb of many varieties grown in U.S.	Leaves	Confections, sauces, jellies, gums
Mustard*	Annual of mustard family from U.S. in Montana and California	Seeds	Sauces, baked meats, processed meats
Nutmeg*	Tropical tree similar to rhododendron from Indonesia or West Indies	Seed	Baked goods, processed meats, meat, fruit sauces
Oregano	Perennial of mint family from U.S., Mexico, Italy, and France	Leaves	Sauces and stews, especially tomato-based
Paprika*	Member of capsicum family from Spain, U.S., and Hungary	Pod or fruit	Red-colored dressings, sauces, meats, condiments as ketchup
Pepper* black	Perennial climbing vine from India and Indonesia	Berry (black); berry with cortex	Virtually all food products, table use
Poppy seed	Annual of poppy family from Poland,	Seed	Toppings for baked goods and
Rosemary	Evergreen of mint family from France, Spain, and Portugal	Leaf	Meats, sauces, gravies
Saffron*	Plant of crocus family from Spain	Stigma of flower	Coloring rice and other specialties
Savory	former Yugoslavia	Leaf	Meats stuffings salade sauces
Sesame	From sesame plant in LLS Nicaragua	Seed	Raked goods and confections
Thuman	Salvador, Egypt, and Brazil	Seeu	Baked goods and connections
Turmeric*	Member of ginger family from India, Haiti,	Rhizome	Yellow color in pickles, sauces, and fish

always logical, series of laws and regulations has been enacted in the United States and most other developed countries to prevent the use of flavorings in a deceptive manner. The basis of the laws and regulations varies from country to country, causing problems in international trade. The *Codex Alimentarius* represents an effort to standardize the composition and identification of many major foods and food ingredients. It is jointly sponsored by the United Nations Food and Agriculture Organization (FAO) and the World Health Organization (WHO) with support from other groups such as the International Organization of Flavor Industries (IOFI).

Laws and definitions differ widely from country to country. In the United States the addition of spices and natural and artificial flavors to food (including beverages and chewing gum) is permitted by law and governed by regulations. These regulations are included in the Code of Federal Regulations. Natural flavors are those which contain only natural flavoring ingredients; artificial flavors are those which contain at least some artificial flavoring ingredients. Natural flavoring ingredients include essential oils, oleoresins, essences or extractives, protein hydrolysates, distillates or any products of roasting, heating or enzymolysis which contain the flavoring constituents derived from spices, fruits, or fruit juices, vegetables or vegetable juices, edible yeast, herbs, barks, buds, roots, leaves or similar plant materials, meats, seafood, poultry, eggs, dairy products, or fermentation products whose significant function is flavoring rather than nutritive. An artificial flavor is any substance whose function is to impart flavor which is not derived as a defined natural flavor ingredient. In addition, any flavor ingredient-natural or artificial-must either be generally recognized as safe (GRAS) or covered by a regulation of the United States Food and Drug Administration. Earl J. Merwin

Bibliography. K. Bauer and D. Garbo (eds.), Common Fragrance and Flavor Materials, 3d ed., 1997; D. B. Birch and M. G. Lindley (eds.), Developments in Food Flavours, 1987; H. B. Heath and G. A. Reineccius, Flavor Chemistry and Technology, 1986; J. K. Jank, Spices: Their Botanical Origin, Their Chemical Composition, Their Commercial Use, Including Seeds, Herbs, and Leaves, 1980; D. B. Min (ed.), Flavor Chemistry of Lipid Foods, 1989; I. D. Morton and A. J. MacLeod (eds.), Food Flavours, pt. A: Introduction, 1982, pt. B: The Flavour of Beverages, 1986; R. Taranishi, R. G. Buttery, and F. Shadidi, Flavor Chemistry: Trends and Developments, 1989.

Spider silk

Spider silks are fine fibers produced from fibrousprotein solutions for use by spiders in the various functions required for their survival. Spiders are the only animals that use silk for a variety of functions in their daily lives, including reproduction, food capture, and construction. Silk fibers produced by silkworms to build their cocoons have been used in the textile industry for more than 5000 years. But in recent years, spider silk, particularly silk from the major ampullate glands used to make elaborate "orb" webs, has attracted the attention of scientists and engineers in many fields from evolutionary biology to mechanical engineering and material science. The exceptional toughness, stiffness, and extensibility of spider-silk fibers outperform even the best synthetic materials. With the recent and expected advances in molecular genetics and proteomics, as well as in the understanding of material properties in the nanoscale size range, orb-web-weaving spiders are ideal organisms for studying the relationship among the physical properties of protein-based materials in terms of their function and evolution. See ARANEAE; FIBROUS PROTEIN; NATURAL FIBER; SILK.

Mechanical properties. Over the past 400 million years, many different kinds of silk have evolved in the nearly 37,000 known spider species. For every

kind of spider silk that has been characterized in any detail, there are more than 1000 that have not been characterized. Silk from only a few species, such as Nephila clavipes (the golden orb-weaving spider) and Araneus diadematus (common garden spider), have been examined in detail. Orb-weaving spiders produce silk for different functions from as many as seven different glands (see illustration and Table 1). Dragline and viscid silk fibers have been the focus of recent research because they are among the toughest materials known. Toughness is a measure of a material's energy absorption capacity. A single silk fiber is only 1-10 micrometers in diameter, compared to a human hair which typically is close to 50 μ m. Dragline silk will support the spider's weight and is used to make the radial threads in webs. Viscid silk is the sticky, spiral threads in webs used for capturing prey. Table 2 lists the characteristic mechanical properties of dragline and viscid silk, compared to other materials. Most of the other materials are either very strong but not tough and extensible (such as steel) or are highly extensible but not strong (such as rubber). The extraordinary characteristic of dragline silk is that it has an exceptional combination of high stiffness, strength, extensibility, and toughness. The breaking stress or strength of dragline silk is comparable to steel, Kevlar® (used in bullet-proof clothing), and carbon fibers, but it is more extensible and tougher. Viscid silk is even more extensible, with stiffness comparable to rubber, but it is much stronger than rubber. The high stiffness of dragline and the rubberlike extensibility of viscid silk with the exceptional ability to absorb energy appear to be optimized for their required functions. Scientists seek to understand how spiders are able to spin silk fibers from an aqueous protein solution with the desired properties. Using spider silk as a guide, it should be possible to make the next-generation high-performance (very strong and yet very extensible) materials. See STRENGTH OF MATERIALS; STRESS AND STRAIN; YOUNG'S MODULUS.

Structure. All types of silk have the same basic building blocks-protein polymers made of amino acids. Although the different silk glands in an evolutionarily advanced orb-weaving spider, such as N. clavipes, possibly evolved from a single type of gland, silk glands now differ significantly from one another in morphology and composition, and produce different kinds of silk. The best-characterized silk protein and associated fiber is the dragline silk of N. clavipes. Dragline-silk fibers are composed primarily of two protein polymers: spidroin I and spidroin II. Both spidroin I and II consist of repeated alternating protein sequences rich in alanine (Ala) and glycine (Gly) amino acids. These proteins are stored in the glands as concentrated solutions (30-50 wt %). A combination of physicochemical and rheological processes (including ion exchange and elongational flow) cause the protein molecules to self-assemble and transform from an unaggregated and soluble state in the gland, through an intermediate liquidcrystalline state, into an insoluble fiber, while passing through the spinning canal.



Silk fibers and silk glands in Araneus diadematus. (Reprinted from F. Vollrath, Strength and structure, Rev. Mol. Biotechnol., 74(2):67–83, 2000)

After spinning, the dry silk fibers consist of stiff nanometer-scale crystals embedded in a softer protein matrix. The nanocrystals are formed by the aggregation of hydrophobic polyalanine regions into a highly ordered crystalline β -sheet via self-assembly, while the glycine-rich protein regions form the softer rubbery matrix. This combination of hard crystallites embedded in a soft matrix is considered the primary reason for the exceptional mechanical strength and extensibility of dragline fibers. It is still not clear how spiders and silkworms prevent and regulate the formation of β -sheet aggregates during the spinning process, but understanding of the mechanism is of great relevance for protein folding. An understanding of this mechanism could help in developing therapies for protein-folding diseases, such as Alzheimer's, which result from the uncontrolled aggregation of proteins into β -sheet aggregates known as amyloid plaques. See AMINO ACIDS; POLYMER; PRO-TEIN.

Mimicking spider silk. Although spidroin I and spidroin II have been sequenced, a complete under-

standing of the various physicochemical processes during the spinning of silk fibers is still lacking. Before we can completely mimic spider silk, we must fully understand the biochemistry, processing strategy, and molecular basis of the mechanical properties of silk. Advances are being made on different fronts to emulate different aspects of spider silk. Three main approaches are being pursued to make materials with mechanical properties comparable to dragline silk. The first approach is to use recombinant DNA methods to produce proteins with structure similar to spidroin I and II in other organisms, including potato, tobacco, Escherichia coli, and goat's milk. The second approach is to synthesize high-molecular-weight block copolymers that mimic the molecular structure of silk proteins. The third approach is to develop strong, yet extensible synthetic composites by reinforcing elastomeric polymers with inorganic nanoparticles. To date, these approaches have had limited success. One resulting application is the spinning of fibers from recombinant silk proteins.

TABLE 1. Types of common spider silk and their function			
Silk	Gland	Function	
Dragline Viscid Gluelike Minor ampullate Cocoon Wrapping Attachment	Major ampullate Flagelliform Aggregate Minor ampullate Cylindriform Aciniform Pyriform	Radial threads of web frame safety line Prey capture Prey capture, attachment Orb-web frame construction Reproduction Wrapping captured prey Attachment to environmental substrates	

Spi	lite	271

Material*	Stiffness or Young's modulus, GPa (kilo psi)	Strength or breaking stress, GPa (kilo psi)	Extensibility or maximum strain, %	Toughness or maximum energy absorption, MJ/m³ (Btu/in.³)
AD dragline silk	10 (1450)	1.1 (159)	27	160 (2.48)
AD viscid silk	0.003 (0.43)	0.5 (72)	270	150 (2.32)
NC dragline silk	22 (3190)	1.1 (160)	9	150 (2.32)
NE dragline	8 (1160)	1.1 (160)	28	165 (2.56)
Silkworm silk	7 (1015)	0.6 (87)	18	70 (1.08)
Tendon, collagen	1.5 (217)	0.15 (22)	12	7.5 (0.12)
Bone	20 (2900)	0.16 (23)	3	4 (0.06)
Elastin	0.001 (0.145)	0.002 (0.29)	150	2 (0.03)
Synthetic rubber	0.001 (0.145)	0.05 (7.25)	850	100 (1.55)
Nylon fiber	5 (725)	0.95 (137)	18	80 (1.24)
Kevlar 49	130 (18850)	3.6 (522)	2.7	50 (0.77)
Carbon fiber	300 (43500)	4 (580)	1.3	25 (0.39)
High tensile steel	200 (29000)	1.5 (217)	0.8	6 (0.09)

In addition to the material aspects, the silkspinning process is far superior and more sophisticated than industrial spinning processes for making polymeric fibers. Conventional industrial processes, such as wet spinning or dry spinning, operate only under very controlled conditions during which the polymer filaments are rapidly solidified or "quenched" to provide sufficient mechanical strength to be drawn into fibers. Spiders are able to spin silk fibers under widely different ambient conditions, and the silk filaments rarely break during spinning. Even in the wet state during spinning, the silk filaments are strong enough to support the weight of the spider. Microscopic analysis shows that the fibers have radially graded mechanical properties, with aligned, stiffer molecules surrounding a softer elastic core. Spiders also produce different silk fibers using the same apparatus through highly regulated protein self-assembly. A complete understanding of spiders' spinning processes would help in developing new technologies that produce nanometer- and micrometer-scale fibers from selfassembling polymers and nanocomposite materials. See ELASTICITY; MANUFACTURED FIBER; NANOSTRUC-TURE. Nitin Kumar; Gareth H. McKinley

Bibliography. C. L. Craig, *Spiderwebs and Silk: Tracing Evolution from Molecules to Gene to Phenotypes*, Oxford University Press, New York, 2003; J. Gosline et al., The mechanical design of spider silks: From fibroin sequence to mechanical function, *J. Exp. Biol.*, 202:3295–3303, 1999; D. Kaplan et al., *Silk Polymers: Material Sciences and Biotechnology*, American Chemical Society, Washington DC, 1994; F. Vollrath, Strength and structure of spiders' silks, *Rev. Mol. Biotechnol.*, 74:67–83, 2000; F. Vollrath and D. P. Knight, Liquid crystalline spinning of spider silk, *Nature*, 410:541–548, 2001.

Spilite

An aphanitic (microscopically crystalline) to very fine-grained igneous rock, with more or less altered appearance, resembling basalt but composed of albite or oligoclase, chlorite, epidote, calcite, and actinolite.

In spite of the highly sodic plagioclase, spilites are generally classed with basalts because of the low silica content (about 50%). Spilites also retain many textural and structural features that are characteristic of basalt.

Spilites are found most frequently as lava flows and more rarely as small dikes and sills. Spilitic lavas typically show pillow structure, in which the rock appears composed of closely packed, elongated, pillow-shaped masses up to a few feet across. Pillows are typical of subaqueous lava flows. Vesicles, commonly filled with various minerals, may give the rock an amygdaloidal structure. *See* SYN-CLINE.

Once thought by some petrologists to be of uniquely igneous origin, spilites are now recognized as metamorphosed basalts. The metamorphism which takes place consists of the reaction of primary Ca-rich plagioclase to produce the assemblage of albite and one or more calcic minerals such as epidote, calcite, prehnite, and pumpellyite. Primary pyroxene and olivine in the basalt convert typically to chlorite, actinolite, and serpentine. In many cases, except for the filling of vesicles, the texture of the original basalt is essentially preserved. Some basalt flows have endured spilitization of their permeable, vesicular tops, yet remain unaltered in their denser interior. The conversion of basalt to spilite is a hydrothermal metamorphism that results from the action of hot briny water which permeates the rock, taking advantage of the pillow or vesicular structures to gain access to the primary minerals. In places where former oceanic crust is now exposed on land, spilite and keratophyre (spilitized andesites) pillow lavas are seen repeatedly. This indicates that low-grade hydrothermal metamorphism is common in the oceanic crust. See BASALT; IGNEOUS ROCKS; METAMORPHISM; METASO-MATISM: PETROLOGY. William Indersoll Rose, Jr.

Bibliography. G. C. Amstutz et al. (eds.), *Spilites and Spilitic Rocks*, 1974; M. G. Best, *Igneous and Meta-morphic Petrology*, 2d ed., 2002.

Spin (quantum mechanics)

The intrinsic angular momentum of a particle. It is that part of the angular momentum of a particle which exists even when the particle is at rest, as distinguished from the orbital angular momentum. The total angular momentum of a particle is the sum of its spin and its orbital angular momentum resulting from its translational motion. The general properties of angular momentum in quantum mechanics imply that spin is quantized in half integral multiples of \hbar ($b/2\pi$, where *b* is Planck's constant); orbital angular momentum is restricted to half *even* integral multiples of \hbar . A particle is said to have spin $\frac{3}{2}b$. See ANGULAR MOMENTUM.

A nucleus, atom, or molecule in a particular energy level, or a particular elementary particle, has a definite spin; for instance, a deuteron has spin 1, a $_{3}\text{Li}^{+5}$ nucleus in its ground state has spin $^{3}/_{2}$, and an electron has spin $^{1}/_{2}$. The spin is an intrinsic or internal characteristic of a particle, along with its mass, charge, and isotopic spin. *See* I-SPIN; SYMMETRY LAWS (PHYSICS).

A particle of spin *s* has 2s + 1 spin states, since according to quantum mechanics the projection of an angular momentum of magnitude *j* along an axis can have the 2j + 1 integrally spaced values $j, j - 1 \cdots -j + 1, -j$. These spin states represent an internal degree of freedom of a particle, in addition to its external freedom of motion in three-dimensional space.

In field theory, in which particles are regarded as quanta of a field, the spin of the particle is determined by the tensor character of the field. For instance, the quanta of a scalar field have spin 0 and the quanta of a vector field have spin 1. A celebrated theorem of quantum field theory, proved first by W. Pauli, states a connection between spin and statistics: A particle with half even integral spin obeys Bose-Einstein statistics and is called a boson; a particle with half odd integral spin obeys Fermi-Dirac statistics and is called a fermion. *See* ELECTRON SPIN; QUANTUM FIELD THEORY; QUANTUM MECHAN-ICS; QUANTUM STATISTICS. Charles J. Goebel

Spin-density wave

The ground state of a metal in which the conductionelectron-spin density has a sinusoidal variation in space, with a wavelength usually incommensurate with the crystal structure. This antiferromagnetic state normally occurs in metals, alloys, and compounds with a transition-metal component. It occurs also, however, in quasi-one-dimensional organic conductors. *See* ANTIFERROMAGNETISM; CRYSTAL STRUC-TURE; ELECTRON SPIN.

In a metal, the conduction-electron charge densities for up and down spin are ordinarily equal. Their sum, $\rho_0(\mathbf{r})$, exhibits a dependence on \mathbf{r} having the same spatial periodicity as that of the lattice. A metal with a spin-density wave has, instead, charge densities given by Eq. (1). The spin density $\sigma(\mathbf{r})$ is the

$$\rho_{\pm}(\mathbf{r}) = \frac{1}{2}\rho_0(\mathbf{r})[1 \pm p\cos\mathbf{Q} \cdot \mathbf{r}]$$
(1)

difference between $\rho_+(\mathbf{r})$ and $\rho_-(\mathbf{r})$, and is therefore given by Eq. (2),

$$\sigma(\mathbf{r}) = \hat{\varepsilon} p \rho_0(\mathbf{r}) \cos \mathbf{Q} \cdot \mathbf{r}$$
(2)

where $\hat{\varepsilon}$ is a unit vector defining the axis of spin quantization, and *p* is the amplitude of the spindensity wave. The wave vector **Q** is determined by the conduction-electron Fermi surface. The wavelength, $\lambda = 2\pi/Q$, of the spin-density wave is thus in general not commensurate with the lattice period *a*. The ratio λ/a varies with temperature, pressure, alloy composition, and other parameters, because of changes in the electronic structure or in the population of electron states as a function of these parameters. *See* FERMI SURFACE.

Origin. The Pauli exclusion principle keeps electrons with parallel spin apart. The resulting reduction in Coulomb interaction is called exchange energy. This reduction is increased when the conduction-electron charge densities are modulated according to Eq. (1). However, modulation requires an increase in conduction-electron kinetic energy. According to the spin-density-wave instability theorem, energy can always be further reduced if the quantum theory is treated in an approximation that neglects electron-electron scattering. Were it not for these latter correlations, every metal would have one or more spin-density waves. *See* EXCHANGE INTERACTION; EXCLUSION PRINCIPLE.

The correlation energy, which contributes to a metal's stability, arises primarily from scattering of up-spin electrons by down-spin electrons. Since a spin-density wave tends to stratify opposite spin densities in alternating layers, as indicated in Eq. (1), the scattering is reduced, thereby tending to suppress the spin-density-wave instability. This contrasts with a charge-density wave, where shifts of exchange and correlation energy act in unison. Consequently, spin-density-wave states are less likely to occur than charge-density-wave states. *See* CHARGE-DENSITY WAVE.

Detection. Since the total electronic charge density, the sum of ρ_+ and ρ_- in Eq. (1), is equal to $\rho_0(\mathbf{r})$, a spin-density wave cannot be detected in an x-ray diffraction experiment. It leads, however, to a sinusoidal magnetic field, from which neutrons, having a magnetic moment, can experience Bragg diffraction. At incommensurate points in wave-vector space, two magnetic satellites are observed on opposite sides of each crystallographic Bragg reflection.

The magnetic origin of spin-density-wave satellites can be demonstrated in two ways. The first is their absence in an x-ray experiment. The second is by means of neutron diffraction with a spin-polarized beam, since only a magnetic reflection can cause reversal of neutron polarization during diffraction. *See* NEUTRON DIFFRACTION; X-RAY DIFFRACTION.

Occurrence. There are well over 100 materials which, over a temperature range, support a



Types of spin-density wave observed in metals: SS = spiral (helical); FS = ferromagnetic (cone); SS = skewed; LSW = longitudinal; TSW = transverse; FAN = stable only in a magnetic field B in the direction shown.

spin-density wave. These include some of the rareearth elements of the lanthanide series and the 3*d* transition metals, manganese and chromium, the latter being the prototypical itinerant electron antiferromagnet. Many types of spin-density wave occur, since in general not only may the vector $\hat{\varepsilon}$ in Eq. (2) be parallel or perpendicular to the wave vector \mathbf{Q} , corresponding to a longitudinal or transverse spindensity wave, respectively, but in the latter case $\hat{\varepsilon}$ may rotate about \mathbf{Q} , corresponding to a helical spindensity wave. A uniform ferromagnetic magnetization may be superposed on the spin-density wave (see **illus.**).

The occurrence of inelastic neutron-scattering peaks at incommensurate points indicates the existence of spin-density-wave fluctuations in some metals thought to be nonmagnetic (for example, copper and yttrium) when doped with magnetic impurities (manganese and gadolinium, respectively). This behavior suggests that the spin-density-wave instability may be common, even in nontransition metals.

Spin-density-wave fluctuations have been proposed as an essential feature for some forms of high-temperature superconductivity. In one class of quasi-one-dimensional organic conductors, the Bechgaard's salts, the spin-density-wave states give way to a superconducting state under pressure. In another case, $(TMTSF)_2ClO_4$ (TMTSF = tetramethyltetraselenafulvalene) is superconducting at low temperatures and at ambient pressure but develops a spin-density wave in a magnetic field. *See* ORGANIC CONDUCTOR; RARE-EARTH ELEMENTS; SUPERCONDUCTIVITY; TRANSITION ELEMENTS.

Bibliography. E. Fawcett, Spin-density-wave antiferromagnetism in chromium, *Rev. Mod. Phys.*, 60:209– 283, 1988; E. Fawcett, Spin-density waves and SDW fluctuations, *Sov. J. Low Temp. Phys.*, 15:441–444, 1989; Yu. A. Izyumov, Modulated, or long-periodic, magnetic structures of crystals, *Sov. Phys. Usp.*, 27:845–867, 1984; D. Jérome and H. J. Schulz, Organic conductors and superconductors, *Adv. Phys.*, 31:299–490, 1982; N. I. Kulikov and V. V. Tugushev, Spin-density waves and itinerant antiferromagnetism in metals, *Sov. Phys. Usp.*, 27:954–976, 1984; A. W. Overhauser, Spin-density waves in an electron gas, *Phys. Rev.*, 128:1437–1452, 1962.

Spin glass

One of a wide variety of materials that contain interacting atomic magnetic moments and also possess some form of disorder, in which the temperature variation of the magnetic susceptibility undergoes an abrupt change in slope, that is, a cusp, at a temperature generally referred to as the freezing temperature, T_{f} . At temperatures below T_{f} the spins have no long-range magnetic order, but instead are found to have static or quasistatic orientations which vary randomly over macroscopic distances. The latter state is referred to as spin-glass magnetic order. Below T_f these systems also exhibit magnetic hysteresis and slow relaxation phenomena. Analogous effects are observed in measurements of torque in an applied magnetic field and in electron paramagnetic resonance.

Spin-glass ordering is usually detected by means of magnetic susceptibility measurements, although additional data are required to demonstrate the absence of long-range order. Closely related susceptibility cusps can also be observed by using neutron diffraction. It is not generally agreed whether spin glasses undergo a phase transition or not. *See* MAGNETIC SUSCEPTIBILITY; NEUTRON DIFFRACTION; PHASE TRANSITIONS.

Spin-glass transition. An example of a spin-glass ordering transition is shown in **Fig. 1**, where magnetic susceptibility χ is plotted as a function of absolute temperature *T* for the case of dilute alloys of iron in gold. The iron solute atoms carry a magnetic moment and occupy sites at random in the gold lattice. Neighboring moments interact with one another via the Ruderman-Kittel-Kasuya-Yosida (RKKY) exchange interaction *U*, given byEq. (1), which is

$$U = -\sum_{ij} J_{ij} \vec{S}_i \cdot \vec{S}_j \tag{1}$$

mediated by the conduction electrons of the gold. The exchange constant J_{ij} oscillates rapidly in sign with increasing distance between spins \vec{S}_i and \vec{S}_j . Anisotropic interactions analogous to the RKKY exchange as well as dipolar and crystal field interactions are also believed to have important consequences for spin-glass properties. The sharp susceptibility peak in Fig. 1 occurs at a temperature T_f such that $k_B T_f$ has the value of a typical exchange constant J_{ij} , where k_B is Boltzmann's constant. In this case the necessary disorder arises from the random placement of the magnetic moments. The resulting configuration of exchange couplings is such that at low temperatures the minimum energy state is one where the moment orientations are only correlated locally and are effectively random over macroscopic distances.

As seen in Fig. 1, the sharpness of the transition



Fig. 1. Magnetic susceptibility versus absolute temperature for two different concentrations of iron alloyed with gold. The data curves indicate measurements in several applied field strengths as shown. 1 Oe = 79.6 units. Magnetic susceptibility in SI units = magnetic susceptibility in emu multiplied by 4π = 12.57. (*After V. Cannella and J. A. Mydosh, Magnetic ordering in gold-iron alloys, Phys. Rev.*, *B6:4220–4237*, 1972)

is destroyed by the application of a relatively small magnetic field, $H_0 \sim 100$ oersteds (1 Oe = 79.6 ampere-turns/m). This is evidence that the transition is a cooperative effect involving a large number of moments, since the motion of an isolated moment would be largely unaffected by a field of 100 Oe at this temperature.

Some of the basic elements of the spin-glass transition are reproduced in a mean-field model calculation. In this model one calculates the thermodynamic behavior of a system of classical spins which interact via exchange couplings having the form given by Eq. (1) with, however, a distribution of J_{ij} values which is taken to be a gaussian random function. Below the temperature T_{f_i} the system is found to be ordered in the sense that each spin \vec{S}_i will retain some memory of its orientation $\vec{S}_i(t_1)$ at time t_1 at a much later time t_2 , even though $\langle \vec{S}_i(t_1) \rangle = 0$, where $\langle \rangle$ denotes an average over all spins. The ordering effect is expressed by the Edwards-Anderson order parameter $q = \langle \vec{S}_i(t_1) \cdot \vec{S}_i(t_2) \rangle$. Their theory gives Eq. (2) for the magnetic susceptibility χ . Here $\chi_c = a/T$ is the Curie

$$\chi = \chi_c (1 - q) \tag{2}$$

susceptibility, where *a* is the Curie constant. For $T > T_f$, q = 0; and for $T < T_f$, $q = \frac{1}{2} [(T_f/T)^2 - 1]$. Thus $\chi(T)$ is just the Curie susceptibility for $T > T_f$, and for $T < T_f$ is given by Eq. (3) for $T_f - T < T_f$, where *b* is a

$$\chi(T) \cong \frac{a}{T_f} - b(T_f - T)^2 \tag{3}$$

constant. Equation (2) therefore exhibits a cusp similar to that shown in Fig. 1. The Edwards-Anderson model also predicts a singularity in the specific heat at T_{f} . However, this is not observed experimentally.

Spin-glass ordering. Spin-glass ordering has been found to occur in a wide variety of disordered magnetic materials, exhibiting several distinct types of disorder. The following list is intended to be illustrative but not exhaustive. The dilute alloy Au:Fe cited above is disordered as a consequence of random placement of iron atoms in the gold lattice. For similar reasons other dilute alloys with noble-metal hosts (for example, Cu:Mn, Ag:Mn, and Au:Mn) are also spin glasses. Some of these systems show spinglass behavior over a wide range of composition, limited only by the Kondo effect and possible ferromagnetic behavior at low and high concentrations of the magnetic ion, respectively. At low concentrations the exchange interactions of these alloys are well characterized as to their strength and distribution. Another type of metallic spin glass is one which contains a relatively high concentration of magnetic ions and is structurally amorphous. Examples of this are GdAl₂ and $(Fe_xMn_{1-x})_{0.75}(P_{16}B_6Al_3)_{0.25}$. In the former case the gadolinium ions are the magnetic constituent. Special techniques are required to prepare these materials in a structurally amorphous state. The second example cited is from a family of materials known as metglass and is a reentrant spin glass, as discussed below. See AMORPHOUS SOLID; KONDO EFFECT.



Fig. 2. Magnetic susceptibility versus absolute temperature for the compound $(Pd_{0.9965}Fe_{0.0035})_{1-x}Mn_x$ for three values of the manganese atomic fraction x. The scale for the curve marked x = 0.065 has been magnified four times. (After G. J. Nieuwenhys, B. H. Verbeek, and J. A. Mydosh, Towards a uniform magnetic phase diagram for magnetic alloys with mixed types of order, J. Appl. Phys., 50:1685–1690, 1979)

There are also a number of interesting cases of insulating spin glasses. The best-characterized and most thoroughly studied insulating spin glass is the system $Eu_xSr_{1-x}S$. This compound is a ferromagnet for x = 1 as a consequence of the ferromagnetic nearest neighbor exchange interaction among the europium, while the original crystalline structure is retained. At sufficient dilution (x < 0.5) the system becomes a spin glass through the agency of antiferromagnetic second-neighbor exchange couplings. At higher europium concentrations this system also becomes reentrant.

Ferromagnetic to spin-glass transition. A number of compounds are found to form ferromagnetic order at some Curie temperature T_c , then exhibit a collapse of that order as the temperature is lowered below T_c . Such a system is termed a reentrant spin glass, because the spin-glass phase appears below a temperature where the magnetically ordered higher temperature phase collapses. An example of this behavior is shown in Fig. 2, where magnetic susceptibility versus absolute temperature is plotted for three concentrations of manganese (Mn) in the compound $(Pd_{0.9965}Fe_{0.0035})_{1-x}Mn_x$. For x = 0.01 the compound is a ferromagnet, with χ saturated for all temperatures below 13 K. For x = 0.05 the compound becomes ferromagnetic at $T \sim 9$ K, but the ferromagnetism collapses into spin-glass behavior for T < 3K. This is the reentrant-spin-glass case. For higher manganese concentrations (for example, x = 0.065) the ferromagnetism disappears, and the material is a simple spin glass as in Fig. 1. Whether the spinglass transition of the ferromagnetic to spin-glass transition is a true phase transition is a matter of some controversy. Reentrant-spin-glass behavior has been observed in a variety of other systems as well, including Eu_xSr_{1-x}S, Au_{1-x}Fe_x, Fe_xCr_{1-x}, and amorphous $(Fe_{1-x}Mn_x)_{0.75}(P_{16}B_6Al_3)_{0.25}$. The first of these is a nonconductor, and the others are magnetic al-R. E. Walstedt loys. See FERROMAGNETISM.

Spin label

A molecule which contains an unpaired electron spin which can be detected with electron spin resonance (ESR) spectroscopy. Molecules are labeled when an atom or group of atoms which exhibit some unique physical property is chemically bonded to a molecule of interest. Groups containing unpaired electrons include organic free radicals and a variety of types of transition-metal complexes (such as vanadium, copper, iron, and manganese). Molecules with unpaired electron spins are readily detected with electron spin resonance spectroscopy. Through analysis of ESR spectra, rates of molecular motion whose motion is restrained by surrounding molecules can be determined. Measurement of rates of molecular motion and molecular orientation has proved to be very important in the study of a variety of types of biologically important problems.

Electron spin resonance. This is a spectroscopic technique which detects transitions between electron spin energy levels. When a molecule possessing an unpaired electron spin is placed in an external magnetic field, the external field interacts with the magnetic moment of the unpaired electron spin, producing two distinct energy states. In one state the unpaired electron is oriented with its magnetic moment along the direction of the external field, while the other state has the moment of the unpaired electron oriented against the external field direction. The separation of these two energy states is about 3 cm⁻¹ in an external field of 3000 gauss (0.3 tesla), and an absorption of radiation can be observed when a sample is irradiated with microwave radiation of about 3 cm wavelength. This absorption is observed as a single line in an electron spin resonance spectrum. The wavelength at which absorption of energy is observed depends on the magnitude of the external field and on a molecular property called the g value.

Crystals. If a crystalline sample is used, different gvalues may be observed along the various axes of the crystal. When directional dependence is observed the g value is anisotropic, with principal values along the x, y, and z axes of the crystal denoted by g_{xx}, g_{yy} and g_{zz} . If $g_{xx} = g_{yy}$, the sample has an axial symmetric g tensor, and the g values are denoted as g_{\parallel} and g_{\perp} . The unpaired electron spin is often delocalized over two or more atoms in the spin label. If the nuclei have nuclear spin other than zero, the magnetic moment of the nuclei can interact with the magnetic moment of the electron spin and split the electron spin energy levels. When this is the case, the ESR spectrum shows a series of lines that depend on the number and type of nuclear spins in the molecule. The interaction between the electron and nuclear spins is called a hyperfine interaction, and the separation of the energy levels and ESR lines is called the hyperfine splitting. The magnitude of the interaction between nuclear and electron spins may depend on the orientation of the molecule with respect to the external field. The components of the electron-nuclei anisotropic splitting are denoted by A_{xx} , A_{yy} , and A_{zz} , or A_{\perp} and A_{\parallel} when the molecule has axial symmetry.

Solutions. When a spin-labeled molecule is free to rotate in solution, the anisotropic g value and anisotropic hyperfine interaction are averaged by rapid reorientation of the label, and one observes the isotropic (angle-independent) g value and hyperfine splitting in the ESR spectrum. As rotation of the molecule is slowed, the anisotropic interactions start to contribute, and the ESR spectrum is changed. A continuous change in the ESR spectrum is observed as the sample goes from the case of free rotation to the case of a rigid sample where motion is very slow. Analysis of ESR spectra permits determination of motional correlation times from about 10^{-12} to 10^{-3} s. In many cases, determination of the direction of motion of the labeled group (that is, rotation in all directions or preferred rotation around a single axis) is also possible.

Biological studies. Analysis of the rate and type of motion of a spin label is important for a wide variety of biological problems. The type of label used in these studies is generally a nitroxide free radical. The unpaired electron spin in a nitroxide radical is delocalized over an N-O group, and an electron-nuclei hyperfine splitting is observed from the spin = 1 nitrogen nucleus. The oxygen atom has a spin of zero. In solution an isotropic nitrogen hyperfine splitting of about 15 gauss (1.5 millitesla) is observed, while in a crystalline solid anisotropic splitting is observed with $A_{zz} = A_{\parallel} = 32$ gauss (3.2 millitesla) and $A_{xx} = A_{\perp} = 6$ gauss (0.6 mT). The *z* direction is defined by the 2*p* orbital of the nitrogen atom.

Enzymes. Nitroxide spin labels have been bonded to a number of enzyme substrates with connecting chains of varying length between the label and the substrate. The labeled substrate is allowed to form a complex with the enzyme, and the ESR spectrum is monitored. The receptor site on the enzyme is often in a type of well within the protein, within which the substrate must fit. If the nitroxide label is connected to the substrate by a short chain, it lies within the well and rotational motion is inhibited. When a sufficiently long connecting chain is used, the nitroxide lies outside the protein and rotational motion is relatively free. Changes in motion of the label can be monitored with ESR, and the transition from hindered to free rotation as a function of the length of the connecting chain can be determined. The distance from the substrate to the label can be calculated from molecular models for labeled molecules with different connecting chain lengths. If the distance at which free rotation starts is determined, the depth of the receptor site well of the enzyme can be calculated. Spin-labeling studies of this type provide a powerful technique for the study of the geometry and dimensions of receptors in enzymes. See ENZYME.

Membranes. Spin labels have been used extensively to study the structure of membranes. Biological membranes consist mainly of a phospholipid bilayer with embedded proteins. Other molecules (for example, cholesterol) may also be found in the lipid bilayer, and a variety of types of molecules may be bound to the surface. The phospholipids have polar head groups (hydrophilic) which are on the outside of the membrane, and nonpolar aliphatic carbon chains which are on the inside of the bilayer. The lipids form ordered structures which are readily oriented in an external magnetic field. Spin labels have been bonded to the carbon atoms along the aliphatic chains of the lipids to study internal motion in the membrane bilayer. The ESR spectra of spin labels bound to lipids which have been oriented in an external field show that the lipid can rotate around its long axis, but that head-to-tail rotation is very slow. The relative motional correlation times of labels bound at different positions along the chain can be determined from ESR spectra. These studies show that motion is freer on the inside of the lipid away from the polar head group, indicating greater fluidity on the inside. Membrane structures can be changed by a number of kinds of perturbations, and spin labels have been used to monitor the effect of the perturbation. The structure of the phospholipid bilayer changes with temperature, and spin-labeling studies have been conducted to determine phase transition temperatures. Molecules such as cholesterol or ethanol change the fluidity of the membrane, and spin-labeling studies have been conducted to monitor the effect of chemical perturbations on the fluidity of membranes. See CELL MEM-BRANES.

When nitroxide labels are relatively close to one another, the odd electron spins in the nitroxide groups may be interchanged between molecules. This spin exchange reaction dramatically changes the appearance of the ESR spectrum of the labels, since the energy states are averaged by this interaction. This interaction may be used to study the rate of diffusion of labeled molecules from an area of high concentration (strong exchange) to an area of low concentration (weak exchange). This type of investigation has been conducted to determine the rate of lateral diffusion of lipids in membranes. A high concentration of labeled lipid is initially introduced into a membrane, and an ESR spectrum from strongly exchange-coupled spin labels is observed. As the labeled molecules diffuse apart, the ESR spectrum changes to that characteristic of labels with no exchange coupling. By monitoring the change in the ESR spectrum of the label as a function of time, it is possible to determine how fast lipids diffuse through membranes. Spin-labeling studies of these types have provided important information about the structure, organization, and rates of motion in membranes.

Other applications. Spin labels have also been used to study the structure and organization of synthetic polymers and to study phase transitions. Investigations using deuterium substitution for hydrogen with nuclear magnetic resonance detection have been conducted to study problems similar to those investigated with nitroxide labels. Spin-labeling studies allow investigation of a variety of types of problems which cannot readily be studied with other

techniques. See ELECTRON PARAMAGNETIC RESO-NANCE (EPR) SPECTROSCOPY; FREE RADICAL; TRIPLET STATE. Robert Kreilik

Bibliography. L. J. Berliner, *Biological Magnetic Resonance*, vol. 14: *Spin Labeling*, 1998; L. J. Berliner (ed.), *Spin Labeling: Theory and Applications*, vols. 1 and 2, 1976, 1979; C. P. Poole, Jr., *Electron Spin Resonance*, 2d ed., 1997.

Spinach

A cool-season annual of Asiatic origin, *Spinacia oleracea*, belonging to the plant order Caryophyllales. It is grown for its foliage (**Fig. 1**) and served as a cooked vegetable or as a salad. New Zealand spinach (*Tetrogonia expansa*) and Mountain spinach (*Atriplex hortense*) are also called spinach but are less commonly grown. Spinach plants are usually dioecious. *See* CARYOPHYLLALES.

Cultivation and harvesting. Propagation is by seed, commonly planted in rows 10-20 in. (25-50 cm) apart. Cool weather favors maximum production. High temperatures and long daylight periods encourage seedstalk formation and reduce vegetative growth. Fall-seeded spinach is overwintered and harvested in the spring in areas where the weather is mild or the crop is protected by snow cover. *See* PHOTOPERIODISM; SEED.

Varieties (cultivars) are classified according to seed (smooth or prickly), leaves (smooth or savoyed), and seeding tendencies (early or late, or long standing). Popular varieties are Long Standing Bloomsdale, Early Hybrid 7, Hybrid 424, Savoy Hybrid 621, and Viroflay.

Harvesting by hand or machine begins when the plants have reached full size and before seedstalks form, usually 40–50 days after planting.

H. John Carew

Disease. Diseases may reduce yield and quality, and be limiting factors in the production of spinach for processing and fresh market. Some of the most important diseases of spinach are Fusarium wilt, caused by *Fusarium oxysporum* f. sp. *spinaciae*; white rust, caused by *Albugo occidentalis*; downy mildew, caused by *Peronospora spinaciae*; blights, mosaics, and stunts caused by viruses and yellows and decline (**Fig. 2**). Several factors may cause a yellows and de-



Fig. 2. Four seedlings with root and hypocotyl decay (left) and three healthy seedlings (right).

cline syndrome; these include soil-borne pathogenic fungi and nematodes, nutrient imbalances, excess soil moisture, abrupt temperature changes, and prolonged periods of cloudy, wet weather.

Crop rotations, with spinach being planted only once every 3 years, and the use of disease-free seed are the most effective control measures for white rust, downy mildew, and other leaf spot diseases that survive in seed, plant debris, and soil. The wilt pathogen will survive in the soil indefinitely, so crop rotation is not an effective method of control.

Some varieties are resistant to downy mildew and blight, but market type and quality are not always acceptable. Fungicides can be used to control downy mildew and white rust, but fungicides are not available to control other foliage diseases of less economic importance, such as Cercospora leaf spot and anthracnose. To help reduce the spread of epidemics of foliage diseases, spring spinach should not be planted in the vicinity of winter spinach.

Viral diseases may be produced by destroying volunteer spinach plants and weeds that serve as alternative hosts. Controlling insect vectors may slow the spread of viruses within a field and between fields. Early spring crops can be grown before blight becomes prevalent in the northern United States, but the practice is ineffective in southern Texas, where vectors may be blown in by northerly winds. *See* PLANT PATHOLOGY. Donald R. Sumner

Bibliography. A. F. Sherf and A. A. MacNab, Vegetable Diseases and Their Control, 2d ed., 1986.



Fig. 1. Broad, mature leaves of spinach plants in the field. (Editorial Photocolor Archives)

Spinal cord

The portion of the central nervous system within the spinal canal of the vertebral column, that is, the entire central nervous system except the brain. The spinal cord extends from the foramen magnum at the base of the skull to a variable level of the spinal canal; it terminates at the lumbar level in humans and extends well into the caudal region in fishes. Paired spinal nerves enter the spinal canal between each pair of vertebrae and connect with the spinal cord. The number of spinal nerves varies widely in vertebrates; in humans there are 31 pairs (8 cervical, 12 thoracic, 5 lumbar, 5 sacral, and 1 coccygeal). In tetrapods there are spinal cord enlargements where the large spinal nerves extend to the limbs, a cervical enlargement cranially, and a lumbosacral enlargement caudally.

The internal organization of the spinal cord can be seen in a transverse section (Fig. 1). The outer portion of the spinal cord is made up of nerve fibers, most of which are oriented longitudinally and carry information between parts of the spinal cord, between spinal cord and brain, and between brain and spinal cord. The outer white matter is divided into dorsal, lateral, and ventral columns. The interior of the spinal cord consists of gray matter and is divided into a dorsal sensory horn (or column) and a ventral motor horn (or column). In the thoracic and lumbar regions of the cord there is also a small lateral horn (or column) which contains preganglionic sympathetic neurons. In the very center of the bilaterally symmetrical spinal cord is a small central canal, lined with ependymal cells and containing cerebrospinal fluid. See SYMPATHETIC NERVOUS SYSTEM.

Each spinal nerve divides into a dorsal sensory root and a ventral motor root before entering the spinal cord. The dorsal root contains the spinal ganglion (Fig. 2), and then its axons enter the dorsolateral aspect of the cord. Upon entering the spinal cord, each dorsal root axon divides and ramifies to send terminal branches to one or all of the following neuron groups: sensory neurons of the dorsal horn; motor neurons of the ventral horn; and long terminal branches to neurons in the brain. The motor neurons of the ventral horn, in addition to receiving synapses from dorsal root axons, also receive synaptic endings from neurons in other parts of the spinal cord and from long axons coming from the brain. The axons of the ventral horn neurons leave the cord through the ventral root of the spinal nerve and run with peripheral nerves to innervate the muscles of the body. With this complex synaptic and fiber organization, the spinal cord can act as the integrating center for spinal reflexes (such as the knee jerk reflex), send sensory information from the brain, and receive information from the brain to initiate or inhibit muscular activity. See MOTOR SYSTEMS; SENSATION.





Fig. 1. Human spinal cord sectioned to show the white and gray matter. (*After G. M. Wyburn, The Nervous System, Academic, 1960*)



Fig. 2. Brain and spinal cord, illustrated in ventral aspect. Nerve roots and chief autonomic ganglia are shown. (After H. Morris et al., Comparative Anatomy, Blakiston–McGraw-Hill, 1936)

cords of vertebrates are basically similar. However, in fishes some or all of the visceral motor fibers pass through the dorsal roots rather than the ventral roots of spinal nerves. *See* AUTONOMIC NERVOUS SYSTEM; NERVOUS SYSTEM (VERTEBRATE). Douglas B. Webster

Spinal cord disorders

In addition to those disorders common to the brain, the spinal cord is subject to certain lesions because of its position or structure. A few of the more important are mentioned.

Injury. Spinal cord injury results from dislocations, fracture, or compression in many cases, but a special form, called spinal shock, may result from a severe blow without actual distortion of adjacent tissue. In this case, there is a temporary paralysis which gradually clears in a length of time related to the severity of the paralysis. In direct damage, the cord may be slightly, partially, or completely damaged at one or more levels. Typical motor and sensory losses follow, with a poor prognosis for recovery if the nerve tissue is severely injured. Occasionally, similar damage may result from a hemorrhage produced by injury.

A fairly common type of potential cord injury is seen in a number of cases of slipped disks, in which the inner, soft part of the vertebral column extrudes into the spinal canal. If this compresses the cord, functional loss of temporary or permanent degree can follow; more often, pressure is exerted on spinal roots so that pain, numbness, and some type of muscle weakness intervene.

Tumors. Spinal cord tumors are not infrequent and most of these are of two types, the metastatic, from a primary source elsewhere in the body, and the tumors of the meninges or connective tissue related to the cord. The latter include neurofibromas, meningiomas, and gliomas, which occur most often. The signs and symptoms and the extent of damage relate largely to the physical compression of the cord at a particular level. *See* TUMOR.

Congenital defects. A few of the more common congenital defects involving the cord include an unclosed neural canal, or spina bifida, and reduplicated or otherwise malformed cords, such as those caught in an external sac of other tissues, the meningomyelocoele.

Inflammation. Inflammations may result from known or unknown agents and in meningitis may involve primarily the coverings; in myelitis, the cord itself. The meningococcus, pneumococcus, streptococcus, tubercle bacillus, and other microorganisms frequently cause meningitis. The most widely known cause of myelitis, of course, is the poliomyelitis virus group, although other agents will also produce inflammations, such as typhus, spotted fever, and *Treponema pallidum*, the cause of syphilis. Other diseases such as malaria, amebiasis, trichinosis, elephantiasis, toxoplasmosis, and blastomycosis may display their effects on the spinal cord.

Vascular diseases. Not uncommonly motor or sensory deficits in the spinal cord may result from specific vascular diseases and from such nonspecific conditions as injury, aneurysms, or congenital defects, and rarely, from blood vessel tumors.

Nerve tract degeneration. An ill-defined group of disorders characterized by degeneration of nerve tracts or myelin sheaths of the cord is found more often than one would suspect. In these, some unknown or poorly understood mechanism causes the deterioration of cells and fibers so that function is altered, then lost, and the nervous tissue is either replaced by a scar or a softening of cystlike area remains. Multiple sclerosis, combined degeneration associated with pernicious anemia, Parkinson's disease, postinfectious encephalomyelitis, and syringomyelia are examples. *See* MULTIPLE SCLEROSIS; PARKINSON'S DISEASE.

Drugs. Finally, the effect of drugs or chemicals and metabolic disorders is often seen as cord depression or stimulation with variable prognoses and permanence of effects. These include alcoholism, carbon monoxide poisoning, lead intoxication, vitamin deficiency, and certain familial diseases. *See* ALCOHOLISM; POISON; SPINAL CORD; VITAMIN.

Edward G. Stuart; N. Karle Mottet

Spinicaudata

An order of fresh-water branchiopod crustaceans formerly included in the order Conchostraca. The entire body and its appendages are covered by a bivalve carapace up to about 17 mm (0.7 in.) long, hence the name clam shrimp. The carapace is often bilaterally compressed and usually displays several lines of growth. The head cannot be protruded. It usually bears paired sessile eyes, but these are fused in *Cyclestheria*. Locomotion is by means of large antennae. The mandibles are of the rolling, grinding type.

There are more than 16 pairs of similar foliaceous trunk limbs (at least 32 in some species); these draw in particles of food by beating in metachronal rhythm. In the male, the first two pairs (the first pair only in *Cyclestheria*) are modified as claspers. The trunk terminates in a robust telson that bears a pair of stout claws.

Reproduction is usually bisexual. Usually, the female carries a clump of numerous, very small, drought-resistant eggs, sometimes of complex shape, that are shed at the next molt. In *Cyclestheria* a few larger parthenogenetic eggs are usually brooded in a dorsal brood pouch beneath the carapace. Eggs, except the parthenogenetic eggs of *Cyclestheria*, hatch as nauplii. All species frequent temporary pools in many parts of the world, but *Cyclestheria* also occurs in permanent waters. *See* BRANCHIOPODA.

Geoffrey Fryer

Spinning (metals)

A production technique for shaping and finishing metal. In the spinning of metal, a sheet is rotated and worked by a round-ended tool, controlled manually or mechanically. The sheet is formed over a mandrel.

Spinning operations are usually carried out on a special rigid lathe fitted only with a driving headstock, tail spindle, and tool rest. Surface speeds of 500-5000 ft/min (2.5-25 m/s) are used, depending on material and diameter. The work is rubbed with soap, lard, or a similar lubricant during working. The operation can be set up quickly, and thus is desirable for short runs or for experimental units subject to change. Spinning may serve to smooth wrinkles in drawn parts, provide a fine finish, or complete a forming operation as in curling an edge of a deepdrawn part. Other operations include smoothing, necking, bulging, burnishing, beading, and trimming. Spun products range from precision reflectors and nose cones to kitchen utensils. Such materials as steel, aluminum, copper, and their softer alloys are spun in thicknesses up to 18 in. (3 mm). It may be necessary to anneal the metal during the spinning. See SHEET-METAL FORMING. Ralph L. Freeman

Spinning (textiles)

The fabrication of yarn (thread) from either discontinuous natural fibers or bulk synthetic polymeric material. In a textile context the term spinning is applied to two different processes leading to the yarns used to make threads, cords, ropes, or woven or knitted textile products. Natural fibers, such as wool, cotton, or linen, are generally found as short, entangled filaments. Their conversion into yarn is referred to as spinning. After a carding operation on the raw material to disentangle the short filaments, the filaments are drawn (drafted) to promote alignment in an overlapping pattern and then twisted to form, by mechanical interlocking of the discontinuous filaments, a resistant continuous yarn. *See* COTTON; LINEN; WOOL.

The term spinning is also used for the production of monofilaments from synthetic polymers-for example, polyamides or nylons, polyesters, and acrylics-or modified natural polymers, such as cellulose-rayon. The process involves forcing a pressurized fluid-either a melt, solution, or chemical intermediate-through many small orifices of a spinnerette and subsequently solidifying the emerging monofilaments by cooling, drying, or coagulation. Generally the monofilaments are then stretched (drawn) to increase their strength by promoting molecular orientation and are wound as yarn which can be used directly for threads, cords, or ropes. Such yarn, however, is often cut into relatively short lengths (staple) and reformed by a process similar to that used for natural fibers into a yarn more suitable, in terms of appearance and feel, for making certain textile products. See MANUFACTURED FIBER; NATURAL FIBER; TEXTILE. J. M. Charrier

Bibliography. K. L. Hatch, *Textile Science*, 1993; M. L. Joseph et al., *Joseph's Introductory Textile Science*, 6th ed., 1992; M. Raheel (ed.), *Modern Textile Characterization Methods*, 1996; B. Wulfhorst, T. Gries, and D. Veit, *Textile Technology*, 2006.

Spiral

A term used generically to describe any geometrical entity that winds about a central point or axis while also receding from it. Spiral staircases, helices, nonplanar loxodromes (curves that intersect those of a given class at a constant angle, for example, rhumb lines, in case the curves are on a sphere whose meridians form the given class) are examples of spirals whose windings do not lie in a plane. The planar spiral curve $\rho = a\theta$, in polar coordinates ρ and θ (see **illus.**), was introduced by Archimedes in his book *On Spirals*. Other planar spiral curves are logarithmic, $\rho = e^{\theta}$, the tangents to which make a constant angle with the radii vectors drawn to the points of



Planar spiral curve, or Archimedes' spiral.

contact; hyperbolic, $\rho\theta = a$; and the lituus, $\rho^2\theta = a^2$. See HELIX. Leonard M. Blumenthal

Spiriferida

An extinct order of brachiopods, in the subphylum Rhynchonelliformea, that inhabited shallow seas of the Paleozoic and early Mesozoic. It was the most diverse group of spire-bearing brachiopods (those that contain a spirally coiled calcareous structure called the brachidium used to support the lophophore).





Spiriferida. (a-b) Dorsal and lateral views of Neospirifer, Permian, and (c) with dorsal valve removed showing spiral lophophore supports. (d) Diversity history of spiriferid families. (a-c after R. C. Moore, C. G. Lalicker, and A. G. Fischer, Invertebrate Fossils, McGraw-Hill, New York, 1952; d after J. J. Sepkoski, Jr., A Compendium of Fossil Marine Animal Families, 2d ed., Milwaukee Public Mus. Contrib. Biol. Geol., no. 83, 1992) Spiriferids also possess unequally biconvex valves that are externally smooth or radially ribbed. The valves are generally strophic (straight hinge) with a well-developed interarea commonly limited to the ventral valve (**illus.** *a-c*). Shells may be punctate (possessing small perforations filled by outer epithelial tissue) or impunctate. Punctate shells arose in stratigraphically younger spiriferids; the presence or absence of punctae serves to distinguish two major taxonomic groups.

The evolutionary origin of spiriferids is unclear. One hypothesis has the spiriferids derived from the rhynchonellids via another spire-bearing group called the athyrids. A competing hypothesis has the spiriferids derived directly from the orthids and the other spire-bearing groups derived from the rhynchonellids. Clarification of this problem will require additional analyses of new fossil material.

Spiriferids were sessile, attached, epifaunal suspension feeders. Most had a functional pedicle, used for attachment to the substrate. The earliest spiriferids appeared in deep-water environments in the Late Ordovician and thus did not take part in the Ordovician radiation of marine life. Spiriferids first diversified in the Silurian following the Late Ordovician mass extinction (illus. d). They, along with strophomenids and pentamerids, became dominant members of marine benthic communities, replacing the orthids. They reached their acme in diversity and dominance of communities in the Devonian. They suffered vast extinction in the Late Devonian, but rebounded in diversity and were still dominant members of communities in the late Paleozoic. Spiriferid diversity dropped dramatically following the late Permian mass extinction and never returned to Paleozoic levels. Spiriferids suffered another big decline in the Late Triassic mass extinction, and finally died out in the Middle Jurassic. See BRACHIOPODA; ARTIC-ULATA (ECHINODERMATA). Mark E. Patzkowsky

Bibliography. P. Copper and R. Gourvennec, Evolution of the spire-bearing brachiopods (Ordovician-Jurassic), in P. Copper and J. Jin (eds.), *Brachiopods*, A. A. Balkema, Rotterdam, 1996.

Spirometry

The measurement, by a form of gas meter, of volumes of gas that can be moved in or out of the lungs. The classical spirometer is a hollow cylinder (bell) closed at its top. With its open end immersed in a larger cylinder filled with water, it is suspended by a chain running over a pulley and attached to a counterweight (see **illus.**). The magnitude of a gas volume entering or leaving is proportional to the vertical excursion of the bell. In another type of spirometer, the bell is wedge shaped and is constrained to rotate on an axis located along its apex. Volume changes are proportional to the arc traversed.

Volume changes can also be determined from measurements of flow, or rate of volume change. Flow may be measured continuously by a pneumotachograph. A resistive element, such as a bundle of fine-



Diagram of a spirometer.

bore tubes, is inserted in the tube through which breathing occurs. The pressure drop across the resistance is proportional to flow and can be sensed and recorded continuously by a transducer that generates an electrical signal. The flow signal can be continuously integrated to yield a volume trace.

The volume of gas moved in or out with each breath is the tidal volume; the maximal possible value is the vital capacity. Even after the most complete expiration, a volume of gas that cannot be measured by the above methods, that is, the residual volume, remains in the lungs. It is usually measured by a gas dilution method or by an instrument that measures blood flow in the lungs. Lung volumes can also be estimated by radiological or optical methods.

At the end of an expiration during normal resting breathing, the muscles of breathing are minimally active. Passive (elastic and gravitational) forces of the lungs balance those of the chest wall. In this state the volume of gas in the lungs is the functional residual capacity or relaxation volume. Displacement from this volume requires energy from natural (breathing muscles) or artificial (mechanical) sources. *See* RES-PIRATION. Arthur B. Otis

Spirophorida

An order of sponges of the class Demospongiae, subclass Tetractinomorpha, with a globular shape and a skeleton of oxeas and triaenes as megascleres, and microspined, contorted, sigmalike microscleres.

Internal buds or gemmules may be formed. The eggs become affixed to substratum by means of long pseudopodia and undergo cleavage and morphogenesis without a free larval stage.

Species of *Tetilla* may have basal spicule tufts functioning to anchor the sponge in sandy bottoms. Species of *Cinachyra*, with ostia and oscules restricted to surface depressions, are common inshore inhabitants of coral reef lagoons. Members of this order range down to depths of at least 5900 ft (1800 m). *See* DEMOSPONGIAE. Willard D. Hartman

Spirurida

An order of nematodes in which the labial region is most frequently provided with two lateral labia or pseudolabia; in some taxa there are four or more lips; rarely lips are absent. Because of the variability in lip number, there is variation in the shape of the oral opening, which may or not be surrounded by teeth. The amphids are most often laterally located; however, in some taxa they may be located immediately posterior to the labia or pseudolabia. The stoma may be cylindrical and elongate or rudimentary. The esophagus is generally divisible into an anterior muscular portion and an elongate swollen posterior glandular region, where the multinucleate glands are located. Eclosion larvae are usually provided with a cephalic spine or hook and a porelike phasmid on the tail.

All known spirurid nematodes utilize an invertebrate in their life cycle; the definitive hosts are mammals, birds, reptiles, and, rarely, amphibians.



Fig. 1. *Thelazia platyptera*, a parasitic spiruroid nematode. (a) Anterior end of male, lateral view; (b) caudal end of male, lateral view; (c) caudal end of male, ventral view; (d) anterior end of female, lateral view; (e) caudal end of female, lateral view; (f) egg. (*Parasite Laboratory, Agricultural Research Service, Beltsville, Maryland*)

The order is divided among four superfamilies: Spiruroidea, Physalopteroidea, Filarioidea, and Drilonematoidea.

Spiruroidea. The superfamily Spiruroidea comprises parasitic nematodes whose life cycle always requires an intermediate host for larvae to the third stage. The definitive hosts are mammals, birds, fishes, reptiles, and rarely amphibians; and spiruroids may be located in the host's digestive tract, eye, or nasal cavity or in the female reproductive system (Table 1). Morphologically (Figs. 1 and 2), the lip region is very variable in Spiruroidea, ranging from four lips to none. When lips are present, the lateral lips are well developed and are referred to as pseudolabia. The cephalic and cervical region may be ornamented with cordons, collarettes, or cuticular rings. The stoma is always well developed and is often provided with teeth just inside the oral opening. In birds the nematodes are often associated with the gizzard, and the damage caused results in death, generally by starvation. When the muscles of the gizzard are destroyed, seeds pass intact and cannot be digested.

This superfamily contains the largest of all known nematodes, *Placentonema gigantissima*, parasitic in the placenta of sperm whales. Mature females attain a length of 26 ft (8 m) and a diameter of 1 in. (2.5 cm). They are unusual not only because of their great length but also because the adult female has 32 ovaries, which produce great numbers of eggs. The eggs are not unusually large. Armand Maggenti

Filarioidea. The superfamily Filarioidea contains highly specialized parasites of most groups of vertebrates. They are particularly common in amphibians, birds, and mammals. While they cannot be classified as completely harmless, most of the many hundreds of known species are not associated with any recognized disease. A limited number of species produce serious diseases in humans, and a few others produce serious diseases in domestic or wild animals (Table 2). The filarial parasites of humans are found almost exclusively in the tropics, with some extension into the subtropics. This is by no means the universal pattern. Heartworm disease in dogs has become established in the north temperate zone in the United States and Canada and in the south temperate zone in Australia. Several other enzootic species are known in northern amphibians, birds, and mammals. See HEARTWORMS.

Morphology of adults. Among the Filarioidea there are no conspicuous divisions into distinct body regions. True lips are absent, and the mouth cavity is greatly reduced. The cuticule is devoid of prominent ornaments found on many nematodes. The female genital opening is usually located in the region of the esophagus-intestine junction in the anterior portion of the worm. Sexual dimorphism is the rule; in common with other nematodes, the female filarioid is at least twice as long as the male, and often the difference is much greater. Among the species found in humans or dogs, the females vary from 1 to 28 in. (2.5 to 70 cm) long. Sensory organs, the cephalic papillae, are reduced in size and number. The eggs in a

Parasite	Definitive host	Intermediate host	Distribution	Pathology
Ascarops strongylina (Rudolphi, 1819)	Swine	Beetles	Cosmopolitan	Inflammation of stomach mucosa
Cheilospirura hamulosa (Diesing, 1851)*	Fowl, turkey	Grasshoppers, beetles	N. America, S. America, Asia, Europe	Soft nodules in muscul- ature of gizzard
Dispharynx nasuta (Rudolphi, 1819)	Fowl, turkey, pigeon, guinea fowl, pheasant	lsopods	N. America, S. America, Europe, Asia	Ulcers and gland des- truction in proven- triculus
Echinuria uncinata (Rudolphi, 1819)	Duck, goose, swan	Water fleas	N. America, Europe, Asia, N. Africa	Inflammation and nodules in proventriculus and gizzard
Draschia megastoma (Rudolphi, 1819)	Equines	<i>Musca</i> sp.	Cosmopolitan	Tumors in stomach wall
Draschia microstoma (Schneider, 1866)	Equines	Stable fly, housefly	Cosmopolitan	Catarrhal gastritis, stom- ach ulcers
Habronema muscae (Carter, 1861)	Equines	Musca sp.	Cosmopolitan	Catarrhal gastritis
Oxyspirura mansoni (Cobbold, 1879)	Fowl, turkey, peafowl	Cockroaches	Southern U.S., S. Asia, Australia, S. America	Lesions in eyes, blindness loss of eyeball
Tetrameres americana (Cram, 1927)	Fowl, turkey	Grasshoppers, cockroaches	N. America, S. America, Africa, Asia, Europe	Irritation and inflammation of proventriculus
Thelazia californiensis (Price, 1930)	Sheep, cat, dog	Fannia thelazia	California	Scar tissue formation in eyes, blindness
Thelazia callipaeda (Railliet and Henry)	Dog, rabbit	Unknown	Far East	Scar tissue formation in eyes, blindness
Thelazia rhodesii (Desmarest, 1828)	Cattle, sheep, goat, buffalo	<i>Musca</i> sp.	Europe, Asia, Africa	Scar tissue formation in eyes, blindness



Fig. 2. Oxyspirura mansoni, a common parasite of poultry. (a) Tail of male, lateral view; (b) tail of male, ventral view; (c) tail of female; (d) eggs. (*Parasite Laboratory, Agricultural Research Service, Beltsville, Maryland*)

few species have thick shells but more commonly are covered only by a thin vitelline membrane. The eggs characteristically undergo partial embryonation in the uterus, and precociously active embryos (microfilariae) are discharged. The adult worms are found in a wide variety of places in the body of the vertebrate host, but each particular species has its preferred host and preferred location within that host (Table 2).

Microfilariae. These precociously active embryos are devoid of any alimentary canal (Fig. 3). The different species range from less than 100 to more than 700 micrometers long and 2 to 8 μ m wide; they may or may not have a cephalic spear. The body is filled with a long column of somatic cells. Among them may be distinguished, with special staining, four cells variously termed R cells or G cells from which the alimentary canal and the reproductive system will develop. In some species the microfilariae are discharged still enclosed in the vitelline membrane and are termed "sheathed"; without the membrane they are "unsheathed." Most of the known forms have microfilariae in the blood regardless of the tissue in which the adults occur. These microfilariae may continue to circulate for months to years depending on the species and the level of host immune response. They do not necessarily circulate in constant numbers. Some species show a conspicuous daily periodicity, either nocturnal or diurnal, in the peripheral circulation. It has been shown with microfilariae of Wuchereria bancrofti of humans and with the unidentified microfilariae in the crow that the periodicity can be reversed in less than a week by reversing the active period (work, feeding, light, and so on) and the resting period (sleeping, darkness).

Species	Definitive host	Location in the host	Disease
Wuchereria bancrofti			
Periodic	Human	Deep lymphatics	Elephantiasis
Aperiodic	Human	Deep lymphatics	Elephantiasis
Brugia malayi	Human	Deep lymphatics	Elephantiasis
B. timori	Human	Deep lymphatics	Elephantiasis
B. pahangi	Wild felines, human	Deep lymphatics	Elephantiasis
Onchocerca volvulus	Human	Subcutaneous lymphatics	Fibrous-nodules, blindness
O. cervicalis	Horse	Ligamentun nuchae	Slight irritation
Loa loa	Human	Subcutaneous tissues, eyes	Calabar swellings (nonpitting edema)
Stephanofilaria stilesi	Cattle	Skin	Dermatitis
Elaeophora schneideri	Mule deer, sheep, moose, elk	Distal arteries	Dermatitis, "clear-eyed blindness"
Dipetalonema perstans	Human	Peritoneal and pleural cavities	Minor allergies
D. streptocerca	Chimpanzee	Subcutaneous	Slight edema
D. reconditum	Dog	Subcutaneous tissues	Nonreported
Mansonella ozzardi	Human	Body cavities	Very little
Dirofilaria immitis	Dog	Right heart	Chronic debilitation (heartworm disease
Dirofilaria repens	Dog	Subcutaneous tissues	Eczema
Litomosoides carini	Cotton rat	Pleural cavity	Pleural adhesions

Furthermore, an annual periodicity has been shown to occur with one species, the microfilariae of *Dirofilaria immitis*, in the dog, in the temperate zones; the microfilariae are most plentiful in the periph-



Fig. 3. Microfilariae. (a) Wuchereria bancrofti. (b-g) Posterior ends of the six species normally found in humans: (b) W. bancrofti; (c) Brugia malayi; (d) Mansonella ozzardi; (e) Loa loa; (f) Dipetalonema (Acanthocheilonema) perstans; (g) Onchocerca volvulus. Of these, b-f are found in blood, g is found in skin.

eral blood during the summer with its higher temperatures and longer periods of daylight. When the microfilariae are not in the peripheral circulation, they are sequestered in the capillary bed of internal organs, notably the lungs. They are incapable of further development as long as they remain in the definitive host.

Life cycle. All the known filariae require a bloodsucking arthropod intermediate host, usually an insect and commonly a dipteran, in which to complete embryonation. The microfilariae are ingested as the arthropod feeds. After embryonation is completed, the resulting infective larvae gain entrance into the definitive vertebrate in association with the next feeding of the arthropod (**Fig. 4**).

Filariasis. Filariasis is a disease caused by filarioidea in humans or lower animals. More loosely, the term is used to indicate mere infection by such organisms. In human medicine, filariasis commonly refers to the disease caused by, or to infection with, one of the mosquito-borne, elephantoid-producing Filarioidea—most frequently *W. bancrofti*, less frequently *Brugia malayi*, and more recently including *B. timori*.

1. *Diagnosis.* Serological and intradermal tests do not give any clue to specific identification. The only specific laboratory aid to diagnosis is the detection and identification of the microfilariae. *Brugia timori* is a sheathed microfilaria, but the sheath does not stain with the common blood-parasite stain, Giemsa. This microfilaria has the two detached terminal nuclei seen in *B. malayi* but is nearly half again as long as the latter, with the cephalic space nearly twice

	N	licrofilariae			
Intermediate host	Location	Sheath*	Periodicity	Geographical distribution	
"Domestic" night-biting mosquitoes	Blood	+	Nocturnal	Nearly worldwide in tropics	
Sylvan day-biting mosquitoes	Blood	+	Slightly diurnal	South Pacific islands	
Sylvan night-biting mosquitoes	Blood	+	Nocturnal	Southern Asia	
Domestic night-biting mosquitoes	Blood	+	Nocturnal	Indonesia	
Sylvan night-biting mosquitoes	Blood	+	Nocturnal	Southern Asia	
Simulium	Skin	-	-	Tropical Africa, Mexico, Guatemala, Venezuela, Columbia	
Culicoides	Skin	_	-	Cosmopolitan	
Chrysops	Blood	+	Diurnal	Equatorial Africa	
Unknown	Cutaneous lymph vessels	+	-	Southern and western U.S.	
Unknown	Skin	-	-	Mountains of southwestern U.S.	
Culicoides	Blood	_	-	Tropical Africa, South and Central	
Culicoides	Skin	_	_	Africa	
Eleas ticks (2)	Blood	_	_	Southern Europe North Africa LLS	
Culicoides	Blood			Central and South America. West Indie	
Mosquitoes	Blood	-	Nocturnal and seasonal	Tropics and subtropics, temperate zon in U.S.	
Mosquitoes	Blood	_	-	Southern Europe, North Africa, Vietnar	
Mites	Blood	+	-	Southern U.S.	

as long. The microfilariae of *Dipetalonema streptocerca*, a subcutaneous parasite of chimpanzees, are unsheathed and nearly half again as long as those of *D. perstans*. However, the posterior terminal column of nuclei is more than twice as long (with 9-12 nuclei) as in the latter, and the posterior end is curved like a shepherd's crook. The parasite has been reported frequently in humans in the rainforests of western Africa. Over 40% infection with microfilariae was shown by one set of skin snips from a small group of natives in Ghana.

2. Distribution and history. The species Wuchereria bancrofti has worldwide distribution in the tropics and extends into both north and south temperate zones. However, the very high infection rates and the high prevalence of the resulting grotesque elephantiasis (characterized by cutaneous and subcutaneous tissue enlargement due to lymphatic obstruction) are localized at sea-level areas with year-round uncontrolled high mosquito densities in the tropics. Introduction into the temperate zones commonly results in a few secondary cases before the infection dies out. This parasite is important also in medical and biological history. A century ago Patrick Manson noted that the common occurrence of tropical elephantiasis was restricted to areas where W. bancrofti was common (with 30-50% microfilaremia) but that microfilaremia commonly did not occur in the older subjects with advanced disease symptoms. This was one of the earliest indications of host (immune) response to repeated parasitic infections. Manson's demonstration that the microfilariae after ingestion with a blood meal continued embryonic development within the mosquito vector was also the first incrimination of mosquito transmission of any disease-causing organism.



Fig. 4. Life cycles of the Filarioidea.

3. Periodicity. In most areas of the world, the microfilariae of W. bancrofti have a conspicuous nocturnal periodicity. There are 50 to 150 times as many microfilariae in the midnight blood as in noon blood. Often none may be found in midday blood. However, this well-defined periodicity has been completely reversed within a week by reversing the human quiescent period (sleep and darkness) and the active period (exercising, eating, and daylight). This periodic variety of *W. bancrofti* is transmitted by domestic, anthropophilous, night-biting mosquitoes, which readily enter houses to feed. Culex fatigans (C. quinquefasciatus) and some of the important malaria vectors such as Anopheles darlingi, A. funestus, and A. farauti are important vectors. Nevertheless, even in areas dominated by the periodic parasite, individual cases of the aperiodic form occur. (In 1906 such a case was found in the Philippines, and since it was considered that the case might involve a distinct species, the specific name *philippinensis* was coined.)

On some of the geologically newer islands of the South Pacific east of the 176th meridian (Samoa, Marquesas, Fiji, and the Gilbert and Ellice among them) the dominant mosquitoes are sylvan daytime feeders. Nocturnal mosquitoes are rare or absent. In this environment an evolutionary change has taken place. The microfilariae have either no periodicity or only a slight diurnal periodicity. Aedes polynesiensis and close relatives are the only vectors, and the parasite has almost completely lost its capacity to develop in C. fatigans. This parasite is evidently evolving into a new species. Perhaps it has already evolved into distinct species, despite the fact that the loss of nocturnal periodicity is the only unequivocal distinction. There is some suggestion that acute febrile attacks after repeated bites by the mosquito intermediate host are more common than with the periodic forms. At any event, to the earlier specific name, philippinensis, pacifica was added in 1932.

4. *Control.* The fundamental difference in the epidemiology of the two forms is clearly depicted by the control measures. The periodic form has been successfully reduced or eliminated by use of bednets, household screening, residential insecticides, and control of the readily detected domestic breeding sites. The use of community-wide household spraying to eliminate malaria from Guyana also eliminated transmission of *W. bancrofti*. Acquired resistance to insecticides has complicated this approach.

Neither attempts at larval destruction nor attacks on adult mosquitoes have had any success in controlling the aperiodic infection. Only communitywide medication has had any success. Diethylcarbamazine (Hetrazan) in various regima has reduced the pool of microfilariae for mosquito infection and is active against the infective larvae or early developmental stages.

Brugia malayi occurs in both periodic and aperiodic strains, but it does not appear to be as completely isolated as is the case with *W. bancrofti*. Another parameter is added by the fact that species of *Mansonia* are the principal vectors. The larval stages of these mosquitoes are attached to submerged vegetation; herbicides have some value in control. Diethylcarbamizine has also been used. It appears to be very successful in areas where *B. timori* has been introduced recently.

Onchocerciasis. This disease is caused by Onchocerca volvulus in the subcutaneous lymphatics. It is characterized by subcutaneous nodules which are most conspicuous where the skin lies closely over bony structures, via cranium, pelvic girdle, joints, and shoulder blades. When they are on the head, the microfilariae reach the eyes. Ocular disturbances vary from mild transient bleary vision to total and permanent blindness. This is a major problem in the provinces of Oaxaca and Chiapas in Mexico and the nearby Pacific slopes of Guatemala, where infection is an occupational hazard associated with coffee growing.

Termed "river blindness," onchocerciasis also affects field workers in the Ivory Coast, Ghana, Mali, Togo, and Upper Volta. The World Health Organization has estimated that there are over a million cases in Upper Volta alone. Attempts to control the blackfly vector have had only some local successes. Surgical removal of the nodules is curative and on a communitywide basis eliminates or reduces transmission.

Loa loa. The African eye worm, Loa loa, is the filarioidan worm most commonly acquired by Caucasian immigrants, including missionaries, in Africa. Transmission is by daytime-feeding sylvan deer-flies, Chrysops. The only preventive measures are protective clothing, including head nets. Repellents have some value. Fortunately, serious damage is rare even when the worm gets into the eye. The areas of pitting edema known as calabar swellings are painful and diagnostic. They commonly occur on the wrists, hands, arms, or orbital tissues. Adult worms move freely. When the worm is accessible, surgical removal is indicated. This is one of the few Filarioidea in which adult worms are killed by diethylcarbamazine. See NEMATA. Gilbert F. Otto

Bibliography. H. W. Brown and F. A. Neva, *Basic Clinical Parasitology*, 6th ed., 1996; L. S. Garcia and D. A. Bruckner, *Diagnostic Medical Parasitology*, 4th ed., 2001; P. E. Sartwell (ed.), *Maxcy-Rosenau's Preventive Medicine and Public Healtb*, 11th ed., 1980.

Spitzer Space Telescope

A high-performance infrared telescope that is one of the four Great Observatories launched by the National Aeronautics and Space Administration (NASA) between 1990 and 2003. It takes advantage of dramatic advances in infrared detectors that have occurred over the last 20 years; it utilizes modern detector arrays in space, where they are limited only by the faint glow of the zodiacal dust cloud. Ground-based infrared telescopes can operate only at the wavelengths where the atmosphere is transparent, lying between 1 and 25 micrometers. Even within these windows, the thermal emission of the atmosphere is more than a million times greater than the dilute emission of the zodiacal cloud; there is additional foreground thermal emission from the telescope itself. High-sensitivity detectors are blinded by these bright foreground signals. Operating in space eliminates the atmospheric absorption and emission; also, a telescope in the vacuum of space can be cooled sufficiently to virtually eliminate its emission. *See* ZODI-ACAL LIGHT.

Spitzer was known for most of its life as *SIRTF* (*Space Infrared Telescope Facility*). *SIRTF* was launched on August 25, 2003, and was renamed after Lyman Spitzer, Jr., the first prominent astronomer to advocate putting telescopes into space.

Development. The Infrared Astronomy Satellite (IRAS) and the Infrared Space Observatory (ISO) were important predecessors of the Spitzer Telescope. IRAS, launched in 1983, had four sets of infrared detectors, operating at wavelengths of 12, 25, 60, and 100 μ m. The detectors and a 60-cm (24-in.) telescope were mounted inside a liquid-helium Dewar vessel that maintained them at a temperature of $\sim 2 \text{ K}$ (-456°F) for the 10-month mission, allowing a very sensitive all-sky survey. The European Space Agency built ISO to follow up on the IRAS survey. ISO too had a 60-cm telescope mounted in a liquid helium Dewar. It carried four instruments designed for detailed study of individual objects through imaging, photometry, and spectroscopy. ISO was launched in 1995 and had a 28-month mission that returned a wealth of data over the 3-200- μ m spectral range. See DEWAR FLASK.

The *Spitzer Telescope* was also conceived to follow up on *IRAS*. Although its science team was selected in 1984, the mission was repeatedly delayed. Throughout this period, the mission was revived repeatedly with new technical concepts. Finally, in 1996 the telescope got started officially toward construction under the direction of the Jet Propulsion Laboratory. By that time, NASA had downsized all its science missions under the guidelines of its "fasterbetter-cheaper" philosophy. Spitzer had been descoped by about a factor of 6 in cost from the original plans.

Design. The approved concept introduces a number of innovations developed in response to the pressure to reduce cost. It uses an Earth-trailing orbit to get far from the thermal radiation of the Earth (Fig. 1). The instruments are cooled inside a Dewar containing liquid helium, as with IRAS and ISO. However, the telescope is mounted on the outside of this Dewar and was launched warm. The very cold environment away from the Earth allows the outer shell of the satellite to cool passively, greatly reducing the heat load on the telescope. The thermal loads on the helium Dewar are also very small. This concept of a telescope launched warm that cools by radiating into space provides the technical foundation for future, larger infrared telescopes such as the James Webb Space Telescope (JWST).



Fig. 1. The orbit of *Spitzer*. The Earth orbits the Sun in a counterclockwise direction in this figure, and the frame of reference rotates with the Earth-Sun line to show how *Spitzer* trails behind. Viewed in a solar frame of reference, the *Spitzer* orbit is a normal Keplerian ellipse. In the Earth-centered frame here, the observatory makes a *loop* or *kink* every year when it partially overtakes the Earth. The observatory will fall farther and farther behind during its 5-year mission. The circles centered on the Earth are 0.2, 0.4, and 0.6 astronomical unit in radius, respectively.

The 85-cm (33.5-in.) telescope and 360-liter (95-gallon) Dewar are held off the spacecraft by a thermally insulating truss, and a thermal shield blocks radiation from the spacecraft (Fig. 2). The solar panel is cantilevered off the spacecraft, and the telescope is protected from its heat by another thermal shield. The Dewar protected the liquid helium from thermal loads while the observatory was on the ground. The instruments are mounted on top of the helium vessel, and a vacuum-tight aperture door sealed out the air when on the ground. Upon reaching orbit, the aperture door was opened to allow light from the telescope to reach the instruments. The outer shell cooled to $34 \text{ K} (-398^{\circ}\text{F})$ during the first month of the mission. Venting the cold helium vapor from the Dewar through heat exchangers cools the telescope further, to about 6° above absolute zero, or 6 K (-449° F). The extremely efficient thermal design is expected to yield a 5-year lifetime at these temperatures.

Spitzer has three instruments. The Infrared Array Camera (IRAC) images in bands at wavelengths of 3.6, 4.5, 5.8, and 8 μ m. The Infrared Spectrograph (IRS) provides spectra from 5 to 40 μ m at low resolution (1–2%) and from 10 to 38 μ m at moderate resolution (about 0.15%). The Multiband Imaging Photometer for *Spitzer* (MIPS) provides imaging at 24, 70, and 160 μ m. The instruments use advanced infrared detector arrays that operate at or near the fundamental natural background limit in space. The IRAC arrays are 256 × 256 pixels in format, using indium antimonide (InSb) photodiodes for the two shorter bands and silicon:arsenic (Si:As) impurity band conduction (IBC) devices for the


Fig. 2. Cutaway drawing of *Spitzer*, showing the major subsystems.

longer ones. IRS uses 128×128 arrays of Si:As and silicon:antimony (Si:Sb) IBC devices. MIPS uses a similar Si:As array at 24 μ m and germanium:gallium (Ge:Ga) arrays for the two longer bands. The high performance of these detector arrays allows *Spitzer* to deliver its anticipated breakthrough in science capability, despite its history of descopes. *ISO* had about 50 times more detectors than *IRAS*, and *Spitzer* has about 100 times more than *ISO*. The sensitivity per detector has also improved by an order of magnitude with each new mission.

Objectives. *Spitzer* is operated as a general-access observatory from the Jet Propulsion Laboratory and a science center at the California Institute of Technology. Eighty percent of the observing time is assigned by competitive application from the astronomical community. The science program includes investigations of star formation, planetary systems, and the assembly and evolution of galaxies.

Star formation. Young stars form in dense interstellar molecular clouds, carrying a heavy burden of interstellar dust. Initially, forming stars are cold, only tens

of degrees above absolute zero. Even after gravitational contraction has warmed them to more typical stellar temperatures, they are hidden by the surrounding cocoons of interstellar gas and dust and can be seen only in the infrared. Spitzer is mapping many nearby star-forming regions to identify very new stars that are too cold and faint to be known from other types of observation, documenting the very first stages of star formation. It is finding luminous massive stars and the outflows of gas that occur as accretion stops on these objects. It is also locating very-low-mass substellar brown dwarfs, revealing the processes by which stellarmass clumps in molecular clouds can fragment into much smaller pieces before collapsing. See BROWN DWARF.

Figure 3*a* shows a star-forming region about 14,000 light-years away that has a classic structure. Within the core of a molecular cloud, clumps of gas have collapsed into massive stars. Winds from these stars have cleared away the remaining gas, punching a hole into the cloud. From the *Spitzer* images, more than 300 newly formed stars have been identified in the central hole and surrounding molecular cloud. The remains of the cloud are heated by the young stars and glow at wavelengths of 4.5, 5.8, and 8 μ m in this image. *See* MOLECULAR CLOUD; STELLAR EVOLUTION.

Planetary systems. Embryo planets are believed to form during the first few million years of the life of their star, from a protoplanetary disk made of dusty and gaseous material with too much angular momentum to fall into the star itself. This stage in the planet formation process stops as the disk is dissipated by evaporation of grains of material within it, by accretion of material into the star or the embryo planets, and by ejection of material from the system. There-

after, planets grow by accretion of material from the embryos that is released when they collide with each other. A by-product of these collisions is circumstellar disks of dust and debris. The dust grains in the circumstellar disk are warmed by the star and glow in the infrared, where Spitzer can detect them. Initial Spitzer measurements of debris disks show a broad variety of behavior, indicating a few cases that appear to be dominated by recent and dramatic collisions. The disk signals damp down after about 150 million years. These observations are reminiscent of theories for the formation of the planets in the solar system, and in particular of the extreme bombardment of the Earth over the first few hundred million years of its existence. See EXTRASOLAR PLANETS; IN-TERPLANETARY MATTER; PLANET; PROTOSTAR; SOLAR SYSTEM.

Figure 3b and c shows Spitzer images of Fomalhaut, a bright star about 25 light-years away and about 200 million years old. The image of the star itself has been removed from the figure panels. Figure 3b, at a wavelength of 70 μ m, shows the dusty debris from a shattering collision of asteroid-sized embryo planets. A Hubble Space Telescope (HST) image of scattered light from this ring shows it to be off-center relative to the star, possibly maintained in this configuration by the gravitational action of a massive planet. Because the ring is closer to the star in the lower left of the 70- μ m image (Fig. 3b), the dust is warmer there and glows more brightly than elsewhere. At a wavelength of 24 μ m (Fig. 3c), the ring is filled in with warmer dust that is probably falling into the star. This dust is very dilute and is not seen in the Hubble image.

Assembly and evolution of galaxies. In nearby galaxies, the shortest-wavelength *Spitzer* bands reveal the distribution of the stars that dominate the visible mass,



Fig. 3. Illustrative science results from the Spitzer observatory. (a) Star formation in the nebula RCW 49 (courtesy of NASA/JPL-Caltech; E. Churchwell, University of Wisconsin). (b) Fomalhaut circumstellar disk at wavelengths of 70 μ m and (c) 24 μ m (courtesy of NASA/JPL-Caltech; K. Stapelfeldt, JPL). (d) Dusty, elliptical galaxy Centaurus A (courtesy of NASA/JPL-Caltech; J. Keene, Spitzer Science Center & Caltech).

improving understanding of the structure of the galaxies. The longer-wavelength bands glow in the infrared emission of various forms of interstellar dust, from tiny aromatic hydrocarbon powder heated to hundreds of degrees from the absorption of a single ultraviolet photon, to large grains at about 20 K $(-424^{\circ}F)$.

Figure 3*d* shows how *Spitzer* can see into the nearest large elliptical radio galaxy, Centaurus A, about 10 million light-years away. A striking trapezoidal structure glows in the 8- μ m band, which traces the distribution of polyaromatic hydrocarbon powder. A computer model indicates that this beautiful structure remains from a spiral galaxy that has been engulfed by the larger elliptical galaxy, a process that may have been responsible for triggering the nuclear activity and radio emission.

In the distant universe, Spitzer is locating a population of very luminous, dusty galaxies at redshifts of z = 1-4. These galaxies are identified because of their very red colors in the IRAC bands (resulting from the combination of high redshift and absorption by interstellar dust) or their detection in the MIPS bands, along with information on the redshift from spectra, or inferred from the IRAC colors. They represent a stage when smaller galaxies were merging at a high rate to form what have become large elliptical galaxies and spiral galaxy bulges. Spitzer is also detecting galaxies and quasars at even higher redshifts and will contribute to knowledge of the most distant objects known in the universe. See GALAXY, EXTERNAL; GALAXY FORMATION AND EVOLUTION; QUASAR. George Rieke

Bibliography. J. C. Mather (ed.), *Optical, Infrared,* and Millimeter Space Telescopes, Proc. SPIE, vol. 5487, 2004; G. H. Rieke, The Last of the Great Observatories: Spitzer and the Era of Faster, Better, Cheaper at NASA, University of Arizona Press, Tucson, 2006; Spitzer Space Telescope Mission (special issue), Astrophys. J. Suppl. Ser, 154(1):1-474, September 2004.

Splachnales

An order of the true mosses (subclass Bryidae), whose members are remarkable perennial plants that grow mainly on nitrogenous substrates, such as dung, and show considerable differentiation of neck tissue below the spore-bearing part of the capsule. Many of the species show adaptations that encourage insect dispersal of spores. The order consists of two families with about eight genera.

The plants are gregarious or dense-tufted, erect, and often forked. The leaves are soft, lanceolate or obovate, sometimes bordered at the margins, and commonly toothed. The single costa may be excurrent or terminate at or below the apex. The perichaetia are terminal, with bracts much like the upper stem leaves. The setae are elongate and the capsules erect with a noticeably differentiated neck. Stomata are very numerous in the neck, and have two guard cells. A single peristome is usually present, entire or forked, with 16 teeth sometimes paired or joined in twos and fours. In *Splachnum* the teeth are chambered since segments of the inner peristome lie opposite the teeth and are fused with them. The calyptrae are generally mitrate and sometimes hairy. *See* BRYIDAE; BRYOPHYTA; BRYOPSIDA. Howard Crum

Spleen

An organ of the circulatory system present in most vertebrates, lying in the abdominal cavity usually in close proximity to the left border of the stomach. It exhibits wide variation in size, shape, color, and location, depending upon species and age.

Anatomy. In humans the spleen normally measures about $1 \times 3 \times 5$ in. $(2 \times 7 \times 12 \text{ cm})$ and weighs less than 1/2 lb (230 g). It is a firm organ with an oval shape and is indented on its inner surface to form the hilum, or stalk of attachment to the peritoneum. This mesentery fold also carries the splenic artery and vein to the organ. The spleen is composed of a complex pulp lying between fibrous partitions, or trabeculae. These partitions project inwardly from the dense fibrous capsule that surrounds the organ.

Blood circulation is unusual in that the splenic arterioles open into thin-walled dilations, called sinusoids, which in turn drain into small veins. This arrangement of distensible blood-filled vessels, clusters of cells, and developing follicles accounts for the pulpy character of the spleen.

The pulp consists of a white pulp and a red pulp. The white pulp (lymphatic tissue) forms irregular nodules or follicles, often called Malpighian corpuscles and consisting of aggregations of developing lymphocytes. These nodules form initially around the small branches of the splenic arteries. In addition to the lymphocytes, other white cells such as monocytes, histiocytes, and giant cells are found in quantity. The red pulp consists of the sinusoids filled with blood and of splenic or Billroth cords filling the spaces between the sinusoids. Lymphocytes are found in the splenic cords that merge gradually into the white pulp. The distribution and amount of white and red pulp vary greatly, depending upon certain infections, intoxications, and disturbances in blood cell formation (anemia and leukemia) in the body. See ANEMIA; LEUKEMIA.

Function. The spleen is an important part of the blood-forming, or hematopoietic, system; it is also one of the largest lymphoid organs in the body and as such is involved in the defenses against disease attributed to the reticuloendothelial system. This system consists of many different types of cells in the body having the power to neutralize or engulf foreign particles or bacteria and sometimes to act as filters through which blood or lymph must pass. *See* HEMATOPOIESIS.

Although the chief functions of the spleen appear to be the production of lymphocytes, the probable formation of antibodies, and the destruction of wornout red blood cells, other less well-understood activities are known. For example, in some animals it may act as a reservoir for red blood cells, contracting from time to time to return these cells to the bloodstream as they are needed. In the fetus and sometimes in later life, the spleen may be a primary center for the formation of red blood cells. *See* ANTIBODY.

Another function of the spleen is its role in biligenesis. Because the spleen destroys erythrocytes, it is one of the sites where extrahepatic bilirubin is formed. Bilirubin is the principal pigment of bile formed by the reduction of biliverdin. Bile is normally present in the feces. *See* BILIRUBIN.

Lower vertebrates. In the lower vertebrates splenic tissue appears as scattered lymphoid masses in the wall of the digestive tract and apparently acts as the primary blood-forming organ. Bone marrow appears in certain amphibians and in most higher forms, taking over the formation of red cells in adults under normal conditions. The spleen then produces cells which are principally of the lymphoid type and engages in a number of other activities, some of which have been mentioned above.

This organ is not necessary to life for any vertebrate because its activities may be shared or assumed by other organs under proper conditions. Many of the splenic functions, such as lymphocyte formation, are common to all lymphoid tissue. Splenectomy may result in a temporary depression of antibody formation or by a recrudescence of latent or low-grade infection. *See* SPLEEN DISORDERS. Walter Bock

Spleen disorders

The spleen is rarely the site of primary disorders except those of vascular origin, but it is frequently involved in systemic inflammations, metabolic diseases, and generalized blood disorders.

Among vascular disturbances, acute and chronic congestion are prominent, particularly chronic congestion caused by cardiac failure, cirrhosis of the liver, and obstruction of the blood flow from the spleen by thrombi, scarring, or tumor tissue. Obstruction of the splenic artery or its branches by thrombi may result in an infarct caused by either cardiac or blood disease.

Inflammations include acute and chronic forms. The characteristic engorgement of blood often causes a marked enlargement of the organ. Bacteremias frequently produce this enlargement, or splenomegaly, and inflammation, but any severe infectious disease such as diphtheria or pneumonia may do so. Tuberculosis, syphilis, typhoid fever, and malaria, as well as many other specific infections, may cause splenomegaly, often with characteristic gross or microscopic changes for each disease. *See* MALARIA; SYPHILIS; TUBERCULOSIS.

The spleen is involved in certain types of pigment metabolism and is frequently the site of specific kinds of degeneration when amyloidosis, hemosiderosis, or argyrosis is present.

Similarly, in disorders of lipid metabolism, abnormal kinds or amounts of fat may appear in connection with Gaucher's disease, the hereditary Neimann-Pick disease, and others. *See* LIPID METABOLISM.

Because the spleen normally aids in the destruction of worn-out red blood cells and, under certain conditions, in the formation of red cells, it is not uncommon to find splenic involvement in blood disorders such as sickle-cell anemia, congenital hemolytic jaundice, polycythemia, and Mediterranean anemia. *See* HEMATOLOGIC DISORDERS.

The leukemias, especially when of the lymphocytic or neutrophilic varieties, cause some of the most prominent cases of splenomegaly as well as other changes. In myeloid leukemia, for instance, spleen weight of 13-18 lb (6000-8000 g) is not rare and the spleen may fill the entire abdomen. *See* LEUKEMIA.

Hypersplenism is an obscure disorder in which one or more of the blood cell types are destroyed to excess. Primary hypersplenism results from unknown causes; secondary hypersplenism may follow inflammatory diseases, chronic congestion, or tumor invasion.

Tumors originating in the spleen are rare and usually limited to such benign growths as hemangiomas, lymphangiomas, and fibromas, but malignant lymphomas and lymphosarcomas also occur.

Secondary tumors, which originate elsewhere and metastasize to the spleen, are not uncommon, particularly the lymphoma group. Other varieties are seen less often and include carcinomas, particularly from a bloodstream invasion or growth in the neighboring stomach or intestine. *See* TUMOR.

Trauma to the spleen is more common than suspected, and the consequences of rupture of the organ are often tragic because this may follow an apparently moderate abdominal injury.

Many other conditions produce splenic change, notably splenomegaly of some degree. The presence of an enlarged spleen indicates that a thorough diagnostic evaluation should be made. *See* SPLEEN.

Edward G. Stuart; N. Karle Mottet

Splines

A series of projections and slots used instead of a key to prevent relative rotation of cylindrically fitted machine parts. Splines are several projections machined on the shaft; the shaft fits into a mating bore called a spline fitting. Splines are made in two forms, square and involute (see **illus.**). Since there are several projections (integral keys) to share the force in transmitting power, the splines can be shallow, thereby not weakening the shaft as much as a standard key.

Square splines have 4, 6, 10, or 16 splines. The external part (shaft) may have the splines formed by milling and the internal part (bore) by broaching. Three classes of fits are used: sliding (as for gear



Diagrams of (a) square spline and (b) involute spline profile.

shifting) under load, sliding when not loaded, and permanent fit. Square splines have been used extensively for machine parts. In the automotive industry, square splines have been replaced generally by involute splines, which cost less to make accurately for good fit and interchangeability.

Involute splines are used to prevent relative rotation of cylindrically fitted machine parts and have the same functional characteristics as square splines. The involute spline, however, is like an involute gear, and the spline fitting (internal part) is like a mating internal gear. Profiles are the same as for gear teeth of the stub (fractional pitch) form with 30° pressure angle.

Involute splines on the shaft are generated by a hob or a gear shaper, and internal splines are formed by a broach or a gear shaper. Three classes of fit are standard: sliding, close, and press.

Involute serrations are similar to involute splines except that the pressure angle is 45° and, while there are three standard fits (loose, close, and press), serrations are usually press-fitted and used for permanent assembly. They are used for both parallel and tapered shafts. *See* MACHINE KEY. Paul H. Black

Bibliography. American National Standards Institute (ANSI), *Involute Splines and Inspections, Inch Version*, B92.1-1970; ANSI, *Involute Splines and Inspections, Metric Version*, B92.1M-1970; ANSI, *Involute Splines, Metric Modules*, B92.2M-1981; C. W. Ham et al., *Mechanics of Machinery*, 4th ed., 1958; C. R. Mischke and J. E. Shigley, *Mechanical Engineering Design*, 6th ed., 2000; G. H. Ryder and M. D. Bennnett, *Mechanics of Machines*, 2d ed., 1990. Dakota. This implies the remarkable ability of a single crystal to replace a large variety of preexisiting minerals and yet maintain the integrity of a single crystal, a crystal growth unequaled elsewhere in nature.

Spodumene is usually found as a constituent in certain granitic pegmatites in association with quartz, alkali feldspars, mica, beryl, phosphates, and a large variety of rare minerals. It is also known to occur as disseminated grains in some granite gneisses. Spodumene often alters to a fibrous mass composed of eucryptite LiAlSiO₄ and albite, or eucryptite and muscovite. The emerald-green variety, hiddenite, and a lilac variety, kunzite, are used as precious stones. Spodumene from pegmatites is used as an ore for lithium. *See* LITHIUM; PYROXENE. George W. DeVore

Spongiomorphida

An extinct group of genera of Mesozoic fossils. Informally known as the spongiomorphs, they were established as a family of scleractinian corals by Fritz Frech in 1890 on the basis of specimens from the Zlambach beds of late Triassic age in Austria. He grouped the fossils into four genera (Spongiomorpha, Heptastylopsis, Heptastylis, and Stromatomorpha). The fossils consist of closely spaced longitudinal rods of calcite or aragonite that are perpendicular to the growth surface and that give off lateral expansions which join the rods and in some forms unite to form laminae parallel to the growth surface (see illustration). Frech noted the similarity of his group of fossils to the stromatoporoids, particularly to the genus Actinostroma, and named one of the species Heptastylis stromatoporoides. In tangential section, the rods of the spongiomorphs have a central dark dot, if round, or a line, if oval, and concentric microstructure. Their fibrous microstructure suggests that they

Spodumene

The name given to the monoclinic lithium pyroxene LiAl(SiO₃)₂. Spodumene commonly occurs as white to yellowish prismatic crystals, often with a "woody" appearance, exhibiting the 87° pyroxene (110) cleavages. It is easily identified during heating in a flame by the red color given off, accompanied by a marked swelling of the fragment. Spodumene usually contains an appreciable quantity of hydrogen substituting for lithium. At 720°C (1330°F), spodumene inverts to a tetragonal form, β -spodumene, which is accompanied by a 30% increase in volume. Spodumene is capable of forming immense crystals in nature. A single crystal 47 ft (14.3 m) in length and 5 ft (1.5 m) in diameter, as well as others almost as large, has been found at the Etta mine in South



Spongiomorpha acyclica (Triassic, Austria), length 8 cm. (From J. P. Smith, Upper Triassic Marine Invertebrate Faunas of North America: Spongiomorphs, U.S. Geol. Surv. Prof. Pap. 1927)

were originally aragonite, and their water-jet appearance in longitudinal section is similar to that of the septa of scleractinian corals. Some spongiomorphs from the middle Triassic Cassian beds still preserve the aragonite mineralogy of the living corals.

In some spongiomorphs, the rods in tangential sections seem to be grouped in sixfold symmetry about a central rod with their axial dark lines radiating away from the central rod. This radial pattern was thought by some investigators to be homologous to the astrorhizal systems of the Paleozoic stromatoporoids. To those who believed the stromatoporoids were hydrozoans, this radial arrangement was taken as evidence that the spongiomorphs were hydrozoans. However, the radial pattern is not produced by convergent canals, as in the stromatoporoids, but by the alignment of rods. Further, the hexameral radial symmetry is typical of the Mesozoic corals, not of hydrozoans. Frech recognized the similarity of his specimens to the scleractinian coral genus Astraeomorpha of the family Poritidae. In this family of living corals, the septa tend to break down into a series of rods, and the boundaries of the corallites disappear to produce a structure very similar to the spongiomorphs. The affinity of Frech's genera to the poritid corals has been confirmed by the work of P. Gautret and J.-P. Cuif on the microstructure of these fossils.

The term spongiomorph should thus be used only in an informal sense for the group of genera in which the septa are reduced to closely spaced rods and the radial structure and individuality of the corallites is obscure.

Spongiomorph corals are known from the Middle Triassic of the Alps; the Upper Triassic of Austria, Bulgaria, Romania, the United States (California, Alaska, Oregon), Turkey, Morocco, Iran, Japan, south China, central Asia, and Tajikistan; the Upper Jurassic of China (Tibet), Austria, Spain, and Japan; and the Upper Cretaceous of Sweden and the West Indies. *See* SCLERACTINIA; SCLEROSPONGE; SPHAERACTI-NOIDEA; STROMATOPOROIDEA. C. W. Stearn

Bibliography. E. Flügel and E. Sy, Die Hydrozoen der Trias, *Neues Jahrbuch für Geologie und Paläontologie, Abhandlungen*, 109:1-108, 1959; P. Gautret, F. Ezzoubair, and J.-P. Cuif, Recherche sur les affinités des spongiomorphidae Frech, 1890—I: Charactéristiques microstructurales et minérologiques de *Spongiomorpha acyclica* Frech, 1890, *Geobios*, 25(3):345-355, 1992; J. P. Smith, Upper Triassic marine invertebrate faunas of North America: Spongiomorphs, *U.S. Geol. Surv. Prof. Pap.*, 141:132-135, 1927; G. D. Stanley and M. T. Whalen, Triassic corals and spongiomorphs from Hell's Canyon, Wallowa Terrane, Oregon, *J. Paleontol.*, 63:800-819, 1989.

Sporozoa

A subphylum of Protozoa, typically with spores. The spores are simple and have no polar filaments. There is a single type of nucleus. There are no cilia or flagella except for flagellated microgametes in some groups. In most sporozoa there is an alternation of sexual and asexual stages in the life cycle. In the sexual stage, fertilization is by syngamy, that is, the union of male and female gametes. All sporozoa are parasitic. *See* PROTOZOA.

Taxonomy. The subphylum is divided into three classes—Telosporea, Toxoplasmea, and Haplosporea.

Telosporea. The class Telosporea is by far the largest and most important. It contains two subclasses: members of the subclass Gregarinia, which is composed of three orders (Archigregarinida, Eugregarinida, and Neogregarinida), are parasites of insects and other invertebrates; members of the subclass Coccidia, which is composed of two orders (Protococcida and Eucoccida), are parasites mostly of vertebrates, although a few are found in invertebrates. *See* INSECT DISEASES; TELOSPOREA.

The order Eucoccida is the largest in the subclass Coccidia. Two of its suborders (Eimeriina and Haemosporina) are of great importance. The suborder Eimeriina contains several hundred species of coccidia, which cause disease in various domestic and wild animals. Most of them multiply in the cells of the intestine and cause diarrhea, dysentery, and even death. They are monoxenous, multiplying asexually by schizogony in their host cells and eventually producing gametes which unite to form oocysts that pass out in the feces, mature on the ground, and infect other animals when eaten. A typical species is *Eimeria tenella*, which occurs in the ceca of chickens and may cause their death.

The suborder Haemosporina includes the malaria parasites and related forms. Their life cycle is basically similar to that of the Eimeriina, but they generally multiply asexually in vertebrate blood cells and produce sexual stages in invertebrate hosts. A typical species is *Plasmodium vivax*, which causes benign tertian malaria in humans (called benign because it seldom kills its victims, in contrast to *P. falciparum*, which causes malignant tertian malaria and has a relatively high death rate). *See* MALARIA.

Spores are present in the Telosporea, and reproduction is typically both sexual and asexual, with alternation of generations. Some species occur only in a single type of host (monoxenous), while others require two types of host (heteroxenous), with sexual development in one and asexual in the other. Some groups have flagellated microgametes; if these are missing, locomotion is by body flexion or gliding; pseudopods are ordinarily absent, but if present are used for feeding, not locomotion. The name Telosporea comes from Greek words which indicate that spore formation occurs at the end of the life cycle.

Toxoplasmea. The class Toxoplasmea contains "sporozoa" which lack spores and sex. There are no pseudopods or flagella at any stage, and locomotion is by body flexion or gliding. Cysts or pseudocysts (with a wall formed by the host cell) are formed and contain many trophozoites. Species in this class have a single type of host.



Fig. 1. Fine structure of a coccidian merozoite (*Eimeria miyairii* from the rat). (a) Longitudinal section. Cross sections at levels of (b) nucleus, (c) Golgi complex, (d) large dense granules, (e) cytostome, (f) paired organelle, and (g) conoid. (After J. Andreassen and O. Behnke, J. Parasitol., 54:150–163, 1968)

There seem to be only three genera in this class, all of which parasitize vertebrates. Toxoplasma occurs mostly in mammals. By far the most important species is T. gondii, which is common throughout the world and causes toxoplasmosis in humans and many lower mammals. Most infected individuals show no symptoms, but the disease has various manifestations ranging up to acute, fatal encephalitis in newborn infants. Its normal mode of transmission is still unknown. Sarcocystis is found in many mammals and birds, especially sheep, cattle, swine, and ducks. The cysts, which may be visible to the naked eye, are found in the muscles. Hunters sometimes see them in wild ducks, where they make the breast muscles appear striated. They are harmless when cooked and eaten, but they are esthetically undesirable, and heavily infected ducks are rarely eaten. The third genus, Besnoitia, is relatively uncommon, although one species may cause a skin disease of cattle and another may pit the bones of reindeer. See TOXOPLASMEA.

Haplosporea. It is not certain whether the last class, Haplosporea, really belongs in this subphylum. Its members have spores but apparently do not reproduce sexually. They have no flagella, but pseudopods may be present. Some people think that this class belongs with the Cnidospora, even though its members lack a polar filament, while others think it may be related to the fungi. At any rate, there are only a few species in the group, and they are mostly parasites of mollusks, annelids, and other invertebrates. *See* CNIDOSPORA; HAPLOSPOREA.

Structure. The structure of the Sporozoa varies with the species and stage of development, and

the variations are too diverse to be given here. At one stage or another all (except perhaps the Haplosporida) have a characteristic conoid and polar ring at the anterior end. Subpellicular fibrils or microtubules extend almost the whole length of the body just beneath the pellicle and are presumably responsible for locomotion. There is a Golgi apparatus which presumably has to do with secretion or excretion, one or more mitochondria which have to do with metabolism, a number of micronemes, rodshaped granules, convoluted tubules, sarconemes or lankesterellonemes extending backward in the anterior part of the organism, a paired organelle or series of toxonemes originating inside the conoid and



Fig. 2. Mature (sporulated) oocyst of a coccidium (*Eimeria*). (*After N.D. Levine*, *Protozoan Parasites of Domestic Animals and Man*, *Burgess*, 1961)

running backward in the cytoplasm, one or more micropores (cytostomes, micropyles) in the side of the body, several types of granule in the cytoplasm, and endoplasmic reticulum (**Fig. 1**). The electron microscope is needed to see most of these structures.

The Telosporea have an oocyst which is produced by fusion of male and female zygotes. Its structures can be seen with the light microscope. The characters of the oocyst vary with the species. In the Eimerina there are generally one or more sporocysts within the oocyst, and one or more sporozoites in each sporocyst. The oocyst may have a micropyle, which may be capped. There may be a polar granule within the oocyst and its wall may be composed of one or more layers. There may be residual material in the oocyst and sporocyst, left over after formation of the sporocysts or sporozoites (**Fig. 2**).

Life cycle. The life cycle of a typical telosporean sporozoon involves alternation of sexual and asexual generations. The sexual forms may occur in one type of host and the asexual forms in another, or there may be only a single type of host. The latter situation is characteristic of the eimeriin Eucoccida. A typical example is Eimeria tenella, which lives in the epithelial cells of the chicken cecum (Fig. 3). When a chicken eats a mature (sporulated) oocyst, the oocyst breaks in the gizzard, releasing the sporozoites. A sporozoite (stage 1 in Fig. 3) enters an intestinal epithelial cell (2) and rounds up and turns into a first-generation schizont (3), which forms a number of merozoites (4) by multiple fission. These break out of the host cell (5), enter new host intestinal epithelial cells (6), and turn into second-generation schizonts (7, 8), which produce second-generation merozoites (9, 10), which break out of the host cell (11), and may form either third-generation schizonts (12, 13), which produce third-generation merozoites (14, 15), or gamonts. All third-generation schizonts produce gamonts. The gamonts are either male or female. The male ones (16, 17) produce a large number of flagellated microgametes (18), while the female ones simply grow, becoming macrogametes (19, 20). The microgametes fertilize the macrogametes, and the resultant zygote (21) rounds up, secretes a wall around itself to become an immature oocyst, and leaves the host cell (22). It passes out of the chicken and becomes mature (sporulates) on the ground. The sporont undergoes meiosis and becomes haploid, throws off a polar body, and forms four sporoblasts (23), each of which forms a sporocyst containing two sporozoites. When this sporulated oocyst is eaten by the chicken, the sporozoites are released in the gut and the cycle begins again.

Eucoccida. In the suborder Eimeriina, since each species has a definite number of asexual generations, the infection is self-limiting. This is not true in the suborder Haemosporina, which contains the malaria parasites. In this group the number of asexual generations is indefinite, a moderate number of microgametes is formed, asexual stages occur in a vertebrate host, and sexual stages and sporozoite formation occur in an invertebrate, usually an insect. In human malaria, for instance, the vertebrate host



Fig. 3. Life cycle of the chicken coccidium *Eimeria tenella*. The numbered steps are explained in the text. (After N. D. Levine, Protozoan Parasites of Domestic Animals and Man, Burgess, 1961)

is a human and the invertebrate host an *Anopheles* mosquito. The vertebrate host is infected when the invertebrate host bites it. In the third suborder of this group, the Adeleina, few microgametes are formed and they are produced when the male and female gamonts are attached to each other, that is, when they are in syzygy.

Gregarinia. In the second large telosporean subclass, the Gregarinia, the mature trophozoites (vegetative stages) are large and extracellular. They occur in the digestive tract or body cavity of invertebrates. Typically (in the order Eugregarinida), no schizonts are formed, the trophozoites producing gamonts and then gametes, which fuse to form a zygote which produces an oocyst and then a number of sporozoites.

Toxoplasmea. In the class Toxoplasmea there is no sexual reproduction. The trophoxoites divide by external or internal binary or multiple fission, and cysts or pseudocysts with many naked trophozoites are formed. Norman D. Levine

Bibliography. J. O. Corliss, *The Ciliated Protozoa*: *Characterization, Classification, and Guide to the Literature*, 2d ed., 1979; P. C. C. Garnham, *Malaria Parasites and Other Haemosporidia*, 1966; R. R. Kudo, *Protozoology*, 5th ed., 1977; N. D. Levine, *Protozoan Parasites of Domestic Animals and of Man*, 2d ed., 1973; N. D. Levine, *Textbook of Veterinary Parasitology*, 1978; N. D. Levine et al., A newly revised classification of the protozoa, *J. Protozool.*, 27:35-58, 1980; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982.

Sports medicine

A branch of medicine concerned with the effects of exercise and sports on the human body, including treatment of injuries. Sports medicine can be divided into three general areas: clinical sports medicine, sports surgery, and the physiology of exercise. Clinical sports medicine includes the prevention and treatment of athletic injuries and the design of exercise and nutrition programs for maintaining peak physical performance. Sports surgery is also concerned with the treatment of injuries from contact (human or object) sports. Exercise physiology, a growing field of sports medicine, involves the study of the body's response to physical stress. It comprises the science of fitness, the preservation of fitness, and the role of fitness in the prevention and treatment of disease.

The various aspects of sports medicine touch on many different branches of medicine and surgery, including cardiology, neurology, orthopedics, pulmonary medicine, and endocrinology. Nevertheless, sports medicine is becoming a distinct discipline.

Advances in instrumentation also led to increased understanding of cardiac function and its response to exercise. The pressure drum and sphygmograph mechanically transmit and record arterial pulses. The modern electrocardiogram and other techniques now give information on the mechanical and electrical properties of heart action at rest and during exercise.

Investigations of the effects of exercise on systems other than the heart and lungs were initiated in the 1930s. The notion of exercise as a route to good health became widespread during the latter part of the nineteenth century.

Diagnosis and therapy. The health problems arising from sports include injuries resulting from overuse, such as ill-defined aches and pains, tendinitis, and chondromalacia (softening of the cartilage), one cause of "runner's knee." Such injuries are becoming increasingly common as more people engage in amateur athletics; they also afflict some musicians who develop overuse of muscle-tendon units resulting in pain of the hand and fingers, commonly seen in string players. Sports medicine also treats acute ailments such as dehydration and heat stroke, injury to the musculoskeletal system, the eyes, the brain and spinal cord, and (rarely) cardiac arrest. The usual therapy for overuse injuries is rest. Injuries caused by contact can require urgent surgical care. See PER-FORMING ARTS MEDICINE.

Strategies to prevent injuries and maintain peak performance have become an important part of sports medicine. Careful attention to diet and fluid intake is routine for professional and amateur athletes during both training and sports events. (The use of performance-enhancing drugs, such as steroids and stimulants, by some athletes is an unfortunate result of the competition for improved athletic performance.) Studies of biomechanics have led to preventive measures such as improved helmets, shoes, and supportive clothing. *See* BIOMECHANICS; STEROID.

Exercise physiology. This subdivision of sports medicine is involved with the responses of healthy individuals to physical stress and aging in all parts of the body. Hormonal, cardiovascular, renal, neurologic (including mental function), musculoskeletal, and other bodily systems are studied. Modern technologies are applied, because many are noninvasive and allow repeated measurements.

Exercise and wellness. The role of sports in the prevention of disease and the improvement of health is a major concern of sports medicine. It is widely believed that physical fitness can help ward off many of the illnesses common in industrial society, mainly obesity and cardiovascular disease but also osteoporosis (brittleness of bones) and such chronic complaints as back pain and headaches. In addition, it is believed that a regimen in which regular exercise is accompanied by beneficial habits, such as a diet that is high in fiber and low in fats, moderate consumption of alcohol, and an avoidance of smoking, can lower the incidence of some illnesses, including cancers.

Conclusive evidence of exercise's capacity to prevent disease is lacking, but the circumstantial evidence is strong. Weight loss, lowered cholesterol, and improved neurologic, gastrointestinal, and general body function all seem to contribute to this effect. Otto Appenzeller

Bibliography. O. Appenzeller (ed.), *Sports Medicine: Training, Fitness, Injuries,* 3d ed., 1988; E. Jokl, Running, psychology and culture, *Ann. N. Y. Acad. Sci.*, 301:970–1001, 1977.

Spot welding

A resistance-welding process in which coalescence is produced by the flow of electric current through the resistance of metals held together under pressure. A low-voltage, high-current energy source is required (**Fig. 1**). Usually the upper electrode moves and applies the clamping force. Pressure must be maintained at all times during the heating cycle to prevent flashing at the electrode faces. Electrodes are



Fig. 1. Spot-welding circuit. When electrodes are closed on the workpiece, the circuit is completed for welding.



Fig. 2. Distribution of temperature in local (numbered) elements of a spot-welding operation.

water-cooled and are made of copper alloys because pure copper is soft and deforms under pressure. *See* RESISTANCE WELDING.

The electric current flows through at least seven resistances connected in series for any one weld. They are (1) upper electrode, (2) contact between upper electrode and upper sheet, (3) body of upper sheet, (4) contact between interfaces of sheets,(5) body of lower sheet, (6) contact between lower sheet and electrode, and (7) lower electrode (**Fig. 2**). Heat generated in each of the seven sections will be in proportion to the resistance of each. The greatest resistance is at the interfaces in 4, and heat is most rapidly developed there. The liquid-cooled electrodes 1 and 7 rapidly dissipate the heat generated at the contact between electrodes and sheets 2 and 6, and thus contain the metal that is heated to fusion temperature at the interfaces.

After the metals have been fused together, the electrodes usually remain in place sufficiently long to cool the weld. An exception is in welding quenchsensitive metals, where it is desirable to remove the electrodes as soon as possible to allow the heat to be conducted in the surrounding metal, preventing steep quench gradients. *See* WELDING AND CUTTING OF MATERIALS. Eugene J. Limpel

Bibliography. L. Jeffus, *Welding: Principles and Applications*, 5th ed., 2002; H. Zhang and J. Senkara, *Resistance Welding*, 2005.

Spray flow

A special case of a two-phase (gas and liquid) flow in which the liquid phase is the dispersed phase and exists in the form of many droplets. The gas phase is the continuous phase, so abstract continuous lines (or surfaces) can be constructed through the gas at any instant without intersection of the droplets. The droplets and the gas have velocities that can be different, so both phases can move through some fixed volume or chamber and the droplets can move relative to the surrounding gas. *See* GAS; LIQUID; PARTIC-ULATES; TWO-PHASE FLOW.

Applications. Spray flows have many applications. Sprays are used to introduce liquid fuel into the combustion chambers of diesel engines, turbojet engines, liquid-propellant rocket engines, and oilburning furnaces. They are used in agricultural and household applications of insecticides and pesticides, for materials and chemicals processing, for fire extinguishing, for cooling in heat exchangers, for application of medicines, and for application of coatings (including paint and various other types of layered coatings). Common liquids (such as water, fuels, and paints) are used in sprays. It is sometimes useful to spray uncommon liquids such as molten metals. In the various applications, the approximately spherical droplets typically have submillimeter diameters that can be as small as a few micrometers. See METAL COATINGS.

Formation. Sprays are formed for industrial, commercial, agricultural, and power generation purposes by injection of a liquid stream into a gaseous environment. In addition, sprays can form naturally in a falling or splashing liquid. Injected streams of liquid tend to become unstable when the dynamic pressure (one-half of the gas density times the square of the liquid velocity) is much larger than the coefficient of surface tension divided by the transverse dimension. Typically, the liquid stream disintegrates into ligaments (coarse droplets) and then into many smaller spherical droplets. The breakup (or atomization) process is faster at higher stream velocity, and the final droplet sizes are smaller for higher stream velocities. Spray droplet sizes vary and typically are represented statistically by a distribution function. The number of droplets in a spray can be as high as a few million in a volume smaller than a liter. See ATOMIZATION; JET FLOW; SURFACE TENSION.

Drag force on droplets. The droplets experience a drag force as they move relative to the gas. The magnitude and direction of the drag depend on the magnitude and direction of the relative velocity between the gas and the droplet. The drag force is proportional to the square of the droplet diameter, while the droplet mass is proportional to the cube of the diameter; therefore, the acceleration, which equals the drag per unit mass, is inversely proportional to the diameter. Drag causes the relative velocity to decrease faster for smaller droplets. Droplets of only a few micrometers in size move at the gas velocity. *See* FLUID-FLOW PRINCIPLES.

Flow in and around droplets. A droplet moving relative to the surrounding gas behaves similarly to an aerodynamic body. An observer fixed to the droplet would see that the gas flow first decelerates as it moves toward the droplet and then accelerates as it passes the droplet. The **illustration** shows a cross section of the flow around and within a droplet. The frictional force at the gas-liquid interface retards the flow in the gas layer at the droplet surface. The approximate thickness of this boundary layer is proportional to the droplet radius divided by the square root of the Reynolds number. The nondimensional Reynolds number is the triple product of the gas



Cross section of a vaporizing droplet, showing the relative gas-droplet motion and internal circulation.

density, relative velocity, and droplet radius divided by the coefficient of gas viscosity. For droplets, the range of the Reynolds number is 0-200. The flow stream in the boundary layer can separate behind the droplet surface, causing a wake with recirculating flow in the gas. The wake formation causes a pressure drop across the droplet. This pressure difference and the friction force result in the drag force. The friction force exerted on the liquid surface of the droplet causes the internal liquid to circulate. The streamlines in the illustration show the spherical vortex pattern that emerges, with recirculation through a region called the internal wake. The convective current affects the transport of heat to the droplet interior from its surface region. See BOUNDARY LAYER FLOW; REYNOLDS NUMBER; WAKE FLOW.

Droplet trajectories. The trajectories of the droplets in a spray depend on the aerodynamic forces (lift and drag) and the force of gravity that they experience. The aerodynamic forces on a droplet are modified by the proximity of neighboring droplets. The drag force is reduced for a droplet moving in tandem behind another droplet. The wake behind the first droplet is moving in the same direction as the second droplet, thereby reducing the relative velocity and the drag force. (This same phenomenon occurs for trucks moving in a convoy.) Droplets moving side by side can experience lift forces, causing them to either attract or repel each other. The lift force is caused by the modification of the pressure due to the Bernoulli effect as the gas flows between the two droplets. As the two droplets come closer, the gas velocity first increases, reducing the inside pressure below the outside pressure, thereby pushing the droplets toward each other. As the droplets come still closer, blockage of the inside flow occurs, increasing the inside pressure above the outside pressure and pushing the droplets apart. The aerodynamic forces can cause droplets to collide but, generally, the droplets

in a spray are diverging. The conduction and convection of heat and the diffusion of vapors are also affected by neighboring droplets. *See* AERODYNAMIC FORCE.

Heat transfer and vaporization. In many applications (such as combustors and heat exchangers), a high-temperature environment exists in the gas surrounding each droplet of a spray. Substantial heat transfer to the liquid surface results; some of the heat is required to vaporize the liquid at the surface, and the remainder of the heat is transferred to the droplet interior, thereby raising its temperature with time. The droplet lifetime is defined as the time required to vaporize the total droplet after its introduction to the hot environment. The droplet temperature is lower than the ambient gas temperature, with continuous increase in the gas temperature through a thermal gas layer surrounding the droplet. An increase in the relative velocity, or equivalently in the Reynolds number, causes the thickness of the thermal layer to decrease, thereby increasing the heating rate and the vaporization rate and decreasing the droplet lifetime. See HEAT TRANSFER.

The temperature of a rapidly vaporizing droplet (for example, a fuel droplet in a combustor) will vary throughout its lifetime in a transient manner. The internal liquid circulation provides convective augmentation to the conduction of heat from the droplet surface toward the interior. A temperature variation through the liquid interior can remain during the lifetime of the droplet. For a droplet vaporizing slowly in a low-temperature gas environment, the droplet temperature can be approximately uniform through the interior at the theoretical maximum temperature, which is called the wet-bulb temperature. A submillimeter-size water or fuel droplet in a gas environment with a temperature above 1000 K (1340°F) can vaporize in milliseconds. Vaporization time varies with droplet diameter to a power between 1.5 and 2.0.

The vapor that comes from the droplet is typically different chemically from the surrounding gas. It must convect and diffuse away from the droplet and mix with the surrounding gas. Sometimes the vapor also reacts with the surrounding (for example, fuel vapor in a heated air environment). *See* EVAPO-RATION.

Supercritical behavior. At very high pressures and temperatures, the distinction between a vapor and its liquid disappears. That is, above the critical temperature and critical pressure values, the addition of heat results in a continuous change in density rather than the discontinuous phase change (vaporization) that is experienced at lower pressures and temperatures. In many applications, a liquid stream is injected into a chamber at supercritical pressure. If the initial liquid temperature is subcritical but the pressure through the gas film and the ambient gas temperature are supercritical, heating of the liquid can result, with a continuous change of density as the temperature increases from subcritical to supercritical values. The gas mixture in the film surrounding the droplet can have substantially higher critical pressures than

either the pure vapor or the ambient gas. Sometimes the gas film is subcritical even if the ambient gas is supercritical. Vaporization, with a discontinuous density change, can occur in this situation. As the liquid droplet approaches the critical temperature, surface tension forces are substantially reduced. Therefore, distortion of the droplet shape and secondary atomization become more likely. *See* CRITICAL PHENOMENA; SUPERCRITICAL FLUID.

William A. Sirignano

Bibliography. A. Frohn and N. Roth, *Dynamics of Droplets*, Springer-Verlag, New York, 2000; A. H. Lefebvre, *Atomization and Sprays*, 1989; W. A. Sirignano, *Fluid Dynamics and Transport of Droplets and Sprays*, 1999; Z. Zapryanov and S. Tabakova, *Dynamics of Bubbles, Drops and Rigid Particles*, vol. 50 of *Fluid Mechanics and Its Applications*, 1999.

Spread spectrum communication

A means of communicating by purposely spreading the spectrum (frequency extent or bandwidth) of the communication signal well beyond the required bandwidth of the data modulation signal. Spread spectrum signals are typically transmitted by electromagnetic waves in free space, with usage in both nonmilitary and military systems.

Motivation for using spread spectrum signals is based on the following:

1 Spread spectrum systems have the ability to reject hostile as well as unintentional jamming by interfering signals so that information can be communicated.

2 Spread spectrum signals have a low probability of being intercepted or detected since the power in the transmitted wave is "spread" over a large bandwidth or frequency extent.

3 Since these signals cannot be readily demodulated without knowing the code and its precise timing, a level of message privacy is obtained.

4 The wide bandwidth of the spread spectrum signals provides tolerance to multipath (reflected waves that typically take longer to arrive at the receiver than the direct desired signal so that the two can be distinguished).

5 A high degree of precision in ranging (distance measuring) can be obtained by using one type of spread spectrum signaling called direct sequence, with applications to navigation.

6 Multiple access, or the ability to send many independent signals over the same frequency band, is possible in spread spectrum signaling. *See* BAND-WIDTH REQUIREMENTS (COMMUNICATIONS); COMMU-NICATIONS SCRAMBLING; ELECTRICAL INTERFERENCE; ELECTRONIC WARFARE; RADIO-WAVE PROPAGATION.

There are four generic types of spread spectrum signals: direct sequence (DS) or pseudonoise (PN), frequency hopping (FH), linear frequency modulation (chirp), and time hopping (TH). The first two methods are much more commonly used today than the other two. **Chirp modulation.** Chirp modulation is a spread spectrum method which was developed for radar use in the mid-1940s. The basic idea is to transmit a long rectangular pulse whose carrier frequency is linearly increased. The frequency-modulated signal returned from the target passes through a filter in the receiver at a velocity of propagation proportional to frequency. The result is a pulse that is much shorter in time duration than the transmitted pulse with a larger peak power. Unchirped pulses such as interference or jamming pulses do not "compress" at the receiver, so that this method yields a processing gain or advantage for the chirped signal. *See* RADAR.

Direct sequence systems. Direct sequence modulation is characterized by phase-modulating a sine wave by an unending string of pseudonoise code chips (symbols of much smaller time duration than a bit). This unending string is typically based on a pseudonoise code that generates an apparently random sequence of code chips that repeats only after the pseudonoise code period. If T_b is the duration of a bit expressed in seconds, then typically the duration of a pseudonoise code chip, T_c , would be much smaller than T_b , so that many chips would occur during one bit (typically from hundreds to millions). An important parameter in direct sequence systems is the processing gain (PG), which is given by the ratio T_b/T_c . Larger values of processing gain give greater immunity to interference.

Transmitter. In many systems both the data and the pseudonoise code phase-modulate the carrier by 0° or 180° , which is called binary phase-shift keying (BPSK). In a basic binary phase-shift keyed, pseudonoise-encoded spread spectrum system (**Fig. 1***a*), digital data representing the information to be transmitted are binary phase-shift keyed onto the carrier. Then the pseudonoise code



Fig. 1. Equipment for a basic direct-sequence (pseudonoise-encoded), binary phase-shift keyed, spread spectrum system. (a) Transmitter. (b) Receiver.



Fig. 2. Example of how a narrow-band jammer is despread in the front end of a receiver. (a) System model. (b) Transmitted signal and jammer. (c) Despread signal and jammer.

generator also binary phase-shift keys the carrier, and the composite signal is transmitted. *See* ELECTRI-CAL COMMUNICATIONS; MODULATION; PHASE MODU-LATION.

Receiver. For the receiver to function, it must first align the receiver pseudonoise code generator timing with the received pseudonoise code signal timing (Fig. 1b). This is accomplished in two distinct steps, code acquisition and code tracking. Code acquisition is process of searching in time by correlating the locally generated code against the received signal plus noise. The search continues until the correlation process produces an output above a predetermined threshold. This occurs either when the local and received codes are in time alignment or there is a false alarm due to noise. In either case the code loop is engaged and given a fixed length of time to track the received code. When true tracking occurs, the received signal is said to be despread, leaving only the data to be demodulated.

The code tracking circuit, which is called a delay lock loop, tracks the code (keeps the local receiver code aligned with the received code) once acquisition has been achieved. When the received code is being tracked, it can be multiplied by the locally generated pseudonoise code which then removes the code on the received signal.

At this point a BPSK-data-modulated carrier signal remains to be demodulated. In order to do this, a carrier tracking loop must be utilized, such as a Costas loop or a squaring loop. Their function is to track the carrier and thereby allow phase-coherent demodulation of the data, that is, multiplication by a signal exactly in phase with the received carrier, which leaves only the baseband data to be detected.

A bit synchronizer then tracks (provides bit timing for) the data bit sequence, and a data demodulator utilizes a matched filter to detect the data. The matched filter for binary phase-shift keyed data is a T_b -second integration over each bit time. The output of the integrator produces either a positive or a negative output, which is then interpreted as a binary 1 or a 0. This bit sequence of 0's and 1's determines the message being sent. The data bits transmitted are either 1 or -1, which maps to either a binary 0 or a 1. If the signal-to-noise ratio is sufficiently large, very few errors will occur in the bit detection process.

The above discussion has obviously been oversimplified. For example, the reconstructed pseudonoise code is never perfectly aligned with the received code due to noise and filtering effects, and the locally generated carrier likewise is not exactly in phase with the received carrier for similar reasons. Degradations therefore occur to the system when compared to an idealized system. Other modulations besides BPSK are used.

Protection against jamming. It is instructive to see how the effects of a narrow-band jammer (narrow-band compared to the spreading-code bandwidth) are diminished at the receiver due to the despreading action in a direct sequence system. A jamming signal, which interferes with the reception of the desired signal, may be intentional or unintentional. Figure 2 illustrates a simplified direct-sequence transmitter and receiver including a signal with a narrow-band jammer. For this example it is assumed that the data and direct spread spectrum signal are BPSK-modulated onto the carrier and transmitted via the transmitter antenna. A jammer transmits its narrow-band jammer in the direction of the receiver antenna. At the receiver the sum of the signal and the jammer are captured at the receiver antenna and despread by a synchronous despreading signal, PN(t). The process of synchronously despreading (mathematically multiplying) the received signal returns the signal to the prespread bandwidth. However, the process of despreading the jammer, since it is unrelated to the signal, spreads the jammer to the bandwidth of the original spread signal, and only a small amount of the jammer power enters the receiver in the data bandwidth. Thus the processing gain of the direct-sequence system (PG = T_b/T_c) greatly reduces the effect of a narrow-band jammer on its performance. Furthermore, it follows that the spectral height of the spread signal is much lower than if it were not spread; hence, it is much harder to detect by an unauthorized user in the constant presence of the receiver noise. See JAMMING.

Attributes. The six attributes of spread spectrum systems that were discussed above apply in this direct-sequence, pseudonoise spread spectrum system. Suppose an unmodulated tone at the carrier frequency is present at the receiver along with the spread signal. Then in the pseudonoise code despreading process, the undesired tone is spread or multiplied by the local pseudonoise code, thereby spreading its spectrum to a bandwidth of $2/T_c$. However, the received signal is collapsed to a bandwidth of $2/T_b$ ($T_b \gg T_c$) so that the data detector "sees" only a small fraction (T_c/T_b) of the original tone energy. Low probability of intercept is achieved by virtue of the fact that the spread signal is very wide band and has low spectral height, and therefore

cannot be easily detected in a noise background. Message privacy is enhanced if the actual pseudonoise code sequence used by the transmitter and receiver is cryptographically derived and is unknown to unwanted listeners. Multipath tolerance is achieved by searching first with minimum delay and then expanding the search to larger delays. The desired signal will have less delay than the multipath signal and will be acquired first. Additional processing can be used to greatly minimize the multipath problems. If a set of essentially orthogonal codes, such as Gold codes (a special type of pseudonoise code), are used, many users can occupy the same bandwidth and only transmitter-receiver pairs with the same code can communicate. Since the pseudonoise chips are very short in time, ranging can be made accurate to a small fraction (5%) of a chip time. The higher the chip rate, the greater the range accuracy. See CRYP-TOGRAPHY.

Time hopping. In time hopping, a time interval Tis selected to be large compared to the data bit time, T_b , and is divided into, say, K slots. Then, during one slot time (T/K seconds), T/T_b bits are all burst-transmitted. The slot time that is transmitted is selected pseudorandomly so that only the transmitter and the receiver know the slot time code. In time hopping, the power increases by K times compared to what a continuous signal would require, but the time-hopping signal is on for only 1/K of the time, resulting in the same average power. The processing gain is K for BPSK modulation in this case, since bits are transmitted at K times the rate before time hopping. This scheme has high bandwidth efficiency, but high peak power and critical timing are required.

Frequency-hopped systems. In a direct-sequence system, the phase of the carrier changes pseudorandomly with the pseudonoise code. In a frequencyhopping system, the frequency of the carrier changes according to a pseudonoise code with a consecutive group of pseudonoise code chips defining a particular frequency. Typically either multiple frequencyshift keying (MFSK) or differential phase-shift keying (DPSK) is used as the data modulation. Multiple frequency-shift keying is a modulation scheme in which one of a number of tones (2, 4, 8, and so forth)is transmitted at a given time according to a group of consecutive data bits (*n* bits produce 2^{+n} tones). At each hop frequency one of the 2^{+n} tones is selected according to the *n* bits, and one of 2^{+n} corresponding frequencies, centered about the hop frequency, is transmitted. In conjunction with frequency hopping, multiple frequency-shift keying would imply, at each instant of time, a given carrier frequency that depends on the hop-pseudonoise code sequence and the consecutive group of the most recent *n* data bits. Differential phase-shift keying is similar to phase-shift keying except that only the differences of the phases (not the actual phases) are encoded and noncoherent techniques (not requiring a carrier loop) can be employed at the receiver.

Frequency-hopped systems are classified as fast or slow. Slow hopping occurs if one or more data symbols are transmitted in the time interval of one hop (T_b) . Fast hopping occurs if two or more hops occur per data symbol. Although fast hopping is more complex to implement, it offers greater jammer protection. The processing gain for a frequency-hopped system is defined by PG = W/R_b , where W is the hopped bandwidth and R_b is the data rate.

Figure 3 illustrates a simple frequency-hopped system with binary frequency-shift keying (BFSK) data modulation. In this example a data bit 1 is denoted by a positive offset of Δf hertz, and a 0 by a negative offset of Δf hertz from the selected hop frequency. In Fig. 3 a 0 is sent in the first hop period since it occurs in the lower portion of the frequency bin, and in the second hop period a 1 is sent (at a different hop frequency). This process continues indefinitely. A well-designed hopping pattern will cover all parts of the hopping bandwidth with the same frequency of occurrence.

A device called a frequency synthesizer achieves the actual frequency selection. For example, a 12bit segment of the pseudonoise code may correspond to one of 212 different frequencies, so that one of approximately 4000 (212) frequencies is selected each hop time. Frequency synthesizers are used in both the transmitter and the receiver (Fig. 4). The transmitter modulates the data by typically using either multiple frequency-shift keying or differential phase-shift keying modulation, which in turn is frequency-hopped by the frequency synthesizer. At the receiver, an acquisition process is utilized to synchronize the receiver frequency synthesizer with the received hopping signal, and then a tracking system maintains synchronism. Finally a bit synchronizer provides timing for the data demodulator which demodulates the original, transmitted data bits.

Hybrid systems. Besides the four basic systems discussed above, hybrid schemes combining two or more of these systems have been used. The most



Fig. 3. Example of a frequency-hopped signal pattern.



Fig. 4. Frequency-hopping equipment for (a) transmitter and (b) receiver.

common hybrid system is frequency hopping in conjunction with direct sequence; that is, each frequency is "spread" by a direct-sequence code.

Code division multiple access. An important aspect of spread spectrum communications is multiple access. Code-division multiple access (CDMA) is a method by which spread spectrum signals are utilized to allow the use of multiple signals over the same frequency band. The two common types are direct-sequence CDMA (DS/CDMA) systems and frequency-hopped CDMA (FH/CDMA) systems. *See* MULTIPLEXING AND MULTIPLE ACCESS.

DS/CDMA. This method is based on modulating the carrier phase via the multiple access code and the data modulation. Multiple codes can be used over the same channel because the different codes are nearly uncorrelated. That is to say, when any distinct pair of codes are multiplied together and the product is compared chip by chip over one data symbol, the code chips (+1's or -1's) have about the same number of "plus ones" as "minus ones." Thus two distinct codes are uncorrelated. However, when the same code is correlated against itself, its product is all ones, and thus is completely correlated is required for each code pair used in the multiple access system in order to minimize crosstalk or mutual interference

effects. When multiple CDMA signals exist on the same channel, CDMA noise (background noise) occurs. *See* CROSSTALK.

FH/CDMA. This method uses spread spectrum codes to control the hop frequencies at various values across the spread bandwidth in such a way that, most of the time, they do not collide. Typically the data modulation is multiple frequency-shift keying, as discussed above. In many applications it is not possible to construct codes that will not collide (or will have zero correlation in the DS/CDMA case). These occasional collisions cause a degradation in the bit error rate performance. However, coding and interleaving/deinterleaving mitigate the effects of the collisions. The more channels that are used on the same band, the greater the interference level, which is sometimes viewed as background noise. *See* ELEC-TRICAL NOISE.

Personal communication systems. In 1983 an analog cellular radio system, called the Advanced Mobile Phone System (AMPS), was introduced in the United States. Other analog systems were later introduced in Europe and Asia. In 1993 a CDMA spread spectrum personal cellular service was introduced in North and South America, Asia, and elsewhere. This standard, called TIA IS-95, is a second-generation (2G) standard. Each mobile user is assigned one member of a set of 64 orthogonal functions know as Walsh functions. The Walsh functions are generated as rows of special square matrices called Hadamard matrices. On the forward channel (from the base station to the mobile user) the Walsh functions are used for spreading, and on the reverse channel (user to base station) they are used as orthogonal functions for the modulation. Third-generation (3G) CDMA systems are in the beginning stages, and offer more features and greater data rates than second-generation systems. See MOBILE COMMUNICATIONS.

Other applications. Although the early evolution of spread spectrum systems was motivated primarily by military interests, nonmilitary applications have enjoyed considerable development. One important example that has both military and nonmilitary users is the Global Positioning Systems (GPS), which is a direct-sequence, CDMA, spread spectrum system for transmitting the satellite ranging codes. *See* SATEL-LITE NAVIGATION SYSTEMS.

The space shuttle utilizes a direct-sequence spread spectrum communication system on its forward link. It relays data through the geostationary *Tracking and Data Relay Satellites (TDRS)* by using either S-band or Ku-band spread spectrum links. Another example in the military arena is the Milstar system, which utilizes frequency-hopping spread spectrum communication over a very large bandwidth to achieve considerable immunity from unfriendly jamming signals. *See* MILITARY SATELLITES; SPACE COM-MUNICATIONS; SPACE SHUTTLE.

Globalstar is a commercial satellite system that utilizes a CDMA signal structure along with a bent-pipe transponder to provide communications much like a cell phone, except that Globalstar is satellite-based. Its signal structure is similar to IS-95. Currently the handsets are considerably more expensive than cell phones. However, these phones have the advantage that they can be reached beyond the range of cell phones. *See* COMMUNICATIONS SATELLITE.

The main reception source of news stories for radio and newspapers is small spread spectrum ground stations. Another application of codedivision multiple access is transmission directly by satellite from a bank's automatic teller machine to that bank's computer facility. A home security system using spread spectrum techniques imposed on the ac power line has been used in Japan. Many cordless telephones utilize direct-sequence spread spectrum techniques. Jack K. Holmes

Bibliography. S. Glisic and B. Vucetic, *Spread Spectrum CDMA Systems for Wireless Communications*, Artech House Publishers, 1997; J. K. Holmes, *Coherent Spread Spectrum Systems*, Krieger Press, 1990; R. L. Peterson et al., *Introduction to Spread Spectrum Communications*, Prentice Hall, 1995; M. K. Simon et al., *Spread Spectrum Communications Handbook*, rev. ed., 1994.

Spring (hydrology)

A place where groundwater discharges upon the land surface because the natural flow of ground water to the place exceeds the flow from it. Springs are ephemeral, discharging intermittently, or permanent, discharging constantly.

Springs are usually at mean annual air temperatures. The less the discharge, the more the temperature reflects seasonal temperatures. Spring water usually originates as rain or snow. Meteoric water compositions for the deuterium (D) and oxygen-18 (¹⁸O) are given by the empirical equation

$$\delta D = 8\delta^{18}O + 10$$

where δ is the isotope composition in parts per thousand referred to standard mean ocean water.

Standard mean ocean water is defined as zero in both O^{18} and D. Meteoric water from evaporation of seawater is depleted in D and ¹⁸O for kinetic reasons and hence the values of δD and $\delta^{18}O$ are negative and increasingly so the farther from the Equator.

Hot springs. Hot-spring water may differ in composition from meteoric water through exchange between the water and rocks. Common minerals consist of component oxides. Oxygen of minerals has more ¹⁸O than meteoric water. Upon exchange, the water is enriched in ¹⁸O. Most minerals contain little deuterium, so that slight deuterium changes occur. Water in hot rock is buoyant relative to surrounding water and isues as hot springs. Temperatures of the deep water may be estimated. Estimates stem from mineral solubilities with temperature. If saturation with a mineral is assumed and the chemical composition of the water known, a temperature may be calculated. The solubility of quartz, SiO₂, is commonly used for estimating the deep temperatures of hot springs. Common minerals also contain silica, so that temperature estimates are ambiguous. Some hot-spring waters are acid from the oxidation of hydrogen sulfide to sulfate.

Mineral springs. Mineral spring waters have high concentrations of solutes. Mineral springs have long been used for therapy; the study of such use of springs is balneology. Pliny the Younger reported mineral springs in Belgium and Anatolia, which are still in use. Although mineral waters are still valued for therapy in much of the world, in North America such use has declined greatly since the early twentieth century.

Mineral springs have wide ranges in chemistry and temperatures, and hot mineral springs may be classified as hot springs as well as mineral springs. Most mineral springs are high either in sodium chloride or sodium bicarbonate (soda springs) or both; other compositions are found, such as a high percentage of calcium sulfate from the solution of gypsum. The water and the solutes may be of different origins. The stable isotope compositions of the water are the best guides to the sources of the water. For the mineral springs of Europe, the United States, and Asia Minor, deuterium and ¹⁸O compositions show that the waters are meteoric.

Carbon dioxide. Carbon dioxide is commonly found dissolved in or issuing through mineral water, and is identifiable by its ¹³C composition. Carbon of carbonate minerals deposited in oceans is near zero in ¹³C compared to an international standards, the Peedee belemnite. The carbon of the carbon dioxide in most of the mineral springs of Europe, the United States, and Asia Minor is close to zero in ¹³C isotope composition. It is inferred that the carbon dioxide results from the breakdown (metamorphism) of the carbonate minerals that are present in rocks of marine sedimentary origin.

Carbon dioxide from mineral springs of volcanic islands of the Atlantic Ocean seems to come from the source of the volcanic rocks, the Earth's mantle, because of the similarity of the isotope compositions of the carbon dioxide in the springs and in carbon dioxide bubbles in volcanic rocks erupted on the deep ocean floor in the -4 to -8 per thousand range. Carbon dioxide from great depth also discharges through springs in the Sierra Nevada, California, and the Alaska peninsula. In mineral springs from the Rocky Mountains, carbon dioxide comes from both very deep sources and the shallower chemical reactions of marine rocks.

A third source of carbon dioxide is organic material. Springs in the Coast Ranges of California, along the Copper River of Alaska, and in New Zealand yield mixtures of carbon dioxide and methane. The carbon dioxide is depleted in carbon-13 and probably comes from the breakdown of organic (woody) materials in the sedimentary rocks.

Salts. Other solutes in the mineral springs are inferred to come from the rocks, but isotope data do not show origins. Salt deposits are sources for solutes in some instances. There are many mineral springs for which there is no possible salt deposit source for the solutes. Chemical compositions may be used to infer the sources of the solutes. Chloride-rich waters

in marine sedimentary rocks also contain bromide, iodide, and boron. During metamorphism of the rock the waters may be retained as fluid inclusions or as films along grain boundaries. Flushing of the retained brines may result in the saline springs occasionally found in metamorphic terraines.

Chemical equilibrium. The chemical compositions of spring waters are seldom in chemical equilibrium with the air. Ground waters whose recharge is through grasslands may contain a thousand times as much CO2 as would be in equilibrium with air, and those whose recharge is through forests may contain a hundred times as much as would be in equilibrium with air. The high CO₂ content, along with other solutes, makes some spring waters quite nutrient to aquatic plants. If the CO2-rich ground water has dissolved calcite, loss of CO₂ in a spring (chiefly by photosynthesis) may cause the calcite to precipitate to make travertime. Ground water flowing through rocks or sediments containing organic material or reduced minerals such as sulfides loses its dissolved oxygen by oxidizing the organic matter and the sulfides. When the dissolved oxygen is lost, iron enters the groundwater as ferrous ion, Fe^{2+} . The Fe^{2+} is not chemically compatible with air, and springs containing Fe^{2+} lose it as $Fe(OH)_3$ upon oxidation. Thin films of Fe(OH)₃ on the water surface are iridescent and resemble oil films. Unlike oil films, the Fe(OH)₃ films are brittle and if ruptured do not reform. Accumulations of the iron precipitates may lead to bog iron ores. Sulfate in ground water may be reduced in the presence of organic matter to H₂S, giving some springs the odor of rotten eggs. See GEYSER; GROUND-WATER HYDROLOGY. Ivan Barnes

Bibliography. D. K. Todd and L. W. Mays, *Ground-water Hydrology*, 3d ed., 2004; P. Younger, *Ground-water in the Environment: An Introduction*, 2006.

Spring (machines)

A machine element for storing energy as a function of displacement. The flywheel, in contrast, is a means for storing energy as a function of angular velocity. Force applied to a spring member causes it to deflect through a certain displacement thus absorbing energy. *See* FLYWHEEL.

A spring may have any shape and may be made from any elastic material. Even fluids can behave as compression springs and do so in fluid pressure systems. Most springs take on specific and familiar shapes such as helix, flat, or leaf springs. All mechanical elements behave to some extent as springs because of the elastic properties of engineering materials.

Uses. Energy may be stored in a spring for many uses: to be released later, to be absorbed at the instant the energy first appears, and so on.

Motive power. One of the early and still the most frequent uses of springs is to supply motive power in a mechanism. Common examples are clock and watch springs, toy motors, and valve springs in auto engines. In these, energy is supplied to, and stored

in, the spring by applying a force through a suitable mechanism to deflect or deform the spring. The energy is released from the spring by allowing it to push (as in the valve) or twist (as in a clock) a mechanism through a required displacement.

Return motion. A special case of the spring as a source of motive power is its use for returning displaced mechanisms to their original positions, as in the doorclosing device, the spring on the cam follower for an open cam, and the spring as a counterbalance. To a certain extent the springs in vehicles of transportation are in this category. They are designed to keep the car at a certain level with respect to the road or rails, returning the vehicle to this position if displaced by applied forces.

Shock absorbers. Frequently a spring in the form of a block of very elastic material such as rubber absorbs shock in a mechanism. For example, the four legs of a punch press rest on four blocks of rubber. The rubber pads prevent the die-closing inertia forces of the press from transferring down through the legs to the floor with impact or hammer blow proportions. With the rubber pads under the press legs, the force on the floor builds up relatively slowly and no shock is evident. As the acceleration of the die block goes to zero, the inertia force goes to zero and the rubber pad springs, which were deflected by the press blow, are relaxed and ready for the next stroke of the press die. The whole press moves up and down relative to the floor, but by proper selection of the rubber pad the elastic constant is such that this motion is small. See SHOCK ABSORBER.

Vibration control. Springs serve an important function in vibration control by supplying the necessary flexibility in the support of the vibrating mechanism and the required opposing forces as a result of their deflection. In controlling vibration, the body or mass of the mechanism must be freely supported so as to generate forces opposing the vibrating forces. These opposing forces tend to bring the sum of the forces on the vibrating body to zero. Vibration-absorbing mounts are available in a wide range of sizes and spring constants to meet most requirements. These do not prevent vibration, which is a function of speed and the balance of the mechanism, but rather they minimize its effect on the machine frame or the mounting. *See* SHOCK ISOLATION.

Force measurement. Springs have long been used in simple weighing devices. Accurate weighing is usually associated with dead-weight devices or balances, but modern spring scales have received wide acceptance and certification for commercial use. Extremely accurate springs for heavy loads are used to calibrate testing machines, and they are employed in scales over crane hooks for weighing material as it is hoisted. Carefully calibrated springs are used in instruments such as electric meters and pressure gages. *See* BALANCE.

Retaining rings. A relatively modern machine part in which the spring function is used as a holding means is the retaining or snap ring. This device is a split ring of square, rectangular, or special cross section. It fits in a groove on a cylindrical surface or in a bore



Fig. 1. Flat or leaf spring responds to perpendicular forces in either direction.

and stays in place by spring force. A retaining ring prevents or restrains relative axial motion between a shaft or bore and the components on the shaft or in the bore. *See* SNAP RING.

Types. Springs may be classified into six major types according to their shape. These are flat or leaf, helical, spiral, torsion bar, disk, and constant force springs.

Flat or leaf spring. A leaf spring is a beam of cantilever design with a deliberately large deflection under a load (**Fig. 1**). One end of a leaf spring is usually firmly anchored to the frame of the machine and the other end is linked to the moving machine elements by a two-force (pin-ended) link. Force may be tension or compression with no modification of the design. This push-pull feature is the great advantage of a leaf spring, plus the fact that a relatively large amount of energy can be stored in a small space.

Helical spring. The helical spring consists essentially of a bar or wire of uniform cross section wound into a helix. The last turn or two at each end of the spring is modified to a plane surface perpendicular to the helix axis, and force can then be applied to put the helix in compression. The ends of the spring helix may be modified into hooks or eyes so that a force may be applied in tension. In general it is necessary to design helical springs so that force may be transmitted either in tension or in compression but not both ways for the same spring (**Fig. 2**). Where reversing forces occur in a spring mechanism, it is better to use a bar or leaf spring and in some cases a disk spring.

Spiral spring. In a spiral spring, the spring bar or wire is wound in an Archimedes spiral in a plane. Each end of the spiral is fastened to the force-applying link in the mechanism.

A spiral spring is unique in that it may be deflected in one of two ways or a combination of both of them (**Fig. 3**). If the ends of the spiral are deflected by forces perpendicular to the spiral plane, the spiral is distorted into a conical helix. For better stability and ease of applying the forces, the spring is often made as a conical helix to start with and is deflected into a plane spiral.

A spiral spring may also have the forces act tangent



Fig. 2. Helical spring. (a) Wound tight to extend under axial tension. (b) Wound loose to contract under axial compression.



Fig. 3. Spiral spring is unique in responding to torsional or translational forces.

to the spiral as in a clock spring. The spiral is wound quite open and the tangential force in tension on the spiral tends to close the gaps between successive turns. The spring behaves like a beam, bending to a shorter radius of curvature and thus storing energy.

Torsion bar. A torsion bar spring consists essentially of a shaft or bar of uniform section. It stores energy when one end is rotated relative to the other. It is used in the spring system of the chassis of modern motor cars. *See* TORSION BAR.

Disk spring. Where large forces are present and space is at a premium, the disk spring may be used, although it is usually expensive to design and build. It consists essentially of a disk or washer supported at the outer periphery by one force (distributed by a suitable chuck or holder) and an opposing force on the center or hub of the disk (**Fig. 4**). For greater deflections several disks may be stacked with the forces transmitted from one to the next by inner and outer collars.

Constant-force spring. Many mechanisms require that a constant force be applied regardless of displacement. The counterbalancing of vertically moving masses against the force of gravity is a typical example. The Neg'ator spring of Hunter Spring Co. provides such a constant force; it uses a tight coil of flat steel spring stock. When the outer free end is extended and the coil allowed to rotate on its shaft or pintle, the spring presents a constant restoring force.

Design. Each type spring has its special features and design refinements. Common to all forms of springs are the basic properties of elastic materials. Within the elastic range of a material, the ratio of applied force to resulting deflection is constant. Spring systems can be designed to have a variable ratio. The ratio is the spring rate or scale and has the dimension of force per unit length. *See* HOOKE'S LAW.

In a helical spring the elastic action stresses the wire in torsion. The following variables are subject



Fig. 4. Disk spring. (a) Single tapered disk; (b) multiple disk; (c) Belleville spring; (d) slotted curved washer; (e) finger washer; (f) wave spring washer.



Fig. 5. Correction factor for curvature and shear. (After A. M. Wahl, Mechanical Springs, Penton, 1944)

to the designer's action and decision:

free length	L	spring rate	k
mean diameter	D	modules of elasticity in	
wire diameter	d	torsion	G
allowable stress	S	number of active turns	N
applied load	p	deflection at load P	F

These variables are related by Eq. (1), which

$$S = \frac{8PD}{\pi d^3} \tag{1}$$

neglects the effect of coil curvature on the stress, and by Eq. (2). For springs subject to frequent cycling, as

$$F = \frac{8PD^3N}{Gd^4} \tag{2}$$

a valve spring, the correction factor for wire curvature must be included. The generally used correction is the factor proposed by A. M. Wahl. Introducing this factor changes Eq. (1) to the form of Eq. (3), in

$$S_{\max} = \frac{8PD}{\pi d^3} \left(\frac{4C - 1}{4C - 4} + \frac{0.615}{C} \right)$$
(3)

which C = D/d. The portion in brackets is Wahl's factor *K* whence maximum stress is *K* times the torsional stress (**Fig. 5**).

Various combinations of spring dimensions for the variables listed above will produce an acceptable spring. Usually the design is chiefly limited by allowable stress of materials and space.

Where compression and tension springs are cycled at very high frequencies, surging may cause high local stress and result in early fatigue failure. Surging is the inability of all parts of the spring to deflect at the same rate due to the inherent inertia in the coils. This phenomenon is closely associated with the natural frequency of the spring; springs should be designed so that their natural frequency and cyclic rate are as far apart as practical.

L. Sigfred Linderoth, Jr.

Bibliography. Society of Automotive Engineers, Spring Design Manual, 2d ed., 1995; Spring Manufacturers Institute, Handbook of Spring Design, rev. ed., 2002; A. M. Wahl, *Mechanical Springs*, 2d ed., reprinted 1991.

Spruce

Evergreen tree belonging to the genus *Picea* of the pine family. The needles are single, usually four-sided, and borne on little peglike projections; the cones are pendulous. Resin ducts in the wood may be seen with a magnifying lens, but they are fewer than in *Pinus. See* EVERGREEN PLANTS; PINALES; PINE.

Eastern species. The white spruce (*P. glauca*), ranging from northern New England to the Lake States and Montana and northward into Alaska, is distinguished by the somewhat bluish cast of its needles, small cylindrical cones (Fig. 1c), and gray or pale-brown twigs without pubescence (hairs). Of erect pyramidal habit, it usually attains a height of 60-70 ft (18-21 m) with trunk diameters of 1-2 ft (0.3-0.6 m), but occasionally in British Columbia and Alberta the tree reaches heights of 80-140 ft (24-43 m) with diameters up to 4 ft (1.2 m). When bruised, the leaves of some individuals emit a disagreeable odor, hence the names cat or skunk spruce. The pale, straw-colored wood is soft and straight-grained without different coloration of the heartwood. The wood is used for paper pulp, construction, and interior finish. The evenly grained wood is used in the manufacture of musical instruments

Red spruce (*P. rubens*) is a similar tree but with greener foliage; smaller, more oval cones (Fig. 1*b*); and more or less pubescent twigs. Occurring naturally with white spruce in the northeastern United States and adjacent Canada, red spruce extends southward along the Appalachians into North Carolina. As in white spruce, the greatest use of the wood is for paper pulp.

Black spruce (*P. mariana*) ranges from northern New England and Newfoundland to Alaska. However, it occurs sparingly in the Appalachians to West Virginia. The cones are smaller than in the white and red species and are egg-shaped or nearly spher-



Fig. 1. Cones of eastern and western spruce. (a) Black spruce (*Picea mariana*). (b) Red spruce (*P. rubens*). (c) White spruce (*P. glauca*). (d) Colorado blue spruce (*P. pungens*). (e) Norway spruce (*P. abies*). (*Brooklyn Botanic Garden*)



Fig. 2. Blue spruce (*Picea pungens*) in Bryce Canyon National Park, Utah. (*National Park Service*)

ical $(1-1^{1}/_{2}$ in. or 2.5-3.8 cm long) and persistent (Fig. 1*a*). The twigs are pubescent. Known also as bog or swamp spruce in the southern part of its range, black spruce appears at high elevations in the north. The long-fibered wood is ideal for paper pulp and is also used for paddles, oars, construction, and shipbuilding. It is short-lived as an ornamental tree.

Western species. Blue spruce (*P. pungens*) (**Fig. 2**), also known as Colorado blue spruce, is probably the best known of the western species (Fig. 1*d*) because of its wide use as an ornamental tree. The twigs are glabrous (without pubescence). The tree is popular as an ornamental in northern Europe, as well as throughout the United States. Individuals vary in the depth of blue coloration, the most pronounced being that of the Koster variety. Most of the ornamental specimens come from grafts made on Norway spruce.

Engelmann spruce (*P. engelmanni*) has needles usually of a deep blue-green color, sometimes much like those of the blue spruce, but the young twigs are slightly hairy. The cones, although cylindrical, are smaller than in blue spruce, being about $1^{1}/_{2^{-}}$ 3 in. (3.8–7.6 cm) long. This species is also a Rocky Mountain tree like the blue spruce, but it is more widely distributed from British Columbia to Arizona and also in the mountains of Oregon and Washington. It is an important timber species of the Rocky Mountain region.

Sitka spruce (*P. sitchensis*) is the largest spruce in the Northern Hemisphere. The leaves have a pungent odor, are considerably flattened, and stand out from the twig in all directions. Ranging from Alaska to northern California, it occupies a coastal strip about 40–50 mi (64–80 km) wide. Mature trees are 180–200 ft (55–61 m) tall and $3^{1}/_{2}$ –6 ft (1.1–1.8 m) in diameter, but in some instances larger dimensions have been recorded. Timber volume in the United States is nearly all in the states of Washington and Oregon. The wood is used for furniture, doors, blinds, and paper pulp, and in piano manufacture of sounding boards.

The Norway spruce (*P. abies*), the common spruce of Europe, is much planted in the United States for timber, as well as for ornamental purposes. It can be recognized by the dark-green color of the leaves; glabrous, pendent, short branchlets; and cones 4–6 in. (10–15 cm) in length (Fig. 1*e*), usually near the top of the tree. An important source of paper pulp, it has often been planted for that purpose in the northern and eastern United States. It is also used for shelter belts and Christmas trees, although it is not as desirable for the latter purpose as the firs and Douglas-fir, since it sheds its needles quickly in a warm room. There are many cultivated varieties. *See* FOREST AND FORESTRY; TREE.

Arthur H. Graves; Kenneth P. Davis

Bibliography. A. Farjon, *Pinaceae: Drawings and Descriptions of the Genera Abies, Cedrus, Pseudolarix, Keteleeria, Nothotsuga, Tsuga, Cathaya, Pseudotsuga, Larix, and Picea*, Koeltz Scientific Books, Königstein, 1990; R. J. Taylor, Sections on Picea and Tsuga, in *Flora of North America: North of Mexico*, vol. 2, ed. by Flora of North America Editorial Committee, Oxford University Press, 1993.

Sputtering

The ejection of material from a solid or liquid surface following the impact of energetic ions, atoms, or molecules. Sputtering was first noted in 1853 by W. R. Grove and in 1854 by M. Faraday as the occurrence of a metallic coating on the glass walls of a discharge tube. E. Goldstein in 1902 showed that the sputtering effect was caused by positive ions of the discharge striking the cathode and ejecting cathode material. Sputtering now is the basis of a large variety of methods for the synthesis and analysis of materials.

Classification. Sputtering can be classified according to the mode of energy loss of the incident (primary) particle. Nuclear stopping involves billiard ball-like atomic collisions in which a significant momentum transfer occurs; it dominates for incident ion energies below about 1–2 keV per nucleon. Electronic stopping involves collisions in which little momentum is transferred, but significant electronic excitation is caused in the target; it dominates for energies above about 10 keV per nucleon.

Sputtering has also been classified into physical and chemical sputtering. Physical sputtering involves a transfer of kinetic energy from the incident particle to the surface atoms leading to ejection, while chemical sputtering occurs when the incident species react chemically with the target surface leading to the formation of a volatile reaction product which evaporates thermally from the surface.

The main models for physical sputtering are the thermal sputtering and the collision cascade models. The thermal sputtering model suggests that the ion impact instantaneously heats a small region of the surface to a temperature well above the boiling point of the material, so that a small amount of material can evaporate before the heat is dissipated to the surrounding solid. The collision cascade model suggests that the impinging ion collides with surface atoms, transferring kinetic energy to these atoms which then collide with their neighbors, and so on. Atoms at or near the surface which are set in motion out of the surface with a kinetic energy great enough to overcome the chemical binding to their neighbors are sputtered. There is little evidence that thermal sputtering is important for most materials; in contrast, the collision cascade model accounts for the vast majority of observed sputtering phenomena. In particular, sputtering of single crystals leads to ejection in certain preferred directions correlated with the crystal lattice, which clearly would not occur in a random disordered hot zone. See CRYSTAL STRUC-TURE.

Yields. In the nuclear stopping regime, sputtering yields, *Y*, can be expressed as in the equation

$$Y \propto \frac{S_n}{U}$$
 atom/ion

where S_n is the nuclear stopping coefficient (energy deposited in the solid per unit distance traversed) and U is the surface binding energy of the target material (assumed equal to the heat of sublimation per atom). Sputtering threshold energies-the incident ion energy at which sputtering is just detectableare on the order of 10 times the sublimation energy, or some tens of electronvolts, for most elements. Sputtering yields are a maximum for incident ion energies around 200-300 eV per nucleon, where the nuclear stopping coefficient is a maximum. For argon ions (atomic mass 40) this corresponds to energies around 10 keV, at which energy sputtering yields for most materials under argon ion bombardment are in the range 1-10 atoms sputtered per incident ion. For incident ion energies below about 10 keV, a major fraction (80-90%) of the sputtered atoms come from the outermost surface atom layer, most of the remainder from the second layer, while essentially all the sputtered atoms come from no deeper than about four atomic layers; this is the basis of the use of sputtering in surface analysis.

Sputtering yields in the electronic stopping regime cannot be expressed simply, as mechanisms are still obscure. Hard, "billiard-ball" collisions are rare at high energies, because the incident ion travels through the solid as a bare nucleus, or certainly with many of its electrons stripped away, and is thus extremely small. However, highly energetic particles can deposit up to several thousand electronvolts per nanometer into electronic excitation and ionization in the substrate. Mechanisms suggested for coupling of this electronic excitation to the nuclear motion necessary for sputtering include Coulomb repulsion of ionized atoms from one another along the particle track, or excitation to antibonding (repulsive) electronic states.

Complex materials. Sputtering of complex materials-metal alloys, inorganic and organic compounds and polymers, and minerals-can produce complex results. The relative efficiencies with which different elemental species are ejected following ion impact can differ, giving rise to preferential sputtering. When preferential sputtering occurs, the species sputtered with the lower efficiency accumulates to a higher concentration at the surface. Subsurface collisions of the incident ion, which penetrates several nanometers into the solid at kiloelectronvolt energies, cause atomic motion leading to ion beam-induced atomic mixing of surface and subsurface layers over the ion penetration depth. Chemical bonds can be broken, and sometimes new bonds can be formed. Sputtering of solids which have multiple phases, or which are polycrystalline, leads to the development of surface roughness due to the differences in sputtering yields between different regions. See ION BEAM MIXING.

Semiconductor devices. Sputtering is widely used in the manufacture of semiconductor devices; sputter deposition is used to deposit thin films with a high degree of control by sputtering material from a target onto a substrate; sputter etching is used to remove unwanted films in a reversal of this process. Most removal and deposition systems generate the energetic sputtering ions in radio-frequency (rf) or direct current (dc) glow discharges in argon gas. Reactive ion etching is a chemical sputtering process in which chemically active sputtering species form volatile compounds with the target material leading to significantly higher etch rates and great selectivity. For example, fluorine-containing compounds etch silicon rapidly by forming volatile silicon tetrafluoride (SiF₄) but do not etch aluminum or other metals used to make electrical interconnections between devices on a semiconductor chip because the metal fluorides are involatile. Sputter etching and reactive ion etching have the useful advantage of being anisotropicthat is, they etch only in one direction (usually normal to the surface) so that very fine surface features can be delineated. See CRYSTAL GROWTH; GLOW DIS-CHARGE; INTEGRATED CIRCUITS.

Materials characterization. In materials characterization, sputtering is used to remove surface material controllably, allowing in-depth concentration profiles of chemical composition to be determined with a surface-sensitive sampling technique such as Auger electron spectroscopy, x-ray photoelectron spectroscopy, secondary ion mass spectrometry, or ion scattering spectrometry. Sputtering ejects not only atomic but also molecular species from surfaces, and a fraction of these ejected species are ionized as a result of the energetic collisions leading to ejection. Analysis of these secondary ions forms the basis of secondary ion mass spectrometry. Inorganic materials can be analyzed by identification of atomic secondary ions; organic materials can give characteristic molecular ions. Degradation of organic materials by the ion beam irradiation is reduced by dispersal in a liquid matrix or by operation at very low irradiation rates. Projectile energies both in the nuclear stopping and in electronic stopping regimes result in ejection of very large molecular ions up to 5000 atomic mass units for nuclear stopping and over 15,000 atomic mass units for electronic stopping, allowing analysis of very large biomolecules by mass spectrometry. *See* AUGER EFFECT; SURFACE PHYSICS. Peter Williams

Bibliography. R. Behrisch and W. Eckstein (eds.), Sputtering by Particle Bombardment, 2007; B. Chapman, Glow Discharge Processes: Sputtering and Plasma Etching, 1980; M. Ohring, The Materials Science of Thin Films, 2d ed., 2001; K. Wasa, M. Kitabatake, and H. Adachi, Thin Films Materials Technology: Sputtering of Compound Materials, 2004.

Squalene

A C_{30} triterpenoid hydrocarbon. Squalene is made up of six (*trans*-1,4)-isoprene units linked as two farnesy1 (head-to-tail) groups that are joined tail to tail in the center (**Fig. 1**).

Squalene can be isolated in large quantities from the liver oils of the shark (*Squalus* spp.) and other elasmobranch fishes, and is a relatively inexpensive compound. It undergoes oxidation reactions typical of a polyene. Complete hydrogenation of the liver oil gives the saturated hydrocarbon squalane, which is used in lotions and skin lubricants.

The major significance of squalene is its role as a central intermediate in the biogenesis of all steroids and triterpenoids. Thus, although squalene accumulates in a narrow phylogenetic group of primitive fish, it is a ubiquitous metabolite in all organisms that contain steroids.

The biosynthetic pathway begins with acetic acid units and leads via mevalonate and isopentenyl pyrophosphate to the sesquiterpene farnesol. Two farnesyl pyrophosphate units are then coupled tail to tail. This coupling involves the formation of presqualene followed by rearrangement and reduction of the cyclopropane ring (**Fig. 2***a*).

The transformation of squalene to polycyclic triterpenoids and steroids is initiated by epoxidation of one double bond followed by cyclization. The product depends on the conformation, or folding, of the



Fig. 1. Structure of squalene; the tail-to-tail joining is indicated by T.



Fig. 2. Biosynthetic pathways. (a) Formation of squalene. (b) Transformation of squalene to cholesterol.

chain. The pathway from squalene oxide to cholesterol is indicated in Fig. 2b. A concerted sequence of bond formation and methyl group rearrangements leads to the triterpenoid lanosterol, which is a major constituent of sheep wool fat. The final steps to cholesterol include oxidative loss of three methyl groups. Further enzymatic oxidations in specific tissues give rise to the several classes of steroid hormones. *See* STEROID; TERPENE; TRITERPENE.

James A. Moore

Bibliography. S. V. Bhat, B. A. Nagasampagi, and S. Minakshi, *Chemistry of Natural Products*, 2005; T. W. Graham Solomons and C. B. Fryhle, *Organic Chemistry*, 8th ed., 2003; R. K. Murray et al., *Harper's Biochemistry*, 25th ed., 1999.

Squaliformes

An order of sharks comprising six families, 24 genera, and 97 species, commonly known as dogfish sharks. The order differs from all other sharks by the follow-



Spiny dogfish (Squalus acanthias). (Photo courtesy of Alaska Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, U.S. Department of Commerce)

ing combination of characters: two dorsal fins, with or without spines; no anal fin; spiracle behind eye; and lower nictitating eye membrane absent. For the most part, squaliform sharks inhabit continental and insular shelves and slopes, and sea mounts. Some species live at great depths. So far as known, squaliforms are ovoviviparous and produce from one to 22 pups per litter, depending on the species. The dwarf lantern shark (*Etmopterus perryi*), at 17 cm (6.7 in.) in length, probably is the world's smallest shark; the largest squaliform is the Pacific sleeper shark (*Somniosus pacificus*), which is up to 7 m (23 ft) in length. The diet consists of bony fishes, cephalopods, crustaceans and other invertebrates, mammals, and other sharks. *See* ELASMOBRANCHII; SELACHII.

Squalidae (dogfish sharks). Squalidae has a spine (without grooves) preceeding both dorsal fins, teeth in the lower jaw only slightly larger than those in the upper jaw, usually a precaudal pit, and a lateral keel on each side of the caudal peduncle. The family comprises two genera. Cirrhigaleus (two species) is very distinctive in having a robust body and barbels in the form of large nasal flaps that reach the mouth. By comparison, Squalus, comprising eight species, has a more slender body and much smaller nasal flaps. The species S. acanthias (spiny dogfish) (see illustration) is probably is the world's most abundant shark and also the largest of the family, attaining a length of 160 cm (63 in.), although most are much smaller. Squalids are antitropical (found in both hemispheres but not in equatorial regions) in the eastern and western Atlantic and Pacific oceans, but apparently are absent in the Indian Ocean.

Centrophoridae (gulper sharks). The family Centrophoridae is distinguished from other squaliforms by having grooved spines in both dorsal fins, teeth in the lower jaw larger than those in the upper jaw, and precaudal pit and lateral keels absent from the caudal peduncle. Gulper sharks comprise two genera and 14 species, which are represented in the eastern Atlantic from Iceland to South Africa, and in the Indian and western Pacific oceans. They range in size from about 79 to 160 cm (31 to 63 in.).

Etmopteridae (lantern sharks). Etmopteridae is a family of small deep-water sharks comprising five genera and 41 species. The family is characterized by the following combination of characters: both dorsal fin spines grooved; spines usually quite large, with second spine larger and both spines free of fin; upper lobe of caudal fin with subterminal notch; and luminous organs usually present on body. Species range in adult total length from 17 to 84 cm (6.7 to 33 in.). Lantern sharks occur on continental and insular slopes from tropical to temperate zones of the Atlantic (northward to Iceland), Indian, and Pacific oceans. *Etmopterus*, the principal genus, comprises 31 species.

Somniosidae (sleeper sharks). The family Somniosidae comprises seven genera and 17 species of relatively small sharks, ranging in length from about 49 to 114 cm (19 to 45 in.). Sleeper sharks differ from other squaliform sharks by usually lacking dorsal fin spines; however, if present, spines are small and in both fins.

They also have a lateral ridge on each side between the pectoral and pelvic fins, as well as luminous organs in most species. Sleeper sharks occur mostly on continental and insular shelves and slopes (some are oceanic) of the Atlantic, Indian, and Pacific oceans, and on Arctic and sub-Antarctic shelves. One of the sleeper sharks, the Portuguese shark (*Centroscymnus coelolepis*), which occurs in the western North Atlantic and eastern Atlantic from Iceland to South Africa, ranges in depth from 270 to 3675 m (885 to 12,055 ft).

Oxynotidae (rough sharks). Oxynotidae, comprising one genus (Oxynotus) and five species, is the most bizarre family of squaliform sharks. The family is characterized as follows: trunk deep and compressed, triangular-shaped in cross section; prominent ventrolateral ridge between pectoral and pelvic fins; dorsal fins saill-ike, with the first extending forward over the pectoral fins and gill slits; both dorsal fins with spines embedded in the fin, with only tips exposed; skin very rough; and most species with luminous organs. The family occurs on continental and insular slopes of the western and eastern Atlantic (including the Mediterranean Sea) and the western Pacific at depths of 500 to 720 m (1640 to 2360 ft). The species range in length from 49 to 150 cm (19 to 59 in.).

Dalatiidae (kitefin sharks). Dalatiidae, comprising seven genera and 10 species (perhaps more), is characterized as follows: small size, ranging from 25 to 50 cm (10 to 20 in.); dorsal fins without spines except in genus Squaliotus, which may have small spines in the first dorsal; and luminous organs, mostly on belly. Kitefin sharks occur on continental and insular shelves and slopes of tropical and temperate zones of the Atlantic, Indian, and Pacific oceans. The pygmy shark (Euprotomicrus bispinatus), only 27 cm (10.6 in.) in maximum length, occurs at or near the surface at night and may make lengthy daily vertical migrations of 1500 m (4900 ft) or more, a round-trip of almost 2 mi. Species of the genus Isistius are called cookiecutter sharks (after the cookie-shaped bite wounds that they inflict on the bodies of prey). They have a short bulbous snout, and their large, fleshy suctorial mouth allows the shark to attach itself to prey, bony fishes and cetaceans, much in the manner employed by lampreys. When attached, the cookiecutter shark bites circular chunks of flesh from its prey. Its lower teeth, quite different from the uppers, are very large (in fact, in proportion to its size, the largest teeth of all sharks), erect, bladelike, and interlocking, making a formidable weapon for this relatively small shark.

Herbert Boschung

Bibliography. L. J. V. Compagno, FAO Species Catalogue, vol. 4: Sharks of the World: An Annotated and Illustrated Catalogue of Shark Species Known to Date, Part 1, Hexanchiformes to Lamniformes, FAO Fish. Synop. no. 125, pp. 212-249, 1984; L. J. V. Compagno, Checklist of Chondrichthyes, pp. 503-547 in W. C. Hamlett (ed.), Reproductive Biology and Phylogeny of Chondrichthyes: Sharks, Batoids, and Chimaeras, Science Publishers, Enfield, NH, 2005; J. S. Nelson, *Fishes of the World*, 4th ed., Wiley, New York, 2006.

Squall

A strong wind with sudden onset and more gradual decline, lasting for several minutes. Wind speeds in squalls commonly reach 30-60 mi/h (13-27 m/s), with a succession of brief gusts of 80-100 mi/h (36-45 m/s) in the more violent squalls. Squalls may be local in nature, as with isolated thunderstorms, or may occur over a wide area in the vicinity of a well-developed cyclone, where the squalls locally reinforce already strong winds. Because of their sudden violent onset, and the heavy rain, snow, or hail showers which often accompany them, squalls cause heavy damage to structures and crops and present severe hazards to transportation.

The most common type of squall is the thundersquall or rain squall associated with heavy convective clouds, frequently of the cumulonimbus type. Such a squall usually sets in shortly before onset of the thunderstorm rain, blowing outward from the storm and generally lasting for only a short time. It is formed when cold air, descending in the core of the thunderstorm rain area, reaches the Earth's surface and spreads out. Particularly in desert areas, the thunderstorm rain may largely or wholly evaporate before reaching the ground, and the squall may be dry, often associated with dust storms. *See* DUST STORM; SQUALL LINE; THUNDERSTORM.

Squalls of a different type result from cold air drainage down steep slopes. The force of the squall is derived from gravity and depends on the descending air which is colder and more dense than the air it replaces. So-called fall winds of this kind are common on mountainous coasts of high latitudes, where cold air forms on elevated plateaus and drains down fiords or deep valleys. Channeling of the air through narrow valleys increases their force. Squall winds (bora) of 100 mi/h (45 m/s) or more are observed in winter, when arctic air masses from Russia intermittently spill over the mountains of former Yugoslavia into the warm coastal regions of the Adriatic.

Violent squalls also characterize the warm foehn winds of the Alps and the similar chinook winds on the eastern slopes of the Rocky Mountains. Sometimes destructive, especially in winter, they are characteristically very gusty and variable over short distances. Strong surface wind speeds are a response to the distribution of pressure established in a standing breaking wave, somewhat analogous to standing waves in water flowing over and around submerged obstacles. Momentum transferred to the surface layers is dissipated in small-scale turbulence, an effective dissipative process affecting jet streams worldwide. *See* CHINOOK; JET STREAM; STANDING WAVE; WIND. Chester W. Newton; Edwin Kessler

Bibliography. C. D. Ahrens, *Meteorology Today: An Introduction to Weather, Climate, and the Environment*, 6th ed., 1999; C. D. Whiteman, *Mountain Me teorology: Fundamentals and Applications*, 2000.

Squall line

A line of thunderstorms, near whose advancing edge squalls occur along an extensive front. The thundery region, 12–30 mi (20–50 km) wide and a few hundred to 1250 mi (2000 km) long, moves at a typical speed of 15 m/s (30 knots) for 6–12 h or more and sweeps a broad area. In the United States, severe squall lines are most common in spring and early summer when northward incursions of maritime tropical air east of the Rockies interact with polar front cyclones. Ranking next to hurricanes in casualties and damage caused, squall lines also supply most of the beneficial rainfall in some regions. *See* SQUALL.

A squall line may appear as a continuous wall of cloud, with forerunning sheets of dense cirrus, but severe weather is concentrated in swaths traversed by the numerous active thunderstorms. Their passage is marked by strong gusty winds, usually veering at onset, rapid temperature drop, heavy rain, thunder and lightning, and often hail and tornadoes. Turbulent convective clouds, 6–9 mi (10–15 km) high, present a severe hazard to aircraft, but may be circumnavigated with use of radar. *See* AIRBORNE RADAR.

Formation requires an unstable air mass rich in water vapor in the lowest 0.6-2 mi (1-3 km), such that air rising from this layer, with release of heat of condensation, will become appreciably warmer than the surroundings at upper levels. Broad-scale flow patterns vary; **Fig. 1** typifies the most intense outbreaks. In low levels, warm moist air is carried northward from a source such as the Gulf of Mexico. This process, often combined with high-level cooling on approach of a cold upper trough, can rapidly generate an unstable air mass. *See* THUNDER-STORM.



Fig. 1. Successive locations of squall line moving eastward through unstable northern portion of tongue of moist air (dark area) in warm sector of a cyclone. Black arrows show general flow in low levels; colored arrows, axes of strongest wind at about 0.6 mi (1 km) above ground and at 6–7 mi (10–12 km).



Fig. 2. Section through squall-line-type thunderstorm. Precipitation is represented by black broken lines. 1 km = 0.6 mi.

The instability of this air mass can be released by a variety of mechanisms. In the region downstream from an upper-level trough, especially near the jet stream, there is broad-scale gentle ascent which, acting over a period of hours, may suffice; in other cases frontal lifting may set off the convection. Surface heating by insolation is a contributory mechanism; there is a preference for formation in midafternoon although some form at night. By combined thermodynamical and mechanical processes, they often persist while sweeping through a tongue of unstable air, as shown in Fig. 1. Squall lines forming in midafternoon over the Plains States often arrive over the midwestern United States at night. *See* STORM.

Figure 2 shows, in a vertical section, the simplified circulation normal to the squall line. Slanting of the drafts is a result of vertical wind shear in the storm environment. Partially conserving its horizontal momentum, rising air lags the foot of the updraft on the advancing side. In the downdraft, air entering from middle levels has high forward momentum and undercuts the low-level moist layer, continuously regenerating the updraft. Buoyancy due to release of condensation heat drives the updraft, in whose core vertical speeds of 66-132 mi/h (30-60 m/s) are common near the midlevel of the storm. Rain falling from the updraft partially evaporates into the downdraft branch, which enters the storm from middle levels where the air is dry, and the evaporatively chilled air sinks, to nourish an expanding layer of dense air in the lower 0.6-1.2 mi (1-2 km) that accounts for the region of higher pressure found beneath and behind squall lines. In a single squall-line thunderstorm about 12 mi (20 km) in diameter, $1-2 \times 10^7$ lb/s (5-10 kilotons/s) of water vapor may be condensed, half being reevaporated within the storm and the remainder reaching the ground as rain or hail. See MESOMETEOROLOGY. Chester W. Newton

Bibliography. C. D. Ahrens, *Meteorology Today: An Introduction to Weather, Climate, and the Environment*, 6th ed., 1999; E. Kessler (ed.), *Thunderstorm Morphology and Dynamics*, 2d ed., 1992; F. K. Lutgens, E. J. Tarbuck, and D. Tasa, *The Atmosphere: An Introduction to Meteorology*, 10th ed., 2006.

Squamata

The dominant order of living reptiles, composed of the lizards and snakes. The group first appeared in Jurassic times and today is found in all but the coldest regions. Various forms are adapted for arboreal, burrowing, or aquatic lives, but most squamates are fundamentally terrestrial. There are over 7700 extant species: 4800+ lizards and 2900+ snakes. *See* REPTILIA.

The order is readily distinguished from all known reptiles by its highly modified skull. The immediate ancestors of primitive lizards are members of the diapsid order Eosuchia, in which a pair of temporal foramina are present on both sides of the head. A similar pattern is found among living reptiles in the tuatara, order Sphenodonta (Rhynchocephalia), and the crocodylians, order Crocodylia. In the Squamata, however, in the case of the quadratojugal and jugal bones, which border the lower margin of the inferior temporal opening, the former has been lost completely and the latter greatly reduced. In consequence, there is only a single temporal opening present, and the quadrate has become enlarged and movable. Even this temporal opening is lost or reduced in many forms. No other reptiles show these modifications, which allow for great kinesis in the lower jaw since it articulates with the quadrate. In addition, the order is distinct from other living reptile groups because its members have no shells or secondary palates, and the males possess paired penes (hemipenes). Another distinguishing characteristic of the squamates is the synchronous shedding of the outer layer of the integument, called ecdysis. The rate of this shedding is not cyclic, but is based fundamentally on the rate of growth of the individual organism, which in turn is based on resource acquisition. See CROCODYLIA; EOSUCHIA; LEPIDOSAURIA; SPHEN-ODONATA.

Traditionally, the order Squamata has been divided into two major subgroups: the lizards, suborder Sauria (= Lacertilia), and the snakes, suborder Serpentes. The latter group is basically a series of limbless lizards, and it is certain that snakes are derived from some saurian ancestor. There are many different legless lizards, and it has been suggested that more than one line has evolved to produce those species currently grouped together as snakes. Besides this possibility, there is the additional problem of separating the two suborders. No single attribute will suffice for this purpose, but the features in the **table**, used in combination, will definitely allocate any questionable form to one or the other group.

Sauria

The majority of lizards are insectivorous, but a few species feed on plants. Others, notably the Varanidae and allies, feed on larger prey, including birds and mammals. The largest living lizard is the Komodo dragon (*Varanus komodoensis*) [**Fig. 1**], which attains a length of 10 ft (3 m) and a weight of 250 lb (113 kg). Several small geckos (**Fig. 2**) and African chameleons are about 1.5 in. (4 cm) long when



Fig. 1. Komodo dragon (*Varanus komodoensis*), Komodo National Park, Indonesia.

fully grown. See CHAMELEON; GECKO.

Physiology. The senses of lizards include the usual ones found in all terrestrial vertebrates. Thermal, tactile, and pain receptors are distributed over the body surface.

Olfaction. The principal chemoreceptors are the nasal organ proper, and the organ of Jacobson or the vomeronasal organ, a specialized derivative of the olfactory apparatus. The nasal organs open into the mouth cavity through the internal nares and are connected to the outside by the external nares. Inspired air passes through the organ on its way to the throat and lungs. Small particles of substances in the air are trapped by the nasal mucosa and analyzed by sense cells in this lining. The paired organs of Jacobson lie anterior to the olfactory sacs and are separated from them. They open only into the mouth, and particles in watery solution from this region are analyzed by them. In many lizards the tongue has a divided tip, and the two divisions are used to pick up particles outside the mouth and bring them back into it. The tips are then rubbed against the openings of the vomeronasal organs. Jacobson's organs are indicated as separate areas of the nasal organ in Amphibia and turtles, are partially separated from the main sac in Rhynchocephalia, and are vestigial in crocodiles, birds, and most mammals. See CHEMORE-CEPTION; OLFACTION; SENSE ORGAN.

Hearing. The auditory apparatus of lizards shows remarkable variation correlated with the mode of life. In the majority of forms, a well-developed eardrum is present in a slight recess at the posterior margin of the head near the quadrate bone. A ligament attaches the small middle-ear ossicle, the stapes, to the inner face of the tympanum, and the other end



Fig. 2. Western banded gecko (Coleonyx variegatus). (Photo by NPS/Tonto National Monument, Arizona)

Comparison of Sauria and Serpentes				
Determining structure	Sauria	Serpentes		
Limbs Eyelids External ear opening Pectoral girdle Braincase	Usually two or four; sometimes none Usually movable; sometimes immovable Usually present; sometimes lacking Usually present; rarely none Usually open anteriorly, bounded by connective	None Immovable None None Completely bony anteriorly		
Mandible	tissue sheath instead of bone Two halves, usually articulating	Two halves connected by elastic ligament, not articulating		

fits into an opening into the inner-ear region. Sound is transmitted from the eardrum along the stapes to the fluid of the inner ear. In many burrowing lizards, the eardrum is partially or completely covered by skin and scales or is lost entirely. Most of these forms are insensitive to airborne sounds and receive their auditory stimuli from the substratum through the lower jaw to the quadrate and then to the stapes. *See* EAR (VERTEBRATE); HEARING (VERTE-BRATE); PHONORECEPTION.

Vision. Lizards have moderately good vision except for burrowing forms in which the eyes are reduced or covered over by skin, scales, or even bone. Most lizards with unreduced eyes have color vision, but some nocturnal families such as the Gekkonidae (geckos) and Xantusiidae (night lizards) are colorblind. In certain forms, the movable eyelids have a transparent window through which an image may be seen even when the eyes are closed. In several groups, the two eyelids have fused and are completely transparent. The fused lids form a protective spectacle in geckos, night lizards, certain burrowing skinks, and other forms. *See* COLOR VISION; EYE (VER-TEBRATE); VISION.

Locomotion. The majority of lizards are quadrupedal in locomotion and are usually ambulatory scamperers or scansorial (adapted for climbing) [Fig. 3]. Some forms are bipedal, at least when in haste. Perhaps the most famous of the bipedal forms is the Jesus Cristo lizard, *Basiliscus*, of Middle America which runs for some distance across the surface of small streams when frightened. In many burrowing lizards, the limbs are reduced so that sometimes only the anterior pair is present or all are gone. In these species, locomotion is of a serpentine, curvilinear



Fig. 3. Tree lizard (Urosaurus ornatus). (Photo by NPS/Tonto National Monument, Arizona)

type. One genus of Asian lizards, *Draco*, has large membranes between its limbs that can be used as gliding surfaces to increase the distances of leaps and to reduce the rate of descent.

Thermoregulation. Lizards, in common with all other reptiles, are dependent upon external sources for maintenance of their body temperatures. They are ectothermic. Thermoregulation is attained by means of gross behavioral adjustments. A lizard moves into shade if its temperature rises, and moves into areas of higher temperatures as its temperature drops. Individuals are most active between temperatures of 60 and 108°F (16 and 42°C) with optimum activity between 80 and 90°F (27 and 32°C). Temperatures above 110°F (43°C) are lethal, and no lizard can survive in direct desert sunlight for more than 15-30 min. *See* TEMPERATURE ADAPTATION; THERMOREG-ULATION.

Respiration and excretion. Respiration is achieved by means of paired lungs. The excretory product of protein metabolism is an insoluble substance, uric acid. The feces are dark, oblong, and capped by a white mass of uric acid when deposited. *See* EXCRETION; RESPIRATION; RESPIRATORY SYSTEM.

Coloration. The coloration of each species of lizard is characteristic. Most forms exhibit marked differences in coloration between the sexes, at least during the breeding season, and frequently the young are markedly different from the parents. Color changes occur in rapid fashion among some species, and all are capable of metachrosis or changing color to a certain extent. Contrary to popular legend, the changes are not made to match the color of the background; they seem, rather, to be under neurohormonal control and to fluctuate with temperature and light and with the activities of the lizard. *See* CHROMATOPHORE; SEXUAL DIMORPHISM.

Sound production. Many lizards make sounds ranging from hisses that are produced by simple expulsion of air through the glottis, to squeaks of small forms and loud cluckings of the geckos. Sounds are used for defensive bluffing and probably for recognition as well.

Defense mechanisms. The basic defense against discovery utilized by most lizards is "freezing" motionless in position. The majority are protectively colored and when still are extremely difficult to see. However, if they are located, rapid flight becomes the second line of defense. Even if a predator catches a lizard, the game is not over, because many forms bite and claw. Among the more interesting specialized

defensive devices of saurians are the bluffing techniques of certain forms. The Australian frilled lizard, Chlamydosaurus, of the family Agamidae, can expand a huge nuchal collar which when presented with open mouth to a predator is an effective deterrent to aggression. The agamid, Uromastix, of Africa and the iguanid, Ctenosaura, of Middle America have spiny tails that may lash out to injure a potential enemy. The horned lizards, Phrynosoma, of the United States and Mexico are armed with numerous spines, particularly on the head, and make a rather unpleasant mouthful. In addition, they have the capacity of squirting blood from ruptured vessels in the lower eyelids into the mouth or face of a predator. The blood contains chemical substances, secreted into it in the eye region, that are extremely repugnant to coyotes and dogs. A great many lizards are able to throw off their tails voluntarily (tail autotomy), and reflexes in the tail make it thrash about to distract a predator from its real objective. In many cases, the tail is brightly colored to add to the illusion. The tail will regenerate on the lizard, but the new tail will lack a bony support. See PROTECTIVE COLORATION; REGENERATION (BIOLOGY).

Venomous species. There are only two species of venomous lizards, both members of the genus Heloderma, in the family Helodermatidae. The Gila monster (H. suspectum) [Fig. 4] is found in western New Mexico, Arizona, extreme southern Nevada, southwestern Utah, eastern California, and northwestern Mexico. The beaded lizard (H. borridum) is a Mexican species. Venom is produced in glands along the lower jaw and penetrates wounds inflicted by the anterior recurved teeth (that is, teeth that curve back toward the animal). No mechanism for injection of the venom into the wound is present, contrary to the situation in most snakes. Although occasionally fatal, the bites of these large lizards, of which a length of 3 ft (0.9 m) is not uncommon, are not as dangerous as those of snakes, nor are human beings bitten frequently.

Rhythmic activity. Saurians show rather definite rhythms in their activities. In temperate regions, seasonal cycles are correlated with temperature and sexual activity, and hibernation occurs in the winter. General activity continues year-round in tropic climates. Daily activity cycles are typical of most species and include a definite daily pattern of feed-



Fig. 4. Gila monster (Heloderma suspectum). (Photo by NPS/Tonto National Monument, Arizona)

ing, drinking, resting, basking, and excreting. Most lizards are diurnal, and their daily cycles are usually correlated with temperature. Certain families, notably the Gekkonidae and Xantusiidae, tend strongly toward nocturnal habits.

Territoriality. Most lizards spend their entire life within a rather restricted area and return at the end of each day to the same site to sleep. The males of the majority of forms exhibit territorial behavior and defend a central portion of their home range against trespass by other males of the same species. Defense of the territory usually involves much bluffing and noise, and the "defending" lizard is usually the victor. Various color patches, scale crests, and other distinctive marks play a part in intimidation and display at these times. Fights frequently occur and generally consist of attempts at biting. Male monitor lizards, Varanidae, rear up on their hindlegs and bite and slash with the front claws. No truly gregarious lizards are known, and none are migratory. See TERRITORI-ALITY

Courtship activity. Males usually breed with any female that they meet in moving about their home ranges during the breeding season. A complex courtship pattern involving a chase and considerable prenuptial nuzzling and biting is typical. During copulation the male frequently grasps the head or neck of the female in his mouth, and the tails of the breeding pair are entwined so that their cloacas are brought into contact. The male has paired hollow hemipenes which normally lie retracted into the base of the tail. When engorged with blood, these organs are everted to be used singly.

Reproduction. Most species produce eggs that have leathery or calcareous shells. The eggs are buried in the soil or hidden in decaying logs or under bark. In many cases, the female guards the eggs against predators. Some species are ovoviviparous and retain the eggs within the body until they hatch. Others are truly viviparous with a seroallantoic placental connection between mother and embryo. Gestation or incubation is generally 5–8 weeks. In oviparous and ovoviviparous forms, at least, the young have a special egg tooth which grows up from the tip of the upper jaw. This tooth is used to cut through the egg membranes and the shell at the time of hatching and is lost shortly thereafter.

Classification of lizards. The following list indicates the major evolutionary lines, families, and distribution of lizards. Families indicated by an asterisk contain limbless, snakelike species (based on M. J. Benton, 2005).

Iguania line

- Family Iguanidae: the Americas, Madagascar, Fiji Agamidae: Africa, southern and central Asia, Australia
 - Chamaeleonidae: Africa, Madagascar, southern India, Near East, southern Spain

Gekkota line

Family Gekkonidae: circumtropical, all continents and most continental and oceanic islands Pygopodidae: Australia, New Guinea Scincomorpha line

Family Xantusiidae: North and Middle America, Cuba Teiidae: the Americas

Lacertidae: Africa, Eurasia

*Gerrhosauridae: Africa, Madagascar

*Scincidae: cosmopolitan, except frigid areas

Gymnophthalmidae: Central and South America

Cordylidae: southern Africa

Annulata line

*Family Amphisbaenidae: tropics of Africa and America, Iberian Peninsula

*Bipedidae: Mexico

Trogonophidae: northern Africa and Middle East

Rhineuridae: Florida (U.S.)

Anguimorpha line

*Family Anguidae: the Americas, Eurasia *Anniellidae: the Californias

> Xenosauridae: southeastern China, Central America

Helodermatidae: North America Varanidae: Africa, southern Asia, Australia Lanthonotidae: North Borneo

Serpentes

Snakes (Figs. 5 and 6) are basically specialized, limbless lizards which probably evolved from burrowing forms but have now returned from subterranean habitats to occupy terrestrial, arboreal, and aquatic situations. In addition to those features already mentioned as distinguishing the snakes from lizards, the following characteristics are typical of all serpents. There is no temporal arch so that the lower jaw and quadrate are very loosely attached to the skull. This gives the jaw even greater motility than is the case in lizards. The body is elongate with 100-200 or more vertebrae, and the internal organs are elongate and reduced. A spectacle covers the eye, and there is no tail autotomy. In most features of their biology, serpents resemble lizards; the two groups are closely related.

The largest living snake is the Indian python *(Python reticulatus)*, which reaches 30 ft (9 m) in length and a weight of 250 lb (113 kg). The largest



Fig. 5. Western coral snake (Micruroides euryxanthus). (Photo by NPS/Tonto National Monument, Arizona)



Fig. 6. Eastern ribbon snake (Thamnophis sauritus). (Photo by NPS/Little River Canyon National Preserve, Alabama)

venomous snake is the king cobra (*Ophiophagus bannab*) of southern Asia, which is known to attain a length of 18 ft (5.5 m). A number of small tropical burrowing snakes are vastly smaller by comparison, being about 5 in. (12.5 cm) in length when fully grown. The longevity of snakes is not well known, but one captive cobra lived 28 years and many of the larger snakes probably live at least 25 years.

Physiology. The senses of snakes are fundamentally similar to those of lizards. Great dependence is placed upon olfaction and the Jacobson's organs. The tongue of all snakes is elongate and deeply bifurcated. When not in use it can be retracted into a sheath located just anterior to the glottis, but it is protrusible and is constantly being projected to pick up samples for the organs of Jacobson from the surrounding environment. Snakes are deaf to airborne sounds and receive auditory stimuli only through the substratum via the bones of the head. The eyes are greatly modified from those in lizards, and there is no color vision. Some groups are totally blind and have vestigial eyes covered by scales or skin.

Thermoreception. Specialized thermoreceptors, not found in any other vertebrates, are present in many boas and pythons (family Boidae, and in all pit vipers) (subfamily Crotalinae). In the boids, these receptors form a series of depressions in the lip scales and are called labial pits. In the crotalids, there is a single, large loreal pit located about midway between the eye and nostril and somewhat below them on the side of the head. These organs are sensitive to temperature differences of about $2^{\circ}F(1.1^{\circ}C)$ and are apparently used to locate prey whose temperature may be higher or lower than that of the surrounding environment.

Locomotion. Four basic patterns of locomotion are found in snakes, and several may be used by a particular individual at different times. The most familiar type is serpentine curvilinear (lateral undulatory). In this pattern, the snake moves forward by throwing out lateral undulations of the body and pushing them against any irregularity in the surface. Snakes using rectilinear locomotion move forward in a straight line, without any lateral undulations, by producing wavelike movements in the belly plates. Laterolinear locomotion, or sidewinding, is used primarily on smooth or yielding surfaces and is very complex. In essence, the snake anchors a portion of the body in the substratum and lifts the rest out laterally to a new position. When the lifted part is anchored again, the portions of the body left behind are lifted to the new position. By a constant lifting and anchoring of alternate parts, the snake moves in a lateral direction. Concertina locomotion movement resembles the expansion and contraction of that musical instrument.

Other activities. The basic bodily functions of snakes—thermoregulation, respiration, excretion, and elimination—are the same as those in lizards.

Defense mechanisms. Snakes and lizards are alike in utilizing procrypsis and flight as their principal means of defense. Protective coloration is of prime importance in the first of these responses. Although all snakes are capable of some color change, none show the marked changes characteristic of many lizards. Snakes are excellent bluffers, and hissing and striking, sometimes with the mouth closed, may be used to intimidate or discourage possible predators. Among the most effective deterrents to attack are the specialized warning devices of the rattlesnakes (Fig. 7). Rattlesnakes, however, are not bluffing, because they can back up their threats with an injection of venom. These rattles are actually an extension of the integument; moreover, each time the individual organism sheds (ecdysis), a new rattle is added. Contrary to popular lore, the age of the snake cannot be determined by the number of rattles. Some snakes may shed once over a year period, while other snakes may shed three times.

Ecology. The seasonal and daily activity patterns of snakes are similar to those described above for lizards. More snakes are nocturnal than diurnal, and they generally have a larger home range than do lizards. Home sites are found in most species, but no territorial behavior is known. A "combat dance" in which the males rear upright and push and weave around one another occurs frequently. The dance has often been described as being a courtship display, but it always involves two males. Its exact func-



Fig. 7. Arizona black rattlesnake (Crotalus viridis cerberus). (Photo by NPS/Tonto National Monument, Arizona)

tion remains a mystery. Snakes are generally nongregarious, although the sea snakes, Elapidae, travel in large schools. Migrations occur in temperate regions to and from den sites, and the sea snakes migrate annually to breeding grounds.

Courtship. Courtship is distinctive for each form but usually involves tracking of the female by the male, considerable nuzzling and rubbing of the female by the male, and a wrapping of the bodies of the breeding pair tightly around one another for mating. Male snakes utilize their hemipenes one at a time. Eggs are buried or hidden and are leathery or calcareous. Development is oviparous, ovoviviparous, or viviparous. Incubation or gestation takes about 60–100 days.

Nutrition. Food is almost always of animal material such as amphibians, reptiles, birds and their eggs, and mammals. However, some forms eat lizard and snake eggs to a large extent, as well as insects, mollusks, or fish; most snakes probably eat carrion. The majority of species are rather specific in food preference. Prey may simply be grabbed and swallowed, or specialized techniques may be employed, depending upon the species. Constriction of the victim is practiced by many forms. In this case, the food organism is large and is grasped in the snake's mouth while coils of the muscular body are wrapped around the prey. The coils are gradually tightened until the victim's heart stops or suffocation occurs. Venom may be injected into the body of the prey by certain forms. Food articles are always swallowed whole. The process is complex but consists primarily of a series of coordinated movements of the head and jaws. The lower jaws are connected to one another by an elastic ligament which allows each half to be moved independently of the other. The teeth on the jaws are recurved so that by alternately moving the right half of the jaw forward and then the left and hooking the teeth into the prey the snake pulls itself forward over the prey's body.

Venomous snakes. The vast majority of living snakes are harmless to humans, although a number are capable of inflicting serious injury with their venomous bites. The venom apparatus has evolved principally as a method of obtaining food, but it is also advantageous as a defense against attackers. Venomous snakes may be placed in three categories on the basis of their venom apparatus. Opisthoglyphodont snakes have fangs in the rear of the mouth. Proglyphodont snakes have fangs in the front of the mouth. This group can be subdivided into the proteroglyphodont in which the anterior fangs are fixed and immovable-for example, cobras, mambas, kraits, coral snakes, and sea snakes-and the solenoglyphodont in which the fangs are inserted on the movable maxilla so that they may be rotated forward when the mouth is open and folded back against the roof of the mouth when not in use, as in vipers.

Fangs are teeth modified for the injection of venom into the victim, and the venom glands are modified salivary glands connected to the grooved fangs by a duct. Special muscles are present in all proglyphous snakes to force the venom into the wound. The

318 Squash

venom itself is a complex substance containing a number of enzymes. Certain of these enzymes attack the blood, others the nervous system, and some are spreaders. The venom of each species is distinct and contains different kinds, combinations, and proportions of the deleterious enzymes.

The rear-fanged and proteroglyphodont forms have relatively short fangs and require some chewing of the victim in order to inject the venom. Small rearfanged forms are consequently harmless to human beings because they cannot get their mouths around any part of a person's body to bring the fangs into contact. Vipers, however, have very long fangs, and they usually strike or bite with blinding speed. Their injection mechanism is efficient, and a bite usually means that a considerable amount of venom has been injected.

Classification of snakes. The following list indicates the major groups of living snakes and their distribution.

Family Typhlopidae: circumtropical

- Leptotyphlopidae: circumtropical Anomalepididae: southern Central America and northern South America
- Aniliidae: southwestern Asia and Central America

Anomochilidae: Malaysia, Sumatra, Borneo Boidae: circumtropical, western United States Bolyeriidae: Round Island (Indian Ocean) Tropidophiidae: West Indies, Central

- America, South America
- Acrochordidae: India, Solomon Islands, Australia
- Colubridae: cosmopolitan, except frigid areas; some species are opisthoglyphodont Elapidae: Africa, Asia, Australia, the Americas;
- all species are proglyphodont
- Viperidae: Eurasia, Africa; all species are proglyphodont

W. Ben Cash; Jay M. Savage

Bibliography. M. J. Benton, *Vertebrate Paleontology*, 3d ed., Blackwell Publishers, Oxford, 2005; F. H. Pough et al., *Herpetology*, 3d ed., Prentice Hall, Upper Saddle River, NJ, 2004; G. R. Zug, L. J. Vitt, and J. P. Caldwell, *Herpetology*, Academic Press, San Diego, 2001.

Squash

The common name for edible fruits of several species of the genus *Cucurbita*: *C. pepo*, *C. moschata*, *C. maxima*, and *C. mixta*. Those species originated in the Americas but are now grown in most countries around the world. Within squash there is a tremendous variation in size, shape, color, and usage (see **Fig. 1**).

Summer squash. The most clearly defined group is summer squash, fruit of any species of *Cucurbita* eaten as a vegetable when immature. It is most commonly *C. pepo*, and its plants are usually bushy rather than having long runners. In some countries fruits

may be picked as early as the day the flowers open, but they are eaten at all stages up to the time when the rind starts to harden. Fruit color may be white, yellow, or light or dark green, and the green may be solid or striped. Shapes may be flattened disks as in Pattypan, cylindrical as in Zucchini and Cocozelle, or with necks as in the straightneck and crookneck types. Summer squash has mild flavor, high water content (about 95%), and relatively low nutritional value. It is not especially high in any one nutrient but contributes modest amounts of several vitamins and minerals.

Winter squash. Winter squash is fruit of *Cucurbita* eaten when mature and derives its name from its ability to be stored for several weeks or months without consumption. Varieties of winter squash are found in all four species. The Table Queen group,



Fig. 1. Three varieties of squash. (a) Zucchini. (b) St. Pat Scallop squash (W. Atlee Burpee Co.). (c) Crookneck squash (Asgrow Seed Co., The Upjohn Co.).

synonymous with Acorn, is *C. pepo*, Butternut belongs to *C. moschata*, Green-striped Cushaw is *C. mixta*, while *C. maxima* has the widest range of types, including Buttercup, Hubbards, and Delicious of various colors, Banana, and Boston Marrow. Flesh color varies from light yellow to dark orange, and the edible portion ranges from thin to very thick, depending on variety.

Characteristics. The percentage of soluble solids, which are mostly sugars, in cooked winter squash is commonly 8-10% and may be as high as 16%, putting squash in the same range as good melons. Squash is a good source of calories and of vitamin A, particularly those with dark orange flesh, which are high in carotene. Varieties tend to have their own characteristic flavors: the greatest difference is between the Table Queen group and the other winter squashes. Plants of winter squash are characteristically viny in habit with runner lengths of 12 to 20 ft (3.6 to 6 m) under good growing conditions, but bush or semiviny forms have been developed in some of the types. The bush forms of winter squash tend to be lower in solids and quality, especially if they set a heavy load of fruit.

There is some confusion in terminology between winter squash and pumpkin, and the use of these names differs from place to place. Both names refer to mature fruits, and varieties called squash and others called pumpkins may be found in all four *Cucurbita* species. In the United States, varieties eaten as a vegetable are generally called squash, while those used for pies, stock feed, or autumn decorations are usually called pumpkins. Pumpkins tend to be round with orange exterior color and flesh that is paler, more fibrous, and less sweet than squash. *See* PUMP-KIN.

Propagation. Squash is propagated by seed and its flowering habit is monoecious, that is, with separate male and female flowers on the same plant. Male flowers usually appear several days before the first female. Pollination is done naturally by insects, and within a species types which look very different cross readily with each other. However, the fruit is not affected in the year crossing occurs. Natural crossing occurs only rarely between the different species. Separation of the sexes makes controlled crossing relatively easy, and first-generation hybrids are increasingly being used as varieties, especially in summer squash where hybrid vigor is manifested in earlier flowering. One result of hybridization is that Butternut, Table Queen, Zucchini, and others which were originally single varieties have become groups of varieties with similar fruit shape but differing fruit sizes, colors, times of maturity, and plant sizes. See BREEDING (PLANT).

Squash seed is usually planted directly in the field after the soil is warm, in rows 4-6 ft (1.2-1.8 m) apart for bush type and 6-9 ft (1.8-2.7 m) apart for vining types. Plants may be spaced from 2-3 ft (0.6-0.9 m) apart in both cases. In cool regions, plants may be started in a greenhouse, provided they are grown in containers which permit transplanting without much disturbance of roots and before they become too large. The best size for transplanting may be reached in 10–14 days under good growing conditions.

Harvesting and storage. Summer squash begins to bear in 50-60 days from planting. The fruit grows very rapidly and must be harvested at frequent intervals (3-4 times per week) if one wishes the plants to continue bearing for an extended period.

Winter squash should be harvested when fully mature for best quality. Unless planted very early, it is often left in the field until the first light frost occurs, then harvested before the exposed fruit can be damaged by subsequent frost. Storage temperatures of $50-60^{\circ}$ F ($10-16^{\circ}$ C) are best because chilling injury occurs at lower temperatures and shortens storage life. For all varieties except Table Queen, a curing period of 1–2 weeks at 70–80°F ($21-27^{\circ}$ C) may improve quality by hastening conversion of starch to sugar. H. M. Munger

Diseases. There are more than 40 diseases of squash and about 16% of the annual crop is lost due to diseases. Diseases caused by the mosaic group of viruses are of primary importance because of their widespread occurrence in all areas where the crop is grown, both in the United States and elsewhere. These viruses produce such striking symptoms in infected plants, such as malformation and mottling of the leaves (**Fig. 2**) and distortion and abnormal discoloration of the fruits (**Fig. 3**). These viruses can be transmitted by seed, insects, and mechanical contact. Effective control measures are the use of virus-free seed, eradication of weed hosts of the viruses, and elimination of insect vectors.

Seed decay is common when squash is planted in cool wet soil, so treatment of the seed with registered fungicides is essential. The fungus *Fusarium solani* causes stunting, yellowing, and wilting of squash plants. The most effective control measures against this disease are the use of noncontaminated seed and extended crop rotations. In the arid southwestern United States and in squash production areas throughout the world with similar climates, powdery mildew, *Sphaerotheca fuliginea*, is a severe production-limiting disease, This disease can be easily recognized by the abundant production of white masses of spores on the leaves of infected plants so that they appear to have been sprinkled with a



Fig. 2. Squash leaves, mottled by mosaic virus.



Fig. 3. Crookneck squash. (*a*, *b*) Fruit mottled by mosaic virus. (*c*) Healthy fruit.

heavy white powder. In the late 1960s, effective systemic fungicides were developed that gave excellent control of this disease. By the mid-1970s, however, the fungus had begun to develop strains resistant to these fungicides, and growers returned to the use of older, more conventional, protective fungicides. There have been efforts by plant breeders to develop squash varieties and hybrids with genetic resistance to powdery mildew.

When heavy, prolonged dew periods are prevalent during the growing season, the downy mildew fungus, *Pseudoperonospora cubensis*, severely defoliates the plants. Since no effective level of resistance to this disease is available in commercial varieties, the only practical control procedure is the application of protective fungicides at regular intervals throughout the growing season. *See* PLANT PATHOL-OGY. Claude E. Thomas

Bibliography. D. M. Bates, R. W. Robinson, and C. Jeffrey (eds.), *Biology and Utilization of the Cucurbitaceae*, 1989.

Squatiniformes

An order of bizarre sharks, commonly known as angel sharks, comprising one family, one genus (*Squatina*), and 15 species that superficially resem-



Pacific angel shark (Squatina californica). (Photo by Tony Chess/NOAA Fisheries, Southwest Fisheries Science Center)

ble rays (see **illustration**). The following combination of characters identify the angel sharks: body flattened dorsoventrally; pectoral fins winglike and spread forward alongside head but not attached to it; mouth almost terminal; nostrils terminal, with barbels on anterior margin; eyes and large spiracles on top of head; five gill slits partly on sides of head and partly on ventral surface; two small spineless dorsal fins on tail; anal fin absent; and lower lobe of caudal fin longer than upper lobe. Maximum lengths of three species vary from 63 to 89 cm (25 to 35 in.), and 12 species vary from 120 to 200 cm (47 to 79 in.).

Angel sharks are found primarily in cool temperate to tropical zones of the eastern and western Pacific, eastern and western Atlantic, and extreme southwestern Indian Ocean, but they do not occur in most of the Indian Ocean or in the central Pacific reaches of Oceania (that is, the archipelagoes and islands of the central and south Pacific). They have a wide depth range, from nearshore to the upper continental slopes at 1300 m (4265 ft). They feed on a variety of small bony fishes, crustaceans, cephalopods, gastropods, and bivalves by using their highly protrusible jaws to snap up prey at high speed. *See* ELAS-MOBRANCHII; SELACHII. Herbert Boschung

Bibliography. L. J. V. Compagno, Checklist of Chondrichthyes, pp. 503-547 in W. C. Hamlett (ed.), *Reproductive Biology and Phylogeny of Chondrichthyes: Sharks, Batoids, and Chimaeras,* Science Publishers, Enfield, NH, 2005; L. J. V. Compagno, Checklist of living elasmobranches, pp. 471-498 in W. C. Hamlett (ed.), *Sharks, Skates, and Rays: The Biology of Elasmobranch Fishes,* Johns Hopkins University Press, Baltimore, 1999; L. J. V. Compagno, *FAO Species Catalogue*, vol. 4: *Sharks of the World: An Annotated and Illustrated Catalogue of Shark Species Known to Date, Part 2, Carcharbiniformes,* FAO Fish. Synop. no. 125, pp. 251-655, 1984; J. S. Nelson, *Fishes of the World,* 4th ed., Wiley, New York, 2006.

Squeezed quantum states

Quantum states for which certain variables can be measured more accurately than is normally possible.

All matter and radiation fluctuate. Much random fluctuation derives from environmental influence, but even if all these influences are removed, there remains the intrinsic uncertainty prescribed by the laws of quantum physics. The position and momentum of a particle, or the electric and magnetic components of an electromagnetic field, are conjugate variables that cannot simultaneously possess definite values (Heisenberg uncertainty principle). It is possible, however, to have the position of a particle more and more accurately specified at the expense of increasing momentum uncertainty; the same applies to electromagnetic field amplitudes. This freedom underlies the phenomenon of squeezing or the possibility of having squeezed quantum states. With squeezed states, the inherent quantum fluctuation may be partly circumvented by focusing on the less noisy variable, thus permitting more precise measurement or information transfer than is otherwise possible. *See* UNCERTAINTY PRINCIPLE.

Vacuum and laser light. According to quantum electrodynamics, the vacuum is filled with a free electromagnetic field in its ground state that consists of fluctuating field components with significant noise energy. If ϕ is a phase angle and $a(\phi)$ and $a[\phi + (\pi/2)]$ are two quadrature components of the field (for example, the electric and magnetic field amplitudes), the vacuum mean-square field fluctuation is given by Eq. (1), independently of the phase angle.

$$\left\langle \Delta a^2(\phi) \right\rangle = \frac{1}{4} \tag{1}$$

Equation (1) is normalized to a photon; the corresponding equivalent noise temperature at optical frequencies is thousands of kelvins. Equation (1) also gives the general fluctuation of an arbitrary coherent state, which is the quantum state of ordinary lasers. Further environment-induced randomness is introduced in addition to Eq. (1) for other conventional light sources, including light-emitting diodes. *See* CO-HERENCE; ELECTRICAL NOISE; LASER; QUANTUM ELEC-TRODYNAMICS.

Squeezed-state light. In a squeezed state, the quadrature fluctuation is reduced below Eq. (1) for some ϕ , as given in Eq. (2). At that point, squeez-

$$\langle \Delta a^2(\phi) \rangle < \frac{1}{4} \tag{2}$$

ing, that is, reduction of field fluctuation below the coherent state level, occurs. The fluctuation of the conjugate quadrature is correspondingly increased to preserve the uncertainty relation, Eq. (3). In a two-

$$\left\langle \Delta a^2(\phi) \right\rangle \left\langle \Delta a^2\left(\phi + \frac{\pi}{2}\right) \right\rangle \ge \frac{1}{16}$$
 (3)

photon coherent state, or squeezed state in the narrow sense, Eq. (3) is satisfied with equality. As a consequence, for a fixed energy E, the field quadrature signal-to-noise ratio is maximized by a two-photon coherent state. In particular, it becomes 4E(E + 1)compared to 4E for a coherent state. The two-photon coherent states are nothing but a slight generalization of the well-known minimum uncertainty states already studied by E. Schrödinger. The designation squeezed state is partly derived from the fact that the noise circle of Eq. (1) is squeezed to an ellipse when Eq. (3) is satisfied with equality (see illus.). Such states are called two-photon coherent states partly because only even counts of photons are possible for them when the mean field amplitude is removed, which is very characteristically a quantum feature

Generation and detection. Squeezed light can be generated by a variety of processes, especially nonlinear optical processes. To detect squeezing, an improved form of optical homodyne detection is employed called balanced homodyning. The first successful experimental demonstration of squeezing, in 1985, involved a four-wave mixing process in an



Field-amplitude fluctuation $\langle \Delta a^{+3}(\phi) \rangle$ as a function of phase angle ϕ . The noise circle of the vacuum state, Eq. (1) in the text, is squeezed to an ellipse according to Eq. (3).

atomic beam of sodium atoms. Since then, squeezing has been observed in a variety of other systems; the most pronounced was seen in parametric oscillation and amplification with about a factor of 4 below Eq. (1). *See* NONLINEAR OPTICAL DEVICES; NONLIN-EAR OPTICS.

Potential applications. Squeezing was first studied in connection with optical communication, although it was evident that reduced quantum fluctuation might find applications in precision measurements. Utilization of squeezed light has been proposed for interferometry, gravitational-wave detection, and spectroscopy, among other areas. The full realization of these applications is still remote. *See* OPTICAL COMMUNICATIONS.

Intensity squeezing. A phenomenon closely related to squeezing, which is also exhibited by some twophoton coherent states, is antibunching, where the photon number fluctuation in a sub-poissonian state is reduced below the coherent-state Poisson level. This sub-poissonian behavior is sometimes called amplitude squeezing or intensity squeezing. The generation, propagation, and detection of sub-poissonian light and squeezed light have similar problems; applications are similar and complementary. Strongly sub-poissonian light has been observed from constant-current-driven semiconductor lasers. *See* DISTRIBUTION (PROBABILITY).

Coherence and loss. Squeezed light and subpoissonian light constitute the most important classes of nonclassical light, specifically, those lights whose quantum state cannot be represented as a random superposition of coherent states. Nonclassical light with space-time coherence can be generated by nonlinear optical processes. On the other hand, the quantum fluctuation characteristic is sensitive to the losses introduced during propagation and detection. Loss not only attenuates the signal energy on nonclassical light, but it also mixes in vacuum (coherent-state) fluctuation quantum-mechanically, thus degrading the desirable quantum noise characteristic of nonclassical light. This could be a serious problem for obtaining good sources and for many applications. The loss problem could be alleviated in some situations by using appropriate quantum amplifiers, specifically, parametric amplifiers for squeezed light and photon-number amplifiers for sub-poissonian light. These amplifiers, in particular the photon-number amplifier, are not yet sufficiently developed for such applications. *See* PARAMETRIC AMPLIFIER.

Contractive states. In the context of a particle, the position and momentum take the roles of $a(\phi)$ and $a[\phi + (\pi/2)]$. A two-photon coherent state is then just a slightly generalized minimum-uncertainty wave packet. A subclass of these states called contractive states have been found to be of potential importance in gravitational-wave detection via the motion of a free mass. It has been shown that quantum measurements involving such states could, in principle, lead to arbitrarily precise monitoring of the position of a free mass, which was formerly thought to be impossible. *See* NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS.

Bibliography. R. Loudon and P. L. Knight, Squeezed light, *J. Mod. Opt.*, 34:709–759, 1987; Special issue: Quantum noise reduction in optical systems: Experiments, *Appl. Phys. B*, vol. 55, no. 3, 1992; Special issue: Squeezed states of the electromagnetic field, *J. Opt. Soc. Amer. B*, vol. 4, no. 10, 1987; H. P. Yuen, Two-photon coherent states of the radiation field, *Phys. Rev. A*, 13:2226–2243, 1976.

Squid

The common name applied to cephalopods of the order Teuthoidea. They are marine mollusks that inhabit the oceans of the world. Squids (see **illustration**) are characterized by having eight arms and two longer tentacles around the mouth; an elongated, tapered, usually streamlined body; an internal rodor bladelike shield (gladius); and fins on the body (mantle). The arms have two (infrequently four or six) rows of suckers and occasionally clawlike hooks, and the tentacles have terminal clubs with suckers, hooks, or both. The muscular, elastic tentacles are contractile, not retractile into pockets like those of cuttlefishes (Sepioidea). *See* SEPIOIDEA.



Short-finned squid (Illex borealis). (Donald Flescher; image courtesy of Marine Biological Laboratory Archives)

Squids have an exceptionally well-developed brain and organs of the central nervous system that approach in complexity and function those of fishes and even some birds and mammals. The eyes are large and provide acute vision. Thousands of pigment cells (chromatophores) in the skin are individually innervated to accommodate very rapid color and color pattern changes important in camouflage and behavior. Highly innervated photophores, capable of producing bioluminescent light, may be present on the skin, arms, or tentacles, or on the internal organs and eyes, and function to warn predators, eliminate the silhouette of deep-sea forms, or enable recognition of species or sexes. See BIOLUMI-NESCENCE; CHROMATOPHORE; NERVOUS SYSTEM (IN-VERTEBRATE); PHOTORECEPTION.

Squids are active, powerful swimmers, driven by jet propulsion as water taken into the mantle cavity is forcefully expelled through the funnel. Prey, normally shrimps, fishes, or other squids, are captured with the two tentacles and held with the arms while the beaks, like parrot beaks, cut off bites that the radula and tongue shove down the throat. Squids are preyed upon by toothed whales (sperm whales, dolphins), pinnipeds (seals, sea lions), fishes (tunas, mackerels, swordfish, groupers, sharks), and pelagic birds (petrels, albatrosses).

Two groups (suborders) of squids are recognized: Myopsida, like *Loligo pealei* and *Lolliguncula brevis* from the east coast of North America and the Caribbean and *Loligo opalescens* from the west coast of North America; and Oegopsida, the openocean and deep-sea forms like *Illex illecebrosus* of the western North Atlantic ocean and *Dosidicus gigas* of the eastern Pacific Ocean from off Chile to California.

Squids are important in biomedical research, especially that concerning neurophysiology and biophysics, in which the giant nerve axon plays a major role. Squids are eaten around the world; annual catches total about 1 million metric tons. An excellent source of protein, squid is very low in fats and high in essential minerals. *See* CEPHALOPODA; COLEOIDEA; TEUTHOIDEA. Clyde F. E. Roper

Bibliography. L. Bonnaud, R. Boucher-Rodoni, and M. Monnerot, Phylogeny of cephalopods inferred from mitochondrial DNA sequences, *Mol. Phylog. Evol.*, 7:44-54, 1997; D. L. Gilbert, W. J. Adelman, and J. M. Arnold (eds.), *Squid as Experimental Animals*, Plenum Press, New York, 1990; R. T. Hanlon and J. B. Messenger, *Cephalopod Behaviour*, Cambridge University Press, 1996; J. B. Messenger, Cephalopod chromatophores: Neurobiology and natural history, *Biol. Rev.*, 76:473-528, 2001.

SQUID

An acronym for superconducting quantum interference device, which actually refers to two different types of device, the dc SQUID and the rf SQUID.

The dc SQUID consists of two Josephson tunnel junctions connected in parallel on a



Fig. 1. Direct-current (dc) SQUID with enclosed magnetic flux Φ . *I* = applied current; *V* = generated voltage.

superconducting loop (Fig. 1). A small applied current flows through the junctions as a supercurrent, without developing a voltage, by means of Cooper pairs of electrons tunneling through the barriers. However, when the applied current exceeds a certain critical value, a voltage is generated. When a magnetic field is applied so that a magnetic flux threads the loop, the critical value oscillates as the magnetic flux is changed (Fig. 2), with a period of one flux quantum, $\Phi_0 = b/(2e) \approx 2.07 \times 10^{-15}$ weber, where *b* is Planck's constant and *e* is the electron charge. The oscillations arise from the interference of the two waves describing the Cooper pairs at the two junctions, in a way that is closely analogous to the interference between two coherent electromagnetic waves. See INTERFERENCE OF WAVES; JOSEPHSON EF-FECT; SUPERCONDUCTIVITY.

The rf SQUID consists of a single junction interrupting a superconducting loop. In operation, it is coupled to the inductor of an *LC*-tank circuit excited at its resonant frequency by a radio-frequency (rf) current. The rf voltage across the tank circuit oscillates as a function of the magnetic flux in the loop, again with a period of one flux quantum. Although SQUIDs were for many years operated while immersed in liquid helium, ceramic superconductors with high transition temperatures make possible devices operating in liquid nitrogen at 77 K.

SQUIDs have important device applications. Usually with the addition of a superconducting input



Fig. 2. Maximum supercurrent I_c versus Φ/Φ_0 for dc SQUID, where Φ = enclosed magnetic flux, Φ_0 = flux quantum.

circuit known as a flux transformer, both dc and rf SQUIDs are used as magnetometers to detect tiny changes in magnetic field. The output of the SQUID is amplified by electronic circuitry at room temperature and fed back to the SQUID so as to cancel any applied flux. This makes it possible to detect changes in flux as small as 10^{-6} of one flux quantum with SQUIDs based on low-transition-temperature superconductors, corresponding to magnetic field changes of the order of 1 femtotesla in a 1-hertz bandwidth. Suitable modifications to the input circuit enable the SQUID to measure other physical quantities, including voltages, displacement, or magnetic susceptibility. SQUIDs are also used for logic and switching elements in experimental digital circuits and high-speed analog-to-digital converters. See ANALOG-TO-DIGITAL CONVERTER; INTEGRATED CIRCUITS; MAG-NETOMETER; SUPERCONDUCTING DEVICES.

John Clarke

Bibliography. A. Barone (ed.), *Principles and Applications of Superconducting Quantum Interference Devices*, 1992; A. Barone and G. Paterno, *Physics and Applications of the Josephson Effect*, 1982; J. Clarke, SQUIDs, *Sci. Amer.*, 271(2):46-53, August 1994; T. Van Duzer and C. W. Turner, *Principles of Superconductive Devices and Circuits*, 2d ed., 1999.

Squirrel

A rodent in the family Sciuridae. This family includes all tree squirrels and flying squirrels, as well as squirrels that live on the ground such as the woodchuck, marmot, chipmunk, and prairie dog. There are 51 living genera and 272 species living in a wide variety of habitats ranging through tropical rainforests, tundra, alpine meadows, and semiarid deserts in Europe, Asia, Africa, North America, and most of South Africa. They are absent from Australia, Madagascar, Polynesia, and the Sahara Desert. Locomotion is plantigrade (walking with the whole sole of the foot touching the ground). Sensitive vibrissae are present on the head, feet, and outsides of the legs. The dental formula is I 1/1, C 0/0, Pm 1-2/1, M 3/3 for a total of 20 or 22 teeth. The incisors grow continually throughout the animal's life. See CHIPMUNK; DENTITION; MARMOT; PRAIRIE DOG; RODENTIA.

Tree and flying squirrels. Tree and flying squirrels are mostly small and slender-bodied with bushy tails often as long as their bodies (see **illustration**). The tail is primarily used to maintain and correct the balance of the animal in leaping from branch to branch, but it may also be used as a rudder when leaping, as a flag to communicate social signals, and as a wraparound blanket when the animal sleeps. Soft pads are present on the soles of the feet, and sharp claws are present on all toes. These squirrels range in size from the African pygmy squirrel (*Myosciurus pumilio*) of Cameroon, Gabon, Congo, and Equatorial Guinea, which has a body length of about 60-70 mm (2-3 in.), a narrow tail of about 50-60 mm (2 in.), and weighs about 16 g (0.6 oz), to the large Oriental giant


Delmarva fox squirrel (Sciurus niger cinereus). (Photo by W. H. Julian/U.S. Fish and Wildlife Service)

squirrel (*Ratufa*) of India, Nepal, Borneo, and the Malay Peninsula, which may reach a length of over 0.9 m (3 ft) and weigh up to 3 kg (6.5 lb). Habitats vary considerably. The European pine squirrel (*Sciurus vulgaris*) and the North American red squirrel (*Tamiasciurus budsonicus*) dwell primarily in pine and evergreen forests, whereas the gray squirrel (*S. carolinesis*) of eastern North America is a resident of groves of oak, maple, and hickory trees. The largest tree squirrel in eastern North America, the fox squirrel (*S. niger*), favors open upland groves of oak, hickory, beech, and longleaf pine rather than the deep unbroken forest.

Tree squirrels are agile and are active during the day, unlike flying squirrels, which are nocturnal and have large eyes for seeing in dim light. They hoard nuts, seeds, and fruits in holes in trees or in underground storage sites. Food caches are located by memory and also by the keen sense of smell. Nests are constructed primarily of leaves and twigs and are located either in tree cavities or in the fork of a tree. Tree squirrels are active all year, although they may remain in their nests for extended periods during severe winter weather. Breeding usually occurs in February with litters being born in March or April. Newborn squirrels are hairless, blind, and helpless. Their eyes open in about 5 weeks.

Other tree squirrels include the flame and pygmy squirrels of South America; the Indian palm (*Funambulus*) and common oriental (*Callosciurus*) squirrels of Oriental Asia and the Pacific Islands; and the oil palm (*Protoxerus*), pygmy (*Myosciurus*), sun (*Heliosciurus*), and bush (*Paraxerus*) squirrels of Africa.

Flying squirrels are not really fliers but gliders. Their "flying" equipment consists of a thin, loose furry membrane (patagium) extending along each side of the body between the fore and hind limbs. As the animal launches into the air from some lofty bough, it spreads its legs, drawing the membranes taut. This wide expanse gives the squirrel enough support to permit it to glide in a downward course for a considerable distance. Before landing, the squirrel checks its speed by manipulating its tail and comes to rest head up on the trunk of a tree. The flying squirrel cannot sustain its flight for any long period of time; it lacks propelling organs like the wings of a bat or bird. It can, however, glide through the air up to a distance of 80 yards (73 m) or more. Flying squirrels are found in forested regions of Europe, Asia, and North America south to Honduras. They are absent from South America, Africa, and Australia. Nests are usually constructed in tree cavities but may be in birdhouses. Flying squirrels are active all year and feed on nuts, seeds, berries, and insects. Lichens and fungi may be eaten during the winter. On occasion, they kill and eat sleeping birds and bird eggs.

Ground squirrels. Ground-dwelling squirrels are usually larger and more heavy-bodied than tree squirrels. They have powerful forelimbs and large claws for digging. They live in subterranean burrows or have their underground dens among rocks. Burrows usually consist of a nest chamber and several side rooms. Ground-dwelling squirrels are omnivorous. Grasses, clover, forbs, flowers, and bulbs make up the bulk of the diet, but insects, bird eggs, and even carrion may also be eaten. They are diurnal, and some, such as woodchucks and marmots, hibernate. The 14 species of marmots (*Marmota*) are found throughout most of the United States and Canada as well as in the Alps of Europe and eastward in the mountains to China and southeastern Siberia.

The five species of prairie dogs (*Cynomys*) are stout bob-tailed rodents that were famous for their "towns" in the midwestern United States. These were actual underground cities with miles of wellworn tunnels and dens extending in every direction. Some large towns may have had millions of inhabitants. A few towns still remain, but prairie dog populations were decimated by cattle ranchers who killed them with strychnine because they consumed the same forage as the herds of cattle.

Chipmunks (*Tamias*) inhabit sunlit woodlands and open pastures. They favor stone walls, rocks, and fallen timber. The 25 species are found in Asia and North America. Chief foods are seeds, grains, nuts, and berries. Food is transported in their cheek pouches to underground storage sites.

Ground squirrels (*Spermophilus*, *Ammospermophilus*, *Atlantoxerus*, *Xerus*, and *Spermophilopsis*) are robust mammals that have rounded heads, short ears, short legs, and somewhat bushy tails that are short to medium in length. The 51 species are social animals that live in densely populated towns beneath the ground. They feed largely on green vegetation, seeds, grasshoppers, cicadas, and many other insects.

Most ground squirrels hibernate, with some entering hibernation as early as July and August and not coming out until February. One litter of 7 to 10 young is produced annually. *See* HIBERNATION AND ESTIVATION. Donald W. Linzey Bibliography. B. Grzimek (ed.), *Grzimek's Encyclopedia of Mammals*, McGraw-Hill, 1990; D. Macdonald (ed.), *The Encyclopedia of Mammals*, Andromeda Oxford, 2001; R. M. Nowak, *Walker's Mammals of the World*, 6th ed., The Johns Hopkins University Press, 1999.

SS 433

A binary star system in the Milky Way Galaxy consisting of a neutron star or a black hole accreting matter in the form of a disk from an early-type companion star. The most remarkable feature of the system is the presence of bipolar narrow outflows of matter (jets), with a velocity of 26% the speed of light, whose ejection axes change with time. *See* BLACK HOLE; NEUTRON STAR.

Discovery. SS 433 owes its name to a catalog published in 1977 by C. B. Stephenson and N. Sanduleak. They made an optical survey of the sky (the SS survey), searching for sources with a peculiarity in their emission spectrum, namely a strong hydrogen emission line. SS 433 is the 433rd object in their list of sources. Between 1967 and 1981, SS 433 was detected during other observing campaigns at different wavelengths, and appeared in many other catalogs as a radio source, an x-ray source, and as an optical variable star. Yet, the discovery that really drew the attention of scientists to this source was related to the characteristics of its hydrogen emission lines. See ASTRONOMICAL ATLASES; ASTRONOM-ICAL SPECTROSCOPY; RADIO ASTRONOMY; VARIABLE STAR; X-RAY ASTRONOMY.

Further optical observations of SS 433 in 1978 showed, besides the very strong hydrogen emission line in the spectrum (the so-called H α line, that is, the first atomic transition in the hydrogen Balmer series, at a wavelength of 656.3 nm), a few other lines whose wavelengths did not correspond to any known element. The extraordinary discovery, at first incomprehensible, was that the wavelengths of these "strange" lines also moved with time. A well-known effect that can cause shifts in the wavelength of a line is the Doppler effect. If the gas that emits a line is moving toward the observer, the line in the spectrum appears at a shorter wavelength than the wavelength at which it has been emitted at rest with the gas (that is, blueshift of the line), while if the gas is moving away from the observer, the line appears at a larger wavelength (that is, redshift of the line). What was difficult to understand was the fact that these moving lines appeared always in pairs, and moved smoothly in a periodic pattern. The solution of the mystery came one year later. See ATOMIC STRUCTURE AND SPECTRA; DOPPLER EFFECT.

Geometry of the jets. All the above mentioned unusual spectral features of SS 433 can be well fit with a relatively simple geometrical model, the kinematic model (**Fig. 1**). The system is ejecting matter in form of two antiparallel narrow (that is, with an opening angle less than 5°) jets. These jets precess in a 21° half-opening angle cone with a period of about 162



Fig. 1. Doppler shifts of the "moving" $H\alpha$ lines of (a) the approaching jet and (b) the receding jet of SS 433 versus the precession phase. The Doppler shift gives the ratio v/c, where v is the approach or recession velocity and c is the speed of light. The data are shown as open diamonds, and the solid lines are fits of the data using the kinematic model. (From S. S. Eikenberry et al., Twenty years of timing SS 433, Astrophys. J., 561:1027-1033, 2001. © The American Astronomical Society. Reproduced by permission of the American Astronomical Society.)

days. The components of the jets (probably discrete blobs of matter) move in a straight line away from the system. Since the ejection axes rotate on a cone (like a spinning top) and the angle between the precession axes and the line of sight is about 78° , the projection of the jets onto the plane of the sky results in a twisted trace (**Fig. 2**). The blobs of matter have a constant velocity of about 0.26*c*, where *c* is the speed of light. For the first time, relativistic jets (that is, moving at a significant fraction of the speed of light, in this case 26%), which had already been discovered in distant galaxies, were observed relatively close-by, from a stellar object in our own galaxy.

Electromagnetic radiation from jets. The physical processes involved in the first steps of collection, launch, acceleration, and collimation of matter in the form of a jet are still poorly understood. Most jets are thought to originate in the innermost regions of an



Fig. 2. Radio image of SS 433 obtained in 2003 with the Very Large Array. A contour plot of the intensity of radio emission is superimposed on a pixel image. The binary core is the brightest spot in the center, from which matter is outflowing in two anitparallel jets. On each side of the core is a twisted trace that is the projection onto the plane of the sky of the precessing jets.

accretion disk, and the outflowing axis to lie in a direction approximately perpendicular to the disk plane. Jet models for SS 433 that can be tested by observations concern the description of the electromagnetic radiation (light) emitted by the jets when they have already formed, say farther then 10^6 km (1 km = 0.6 mi) from the core of the system.

Emission lines in the spectrum of SS 433 are observed not only in the optical band but also in xray spectra, revealing the presence in the jets of heavy elements (such as iron) and highly ionized atoms. Emission lines indicate that: (1) the jets of SS 433 comprise, besides light particles like electrons, atomic nuclei; and (2) the gas that emits lines is thermal (that is, the state of the matter can be described by a macroscopic quantity like the temperature). The process that emits the light observed in the radio band from the jets of SS 433 is, instead, nonthermal and is due to synchrotron emission: Charged particles wrapping around a magnetic field, accelerate and thus emit light. *See* SYNCHROTRON RADIATION.

The model developed and generally accepted to describe the emission of radiation due to the thermal component of the jets is the adiabatic cooling model. The hypothesis of adiabatic cooling means that the gas ejected in the jets does not exchange energy with the environment. The gas has a certain density and temperature at the time of launch at the base of the jet, and both temperature and density decrease with the distance from the binary core only because of the expansion of the gas in a conical shape. The more dense and hot is the gas, the more it emits at shorter wavelengths. Close to the binary core, at distances less than 10^{6} - 10^{7} km, the jets' gas is at a temperature of about 10⁸ K, and emits light in the x-ray band. Then the gas in the jets cools, and at distances of about 10^{10} km from the core the temperature is of the order of 10⁴ K and the gas emits optical radiation. (This is the only really model-independent constraint on temperature and distance, coming from optical spectral analysis.) Based on the adiabatic cooling model, farther out the gas should be too cool to thermally emit optical or x-ray radiation. The nonthermal radio emission superimposed on this thermal radiation is probably caused by the highly relativistic free electrons in the tail of the thermal distribution of the gas.

The adiabatic cooling model breaks down beyond a certain distance from the binary core. In fact, spatially resolved x-ray observations of the jets of SS 433 showed highly ionized Doppler-shifted iron emission lines, emitted in the jets at distances of about 10^{12} km from the binary core, from gas with a temperature of the order of 10^8 K. This evidence indicates that reheating of atoms takes place in the jet flow, which is still moving with relativistic velocity more than 100 days after launch from the binary core (**Fig. 3**).

A possible explanation has been suggested for this reheating. Random variations of about 15% in the velocity of the blobs launched in the jets are inferred from long-term optical monitoring of the source. Therefore, a faster blob could catch up with a slower blob launched 162 days before with the same precession phase; this would produce a shock and a reheating of matter in the jet a few hundred days after its launch from the binary core. *See* SHOCK WAVE.

Further out, at distances of 10^{14} - 10^{15} km (10-100 light-years) from the binary core, x-ray observations show bright termination shocks, caused by the matter in the cold jets that hits the dense interstellar medium. This process stops the jets' run, releasing frictional energy in form of light. At those distances the two jets deform, like a seashell, the surrounding W50 nebula, which is thought to be a supernova remnant associated with the formation of the neutron star or black hole in SS 433. *See* INTERSTELLAR MATTER; NEBULA.

Binary system. As already noted, SS 433 is an x-ray binary system, that is, a binary star system in which a neutron star or a black hole is accreting matter in the form of a disk from a companion star. It is also an eclipsing binary system with an orbital period of about 13 days. Astronomers agree that the companion star is an early-type high-mass star (O-type or Atype stars are among the current hypotheses), with a mass that should exceed 5 times the mass of the Sun. However, the proper identification and classification of this star, and consequently (using the socalled mass function) the classification of the compact object as a neutron star or a black hole, is still controversial. In fact, the binary system is imbued with a dense envelope of gas probably coming from a strong stellar wind from the high-mass companion star. This envelope makes difficult any deep investigation of the two stars of the system. See BINARY STAR; ECLIPSING VARIABLE STARS.

Equatorial outflow. Radio observations have also shown outflows of matter in a direction perpendicular to the jets' axes, in the disk plane. This equatorial outflow can be associated to a windlike emission from the system (the high-mass companion) or to an extension of the accretion disk. It is not clear



Fig. 3. X-ray observations of SS 433 performed in 2000 with the *Chandra X-ray Observatory*. (a) X-ray image. The binary core of SS 433 is indicated with a cross and is not bright due to an instrumental effect. The lobes of the x-ray jets are visible on either side of the core. The squares indicate the regions of the jets from which the energy spectra are extracted. (b) Energy spectrum of east lobe. (c) Energy spectrum of west lobe. The spectra (indicated by crosses) are consistent with a thermal continuum (indicated by solid color lines) and Doppler-shifted highly ionized emission lines (indicated by color dotted lines).

yet if these outflows consist also of thermal matter, although x-ray observations seem to support this hypothesis.

Based on the principle that the physical processes behind the formation and evolution of relativistic jets are the same regardless the scale of the astrophysical objects involved (either a supermassive black hole in the center of a distant galaxy or a stellar collapsed object in a galactic binary system), the discovery of SS 433 gave astronomers the first near-by laboratory to study in more detail one of the most powerful phenomena of the universe, relativistic jets of matter. Simone Migliari

Bibliography. S. S. Eikenberry et al., Twenty years of timing SS 433, *Astrophys. J.*, 561:1027-1033, 2001; B. Margon, Observations of SS 433, *Annu. Rev. Astron. Astrophys.*, 22:507-536, 1984; S. Migliari, R. Fender, and M. Méndez, Iron emission lines from extended x-ray jets in SS 433: Reheating of atomic nuclei, *Science*, 297:1673-1676, 2002.

Stability augmentation

The alteration of the inherent behavior of a system. As an example, ships tend to exhibit significant rolling motions at sea. To dampen these rolling motions, a roll stabilization (feedback) system can be used. Such a system consists of a set of vanes (that is, small wings) extending outward from the hull, below the waterline. By varying the vane incidence angle relative to the hull, a hydrodynamic lift is generated on the vane. The vanes are driven by a feedback system so that the rolling motions are opposed by creating positive lift on one side of the hull and negative lift on the other side.

As a second example, nearly all satellites require some form of stability augmentation to help in keeping the antennas or sensors aligned with receiving equipment on Earth. The stability augmentation is effected by thrusters which receive their commands from a feedback system. *See* SPACECRAFT PROPUL-SION.

As a third example, stability augmentation systems are used on aircraft while airborne. This is usually achieved by a system which controls one or more flight-control surfaces (or engines) automatically without inputs from the pilot. The inherent stability and response behavior of many modern airplanes tends toward low damping or even instability. The physical reasons have to do with the configuration of the airplane and the combination of flight speed and altitude at which the airplane is operated. Several modern fighters and even some transports are intentionally designed with no or little inherent stability. There are a number of reasons for such a design condition. In the case of fighters, excellent maneuverability in combat is essential. By making a fighter intentionally inherently unstable, it is easy to design the control system so that load factors in pullups or in turns can be built up rapidly. In the case of transports, the motivation to design for little or no inherent stability is to lower the size of the tail and thereby achieve a reduction in drag and weight. See AIRPLANE; MILITARY AIRCRAFT.

Yaw-damper operation. If, for whatever reason, an airplane is deficient in the damping of its response or is divergent instead of stable, the flying qualities are compromised. This can be corrected by installing a stability augmentation system. Yaw dampers were among the first such systems designed into airplanes. The dutch-roll motion of an airplane consists of simultaneous oscillations of the bank (or roll) angle, the sideslip angle, and the heading angle; when this motion becomes poorly damped, it is difficult for a pilot to maintain manual control over the airplane when flying through turbulence. A yaw damper can usually solve this problem. It normally drives the rudder with the help of a servo (electrohydraulic or electromechanical). The servo receives its signal to move from a computer-amplifier combination. The computer has as its inputs the desired yaw rate and the actual yaw rate. The desired yaw rate of the airplane depends on the flight condition. If the desired condition is straight-line flight, the desired yaw rate will be zero. If the desired flight condition is a turn, the desired yaw rate depends on the steepness of the turn but is not zero. The actual yaw rate of the airplane is sensed by a yaw-rate gyro. The desired yaw rate is normally derived from the automatic pilot computer. The error signal, which is the difference between the two inputs (Fig. 1), is gradually driven to zero so that the output rate of the yaw damper equals the input rate. Generically, such systems are called feedback systems. *See* CONTROL SYSTEMS; FLIGHT CHARACTER-ISTICS; GYROSCOPE; SERVOMECHANISM.

Contrast with flight control. The control exercised by the stability augmentation system contrasts with that exercised by the pilot. The pilot may be connected with the flight-control surface via a direct mechanical link. Alternatively, in many modern airplanes the pilot cockpit control movement is sensed by a position transducer. The output of the position transducer in turn is sent, via a computer-amplifier combination, to a hydraulic actuator, referred to as a servo, which drives the flight-control surface (Fig. 2). Command signals which come from the pilot or from the stability augmentation system are sent by wire (fly-by-wire) or by optical conduit (flyby-light) to the electromagnetic valve. A valve distributes high-pressure hydraulic fluid either to the left or to the right of the piston so that the piston is forced to move. The piston in turn moves the flightcontrol surface. See FLIGHT CONTROLS.

Types of aircraft systems. When an airplane flies through turbulence, an important task is keeping the wings level. This becomes a tedious job for a pilot during long flights. For that reason, even many small airplanes are equipped with an automatic control device (called a wing leveler) which keeps the wings level.

Yaw dampers typically address lack of damping around the yaw axis (or z axis) of an airplane. Many airplanes also require improvement in damping around the pitch axis (or y axis). Some even require extra damping around the roll axis (or x axis). These dampers are referred to as pitch dampers and roll dampers, respectively. Several fighter and transport airplanes now have negative or low static longitudinal stability. Such problems can be addressed by an angle-of-attack or load-factor feedback system. The reliability of these stability augmentation systems and the associated computer hardware and software has improved to the point where only three such systems are required in commercial airplanes.

The early digital feedback systems still used fly-bywire systems. Such systems are susceptible to electromagnetic interference and therefore must be protected (that is, hardened) against such interference. Because hardened systems are both costly and heavy, more advanced digital feedback systems are fly-bylight systems.

With the introduction of fast in-flight digital computers, it has become possible to equip airplanes



Fig. 1. Block diagram for a yaw damper. The rudder deflects to oppose the error, and the airplane reacts with a different yaw rate.



Fig. 2. Electrohydraulically powered elevator system. Hydraulic lines to the actuator are not shown. The actuator is attached to a horizontal stabilizer structure.

with so-called full flight envelop protection systems. Such systems are designed to refuse any pilot input which might get the airplane into a flight condition from which recovery is no longer possible. Such systems can easily be arranged to prevent a pilot from rolling a commercial airplane too much or to prevent the pilot from stalling the airplane. Such systems can also be arranged so that loads acting on the wing or tail do not approach dangerously high levels. In that case the system is referred to as a load-alleviation system.

Flight simulations. A major design problem in all of these types of systems is assurance against unforeseen failures or unforeseen actions by the pilot. Very detailed failure cause-and-effect analyses are carried out during the detailed design and development stages of modern flight-control systems. However, it is not always possible to foresee every possible type of failure or unintentional wrong action by the cockpit crew. For that reason extensive so-called iron-bird simulations, with and without pilots in the loop, are carried out long before the first flight. During these simulations, all possible worst-case scenarios are, in effect, test-flown on the simulator, and the reaction of the system to these events in monitored. If necessary, design changes are introduced and tested. See AIRCRAFT TESTING; SIMULATION. Jan Roskam

Bibliography. J. H. Blakelock, Automatic Control of Aircraft and Missiles, 2d ed., 1991; D. McRuer, I. Ashkenas, and D. Graham, Aircraft Dynamics and Automatic Control, 1974, paper 1990; R. C. Nelson, Flight Stability and Automatic Control, 2d ed., 1997; J. Roskam, Airplane Flight Dynamics and Automatic Control, parts I and II, 1995.

Stabilizer (aircraft)

The horizontal or vertical aerodynamic wing surfaces that provide aircraft stability and longitudinal balance in flight. Horizontal and vertical stabilizers (fins) are similar to the aircraft wing in structural design and function of providing lift at angle of attack to the wind. However, stabilizers are not required to supply lift to overcome aircraft weight during flight, and when the wing-fuselage center of pressure is behind the aircraft center of gravity, the aerodynamic load on the horizontal stabilizer may be downward.

Stabilizer arrangements. Stabilizers may be swept back like wings for improved high Mach-number characteristics. The horizontal stabilizer may be found in a conventional or tail-last arrangement or in a canard or tail-first arrangement, or it may be dispensed with to give a tailless arrangement. Vertical stabilizers, however, are invariably present at the rear of the aircraft in single or multiple units. Aircraft designed for very high speeds may have auxiliary vertical stabilizers or ventral fins below the wing or fuselage to avoid the large losses in directional stability that occur with conventional arrangements at positive angles of attack at Mach numbers above about 2. Various stabilizer arrangements are shown in the **illustration**.

Stabilizing function. The stabilizing function of the horizontal and vertical stabilizers may be



Various stabilizer arrangements.

represented by the equation below, where $\partial C_m / \partial \alpha$

$$\frac{\partial C_m}{\partial \alpha} = -\left(\frac{\partial C_L}{\partial \alpha}\right) \frac{S'}{S} \frac{l'}{l} (1 - \epsilon_\alpha) \frac{q'}{q}$$

is the dimensionless rate of change of stabilizing or weathervane moment with angle of attack of sideslip, $\partial C_L / \partial \alpha$ is the dimensionless lift-curve slope of the isolated stabilizer based on stabilizer dimensions, and *S'/S* and *l'/l* are the ratios of stabilizer area to wing area, and of distance from the stabilizer center of pressure to the aircraft center of gravity to a characteristic length, respectively. The quantities ϵ_{α} and q'/q are flow conditions at the stabilizer and are the downwash or sidewash and dynamic pressure, respectively, referred to the free stream.

Control function. On aircraft provided with elevator and rudder control surfaces, the stabilizers may be either fixed to the fuselage or adjustable in incidence at a slow rate (about 1° /s) for trim. In either case, the stabilizers provide the support for the elevator and rudder, carrying their loads to the fuselage. Supersonic aircraft and guided missiles generally dispense with hinged trailing-edge elevator and rudder control surfaces, using the stabilizers for control as well as stabilization. In this case the stabilizers are actuated by the primary flight control system at rates up to about 50°/s. *See* FLIGHT CHARACTERISTICS; FLIGHT CONTROLS.

Bibliography. A. W. Babister, *Aircraft Dynamic Stability and Response*, 1980; B. Etkin and L. D. Reid, *Dynamics of Flight-Stability and Control*, 3d ed., 1995; R. C. Nelson, *Flight Stability and Automatic Control*, 2d ed., 1997.

Stabilizer (chemistry)

Any substance that tends to maintain the physical and chemical properties of a material. Degradation, that is, irreversible changes in chemical composition or structure, is responsible for the premature failure of materials. Stabilizers are used to extend the useful life of materials as well as to maintain their critical properties above the design specifications. Oxygen and water are the principal degradants, but ultraviolet radiation also can have a significant effect (photodegradation).

There are three general approaches to stabilization: surface modification, modification of chemical composition, and use of additives. The relative importance of these approaches varies with the material required to be protected and with the mode of degradation.

Surface modification. Wood and metals are protected against destructive weathering and corrosion by covering with paints or other surface finishes. Primers and rust-inhibiting paints are specialized formulations designed to provide corrosion resistance for metals. Aluminum is stabilized against oxidative degradation by a natural process in which an oxide film forms over the surface which then restricts oxygen penetration into subsurface layers. *See* COR-ROSION; PAINT. Laminated structures are also used to minimize degradation. For example, moisture-resistant polymers have been laminated over metals to provide corrosion resistance. In other instances, naturally resistant polymer films have been laminated over the surface of more sensitive polymers to stabilize the latter against photodegradation. Polymers have also been stabilized by crosslinking at the surface to restrict penetration of reactants into the bulk of the material.

The protection of food products by plastic wrappings is an extension of the approach to stabilization by surface modification. Plastic wrappings limit the amount of oxygen that reaches the food, thus inhibiting those oxidative reactions that lead to spoilage. *See* FOOD MANUFACTURING.

Modification of composition. Stainless steels are stabilized steels. Modifying the structure of steel is done by alloying with chromium or nickel corrosive chemical reactions. *See* STAINLESS STEEL.

Both natural and synthetic polymers degrade during processing and long-term exposure. Different polymers exhibit considerable variation in the level of stability to hostile environments, their stability being directly related to the dissociation energies of chemical bonds within the structure. When strong bonds are present, such as the C-F bonds in fluorinated polymers, a greater level of stability is observed in contrast to that of polymers in which the weaker C-H bonds predominate. This accounts for the superior thermal stability of tetrafluoroethylene over that of hydrocarbon polymers, which exhibit significant differences in the strength of their C—H bonds. The weakest C-H bonds in polypropylene occur at the tertiary carbon atoms where methyl groups are attached. Since one such bond is present in each repeating unit, polypropylene is much less stable to oxidative degradation than is linear polyethylene. The stability of hydrocarbon polymers could be modified by controlling the extent of branching. However, an improvement in stability by this approach will be accompanied by changes in other potentially important properties, for example, modulus.

There is evidence that some synthetic polymers have structural irregularities that could serve as sites for degradation to begin. Reduction of such irregularities during synthesis or replacement with less labile groups could also be used as an approach to stabilization. *See* POLYMER.

Additives. A wide variety of additives has been developed to stabilize polymers against degradation. Stabilizers are available that inhibit thermal oxidation, burning, photodegradation, and ozone deterioration of elastomers. Research in the chemistry of the low-temperature oxidation of natural rubber has revealed that hydrocarbon polymers oxidize by a freeradical chain mechanism. In pure, low-molecular-weight hydrocarbons, an added initiator is required to produce the first radicals. In contrast, initiation of polymer oxidation occurs in the complex molecules of elastomers through impurities already present, for example, hydroperoxides. *See* CHAIN REACTION (CHEMISTRY); FREE RADICAL.

Stabilization of hydrocarbon polymers can be accomplished with preventative or chain-breaking antioxidants. Preventative antioxidants stabilize by reducing the number of radicals formed in the initiation stage. Where hydroperoxides are responsible for initiation, the induced decomposition of these reactive intermediates into nonradical products provides effective stabilization. Suppressing the catalytic effects of metallic impurities that increase the rate of radical formation can also provide stabilization. Chain-breaking antioxidants interrupt the oxidative chain by providing labile hydrogens to compete with the polymer in reaction with the propagating radicals. The by-product of this reaction is a radical which is not capable of continuing the oxidative chain.

Stabilization of polymers against photooxidation, the principal component of outdoor weathering, is accomplished by addition of light screens, ultraviolet absorbers, or radical scavengers. Carbon black is the most effective of all light screens, functioning by preventing the penetration of ultraviolet radiation into the polymer. Correctly applied, this pigment can extend the outdoor life of hydrocarbon polymers from 1-2 years to over 50 years. Ultraviolet absorbers, which absorb and harmlessly dissipate damaging radiation, are not as effective as carbon black. However, another class of additives, known as hinderedamine light stabilizers, has become available. These stabilizers are believed to function by scavenging destructive radicals, and they may approximate the protection that is provided by carbon black. See AN-TIOXIDANT; CARBON BLACK; INHIBITOR (CHEMISTRY); PHOTODEGRADATION. W. Lincoln Hawkins

Bibliography. N. S. Allen and M. Edge, Fundamentals of Polymer Degradation and Stabilisation, 1993; V. A. Goldade et al., Plastics for Corrosion Inbibition, 2005; A. Hamidi, Handbook of Polymer Degradation, 2d ed., 2000; P. Marcus (ed.), Corrosion Mechanisms in Theory and Practice, 2d ed., 2002; Y. A. Shlyapnikov, S. G. Kiryushkin, and A. P. Mar'in, Antioxidative Stabilization Of Polymers, 1996; M. P. Stevens, Polymer Chemistry: An Introduction, 3d ed., 1998.

Stain (microbiology)

Any colored, organic compound, usually called dye, used to stain tissues, cells, cell components, or cell contents. The dye may be natural or synthetic. The object stained is called the substrate. The small size and transparency of microorganisms make them difficult to see even with the aid of a high-power microscope. Staining facilitates the observation of a substrate by introducing differences in optical density or in light absorption between the substrate and its surround or between different parts of the same substrate. In electron microscopy, and sometimes in light microscopy (as in the silver impregnation technique of staining flagella or capsules), staining is accomplished by depositing on the substrate ultraphotoscopic particles of a metal such as chromium or gold (the so-called shadowing process); or staining is done by treating the substrate with solutions of metallic compounds such as uranyl acetate or phosphotungstic acid. This type of staining is not dealt with in this article. *See* ELECTRON MICROSCOPE.

Stain classification. Stains may be classified according to their molecular structure. For example, there are the triphenylmethane dyes, such as the fuchsins and the methyl violets; the oxazine dyes, such as Nile blue; and the thiazine dyes, such as thionine and methylene blue.

The stains may also be classified according to their chemical behavior into acid, basic, neutral, and indifferent. This classification is of more practical value to the biologist. An acid dye, such as Congo red and eosine, is usually the sodium or potassium salt of a dye acid; a basic dye, such as methylene blue and Nile blue, is usually the chloride or sulfate of a dye base; a neutral dye, such as the eosinate of methylene blue, is a complex salt of a dye acid with a dye base; and an indifferent dye, such as Sudan III, is one with little chemical activity. At ordinary pH values, cells stain with basic dyes; at low pH values, they stain with acid dyes.

Staining procedures. The staining of microorganisms usually begins by making a smear. When it is desired to preserve intercellular relations, the smear may be replaced by a microculture. The smear or its substitute is usually fixed by heat and treated with a solution of some suitable biological stain as in simple stains or, in more elaborate procedures, it may be chemically fixed, treated with mordants, stained, partially decolorized, and counterstained. In certain procedures parts of a substrate are chemically removed prior to staining.

Smear preparation. The material containing the microorganisms is smeared or spread on the surface of a glass slide. Liquid material may be spread by means of a freshly flamed and cooled loop of platinum or some suitable alloy. Solid material may be smeared directly, or it may be used to make a suspension in water or some liquid medium, a loopful of which is spread on the slide. The film produced is allowed to dry in the air and is fixed by passing the slide, film side up, three or four times over the flame of a Bunsen burner. Some cytological procedures require chemical fixation. After cooling, the film is stained and examined.

Simple stains. These are procedures in which the smear, or the substrate, is stained with a dye solution for a given length of time, washed with water, air-dried, and examined. The most widely used solutions for simple staining are (1) Löffler's methylene blue, a solution of 0.3 g of methylene blue in 30 ml of 95% ethanol, mixed with 100 ml of 0.01% potassium hydroxide; (2) Ziehl-Neelsen's carbol fuchsin, a solution of 0.3 g basic fuchsin in 10 ml of 95% ethanol, mixed with a solution of 5 g phenol in 95 ml of distilled water; and (3) Hucker's crystal violet, prepared by mixing a solution of 0.2 g of crystal violet in 20 ml of 95% ethanol with a solution of 0.8 g ammonium oxalate in 80 ml of distilled water.

Complex staining procedures. Among complex staining procedures are the flagella and capsule stains. Both include the use of a special mordant to increase the density of their substrates and their affinity for the dye. This also increases the thickness of the flagella to a range that is resolved with the light microscope. There are many procedures for demonstrating both structures. In addition, most flagella stains recommend precautions to prevent destruction of these delicate structures while making the smear.

Nuclear stains. Most nuclear stains are based on the staining properties of nucleic acids. In many organisms, particularly among the bacteria, the cytoplasm is rich in ribonucleic acid. As a result, the entire cell, in young and mature cultures, tends to stain like a nucleus. Consequently, most nuclear stains include a step in which the ribonucleic acid is removed, or the bacteria are grown in media deficient in nitrogen supply to hinder the synthesis of ribonucleic acid.

Ribonucleic acid may be removed by treatment with the enzyme, ribonuclease, or a strong mineral acid. This is the basis of the widely used Feulgen reaction and of the HCl-Giemsa stain. In both methods the fixed smear is placed in normal HCl, usually at 60° C for about 10 min. In the Feulgen procedure this is followed by placing the smear for several hours in a solution of basic fuchsin that has been decolorized with sodium or potassium metabisulfite. The nuclei appear purple to violet in a colorless cell.

In the HCl-Giemsa procedure, the smear is stained with dilute Giemsa's stain. Stock solutions of this stain contain chiefly the eosinates of methylene azure and of methylene blue dissolved in a mixture of methyl alcohol and glycerol. It is a valuable cytological stain because of its metachromatism; it stains chromatin purple to red and the cytoplasm blue.

Differential stains. These are staining procedures that bring out color differences between a substrate and its background or between different parts of the same substrate. This may be accomplished by simple staining with a metachromatic dye; Nile blue, for example, stains the neutral fat droplets red and the rest of the cell blue. Or it may require controlled decolorization or differentiation with acids, alkalies, neutral solvents, or some other agents. Examples of the latter procedures are Gram's stain and acid-fast stain, and the spore stain.

Spore stains. These are based on the fact that the spore does not take up dyes readily, but once stained it resists decolorization. The differentiating agent used may be a dilute solution of an organic acid, an acid dye, or another basic dye.

Macchiavello stain. Another differential stain is that of Macchiavello for demonstrating rickettsiae. The fixed tissue smear is stained with basic fuchsin, differentiated with a 0.5% solution of citric acid, and counterstained with a 1% solution of methylene blue. Rickettsiae stain red, and tissue cells stain blue.

Negative stain. Another differential stain is the negative stain. In this procedure the cells appear colorless against a colored background. The cells are suspended in a droplet of the negative stain, such as India ink, water-soluble nigrosine, or Congo red in 1% aqueous solutions, on a glass slide. The droplet is spread out into a thin film and allowed to dry in the air. When pathogenic organisms are to be examined, the preparation is dipped in a mixture of 1 ml concentrated hydrochloric acid (HCl) and 100 ml of 95% ethanol. The acid alcohol changes Congo red preparations to blue, giving better contrast. Georges Knaysi

Fluorochrome stain. The fluorochrome staining procedure is used for the early and rapid detection of disease-producing Mycobacterium tuberculosis and M. leprae, together with other species of Mycobacteria and Nocardia asteroides. The method has been used worldwide since the mid-1970s. Widely recognized as the truant auramine-rhodamine (A-R) acidfast staining procedure, it not only permits the detection of tubercle bacilli more quickly in clinical specimens, such as sputa, urines, and tissues, but also gives a higher yield of positive findings than the conventional Ziehl-Neelson acid-fast procedure, which employs carbol fuchsin as the primary dye. The truant auramine-rhodamine procedure has been especially popular and relatively inexpensive because of the improvement in the development of fluorescent microscopes. The auramine and rhodamine dyes appear to have structures better suited for penetrating and stacking along the receptor macromolecules of the mycobacterial peptidolipid or peptidoglycolipid than does carbol fuchsin. The intracellular concentration possible with auramine and rhodamine may be greater than with carbol fuchsin, which may result in the superior staining qualities. The interpretation is that the cell wall lipids of mycobacterial cells in their native state containing mycolic acid react with the two fluorescing dye molecules of auramine and rhodamine and therefore absorb the stain. The dyes may well confer the property of acid fastness because they also function as a permeability barrier and impede the penetration of the bleaching agent or decolorizer. J. P. Truant

Bibliography. M. E. Boon and J. S. Drijver, *Routine Cytological Staining Techniques: Background and Practice*, 1986; G. Clark, *Staining Procedures*, 4th ed., 1980; P. R. Lewis and D. P. Knight, *Staining Methods for Sectioned Material*, 1977; P. Murray (ed.), *Manual of Clinical Microbiology*, 7th ed., 1999.

Stainless steel

The generic name commonly used for that entire group of iron-base alloys which exhibit phenomenal resistance to rusting and corrosion because of chromium content. The metallic element chromium (Cr) has been used in small amounts to strengthen steel since the famed Eads Bridge spanned the Mississippi at St. Louis, Missouri, in 1872; but only in the present century was it discovered that contents with chromium exceeding 10%, with carbon (C) held suitably low, make iron effectively rustproof. Onset



Fig. 1. Passivity phenomenon in stainless steel. $1200^{\circ}C = 2200^{\circ}F$.

of the property is striking (**Fig. 1**), and the chromium can then be increased to about 27% with even further chemical advantages before the metal becomes structurally useless. *See* ALLOY; STEEL.

Other alloy elements, notably nickel (Ni) and molybdenum (Mo), can also be added to the basic stainless composition to produce both variety and improvement of properties. Over 100 different stainless steels are produced commercially, about half as standardized grades. Some are more properly classed as stainless irons since they do not harden as steel; others are true steels to which corrosion resistance becomes an added feature. Still others that are neither properly steels nor irons introduce totally new classes of materials, from both mechanical and chemical standpoints.

In fact, the great modern developments of heatresisting superalloys and refractory metals stand as extreme extensions of the two heat-resisting families of stainless steels. That same resistance to rusting at ordinary temperatures shows as resistance to oxidation at elevated temperatures (Fig. 1), producing a family of oxidation-resistant high-temperature alloys; and with this basic feature contributed by the chromium content, further additions of nickel and other elements contribute remarkable properties of workability and strength, which extend not only to high temperatures but to cryogenic temperatures as well. The nickel-base heat-resisting alloys, the most prominent in that field, project from stainless steel in the sense that nickel increasingly replaces the iron (Fe). Similarly, the important titanium (Ti) and cobalt (Co) alloy systems, such refractory metals as niobium (Nb) and tungsten (W), and also chromium and molybdenum, represent the far extremes for use of these alloy elements in stainless steel.

The shaded band in Fig. 1 covers typical values for corrosion penetration of iron alloys with increasing chromium content, the actual data referring to hot nitric acid. The width of the band is due to differences in temperature and concentration. Attack almost ceases near 12% chromium, the alloy type 410; and it becomes scarcely measurable as chromium increases to type 446. The double-dash boundary in Fig. 1 is a plot of actual data for hot-gas scaling at $2190^{\circ}F$ ($1200^{\circ}C$). The sudden drop occurs in nearly the same chromium range as for aqueous corrosion, type 446 again being the most resistant alloy. The reversal at the far left is also similar for aqueous corrosion.

In service, the stainless steels cover three major fields: corrosion resistance at ordinary temperatures, scaling resistance at high temperatures, and mechanical strength over the entire temperature range of metallurgical engineering.

Types. Metallurgically, the alloys fall within four groups: ferritic, martinsitic, austenitic, and precipitation-hardening.

Ferritic. When chromium is added to iron to develop corrosion resistance without steel-type hardening, ferritic alloys are obtained. *See* IRON ALLOYS.

The typical alloys in the table are standards listed by the American Iron and Steel Institute (AISI). The first is the least expensive, and a little aluminum (AI) is usually added to help the chromium suppress steel-type hardening and ensure the ferritic condition, since this avoids welding problems from stray martensite. The second is the most popular commercial ferritic grade, widely used for automotive trim; and some fractional martensitic hardening is deliberately courted for additional strength. Type 446 is primarily used for oxidation resistance, for above this chromium content the engineering value of the chromium-iron alloys rapidly vanishes.

Martensitic. When chromium is added to steels with carefully chosen carbon contents, the martensitic alloys result (see **table**). Again, these cover a range from the least expensive and most widely used type

Typical compositions of standard stainless steels in percent				
AISI no.	Cr	С		
Ferritic				
405	13	0.08		
430	17	0.12		
446	26	0.20		
Martensitic				
410	12	0.12		
420	13	0.25		
440-C	17	1.00		
AISI no.	Cr	Ni	Other elements	
Austenitic				
302	18	8	C = 0.08	
310	25	20	C = 0.20	
316	16	13	Mo = 2.75	
347	18	11	Nb = 0.50	
Precipitation-hardening				
630 (17-4 PH)	17	4	Cu = 3.2, Nb = 0.25	
633 (AM-350)	16.5	4.3	Mo = 2.75, N = 0.10	
(174-4 Cu Mo)	17	14	Cu = 3.0, Mo = 2.5 Nb = 0.50, Ti = 0.25	

410, through the more selected type 420, ideal for cutlery, and on to the highly specialized type 440-C, whose great hardness makes it suitable for ball bearings and surgical instruments.

Austenitic. Nickel becomes the next most important alloving element. Its commonest use is in the renowned "18-8" containing 18% Cr and 8% Ni. Unlike the atomic structure of ferrite, which has the form of a neatly body-centered cube, or the martensite with its atomic disarrangement responsible for strength and hardness, the austenite of this third class of stainless steel has a face-centered cubic form made stable by the nickel. Completely new mechanical properties appear, outstanding for workability, notch toughness, high-temperature strength, and a frequently superior corrosion resistance. These are generally the most prized and popular of all stainless steels and also the most expensive. Only a few of the score of standards appear in the table. The parenthetical C for type 302 refers to a strange sensitivity to carbon content, such that 18-8 today is more commonly type 304 with carbon held below 0.08, and even type 304 with carbon held below 0.03. Type 310 lies well toward the nickel-base alloys and is accordingly notable for its heat resistance. The molybdenum in type 316 confers a special resistance to pitting corrosion, while the niobium in type 347 is a highly specialized addition correcting the carbon problem in welding 18-8.

While these alloys are called steels, no martensite develops on cooling; neither can the alloys be hardened by the usual quenching and tempering. Contrarily, they are put into their softest state by producing austenite at high temperature, with subsequent quenching to retain it. Strengthening is developed instead by cold working, and by adjusting the 18-8 composition slightly downward toward 17-7. The austenite then tends to decompose inside the cold steel; the martensite crowds the slip planes loosened by cold working, locking them against further movement and thereby increasing the strength (Fig. 2). Hence, ferritic and martensitic steels are little strengthened by cold working, while austenitic steels are greatly strengthened by it, notably type 301.



Fig. 2. Graph of the effect of cold-work on tensile strength of stainless steels. 1000 $\rm psi=6.9~MPa.$

Precipitation hardening. Exploiting this approach to strengthening, and following the example of duralumin in a neighboring metallurgical field, metallurgists have developed the fourth group of stainless steels known as precipitation hardening. A score of alloys became available within two decades of their commercial acceptance, and trade names rather than generic designations are still commonly used. The first listed in the table is subclassed martensitic because the metal behaves primarily as martensite which further experiences internal precipitation; the second is typical of what are called semiaustenitic alloys for the reason that they rather closely follow the instability pattern of 17-7; and the third belongs to the austenitic subgroup, with hardening essentially due to precipitation within austenite. The special properties are developed through carefully controlled heat treatments that always begin with a "solution anneal" at some high temperature and are followed by internal precipitation at some lower temperature. The latter may have multiple stages, ranging down to subzero treatments, and effects of cold work are often also exploited, particularly for the semiaustenitic alloys. These precipitation-hardening stainless steels are the hallmark since the third quarter of the twentieth century, bringing mechanical properties to levels unmatched by any other metal or alloy. Tensile strengths are actually exceeding the phenomenal figure of 300,000 psi (2.1 gigapascals).

Fabrication. As with steels in general, the so-called wrought stainless steels come from the melting furnaces in the form of either ingot or continuously cast slabs. Ingots require a roughing or primary hot working, which the other form commonly bypasses. All then go through fabricating and finishing operations such as welding, hot and cold forming, rolling, machining, spinning, and polishing. No stainless steel is excluded from any of the common industrial processes because of its special properties; yet all stainless steels require attention to certain modifications of technique.

Hot working. Hot working is influenced by the fact that many of the stainless steels are heat-resisting alloys. They are stronger at elevated temperatures than ordinary steel. Therefore they require greater roll and forge pressure, and perhaps less reductions per pass or per blow. The austenitic steels are particularly heat-resistant.

Welding. Welding is influenced by another aspect of high-temperature resistance of these metals—the resistance to scaling. Oxidation during service at high temperatures does not become catastrophic with stainless steel because the steel immediately forms a hard and protective scale. But this in turn means that welding must be conducted under conditions which protect the metal from such reactions with the environment. This can be done with specially prepared coatings on electrodes, under cover of fluxes, or in vacuum; the first two techniques are particularly prominent. Inert-gas shielding also characterizes widely used processes, among which at least a score are now numbered. As for weld cracking, care must be taken to prevent hydrogen absorption in the martensitic grades and martensite in the ferritic grades, whereas a small proportion of ferrite is almost a necessity in the austenitic grades. Metallurgical "phase balance" is an important aspect of welding the stainless steels because of these complications from a two-phase structure. Thus a minor austenite fraction in ferritic stainless can cause martensitic cracking, while a minor ferrite fraction in austenite can prevent hot cracking. However, the most dangerous aspect of welding austenitic stainless steel is the potential "sensitization" affecting subsequent corrosion, to be discussed later. *See* WELDING AND CUTTING OF MATERIALS.

Machining and forming. These processes adapt to all grades, with these major precautions: First, the stainless steels are generally stronger and tougher than carbon steel, such that more power and rigidity are needed in tooling. Second, the powerful workhardening effect (Fig. 2) gives the austenitic grades the property of being instantaneously strengthened upon the first touch of the tool or pass of the roll. Machine tools must therefore bite surely and securely, with care taken not to "ride" the piece. Difficult forming operations warrant careful attention to variations in grade that are available, also in heat treatment, for accomplishing end purposes without unnecessary work problems. Spinning, for example, has a type 305 modification of 18-8 which greatly favors the operation; while at the opposite extreme is type 301 or 17-7, for those who wish to take advantage of the strengthening due to cold working. See MACHINABIL-ITY OF METALS.

Finishing. These operations produce their best effects with stainless steels. No metal takes a more beautiful polish, and none holds it so long or so well. Stainlessness is not just skin-deep, but body-through. And, of course, coatings are rendered entirely unnecessary. At this point the higher initial cost of these alloys rapidly falls behind the long-range cost of other metals, often making them cheaper to buy on the market.

Service. Service of the stainless steels covers almost every facet of modern metallurgical engineering. They began in the chemical industry for reasons of their corrosion resistance, then graduated into high-temperature service for reasons of resistance to both surfacial sealing and internal creep. They have finally proliferated in every field of mechanical operation, ranging from high strength and hardness in combination with good ductility, at elevated as well as cryogenic temperatures, through cutlery and spring and nonmagnetic materials, to the purely ornamental applications of modern architecture.

However, when stainless steel does break down, the action is usually dramatic, and nothing exceeds the phenomenology of corrosion in the chemical industry. The situation can be understood by recognizing that the passivity of the stainless steels—that which prevents or restrains their reaction with environmental chemicals—is not the nonactivity or nobleness which characterizes the noble metals of the platinum group. It comes as a surprise to the nonscientists that the chromium content makes stainless steel even more disposed to rust than plain iron. But it is the very tendency which immediately fixes the oxygen or rust so tightly upon the surface that a protective coating results. The same quality applies to titanium and aluminum and their alloys. If something disturbs the quite special conditions needed for this passivity, corrosion takes place. The damage then usually takes one of the three major forms illustrated in **Fig. 3**.

Sensitization. Some phenomenon, not yet fully understood, causes chromium carbide to precipitate in the grain boundaries of the austenitic grades when heated even for short periods of time in the range 800-1500°F (427-816°C); these grain boundaries thereupon become so susceptible to chemical attack that a piece of formerly solid steel can fall apart in the hand like granulated sugar. The most workable explanation is that the carbon has depleted the chromium content in forming the carbide, thus leaving the metal locally nonstainless. Because welding necessarily produces a heat zone alongside the bead which somewhere must cover this critical range, welded steels often are susceptible to sensitization when later subjected to chemical service. The principal cures thus far are as follows: (1) The carbon is lowered to values where carbide cannot form; this has been accomplished with the introduction of type 304L. (2) Other and even more powerful carbide formers-notably titanium (Ti) and niobium (Nb)have been added to the alloys to tie up carbon at high temperatures before chromium carbide forms. Nevertheless, the problem persists as the principal concern in the chemical service of austenitic stainless steel.

Stress-corrosion cracking. Known for a hundred or more years in the brass industry as season cracking, the same phenomenon seems to have been discovered in stainless steel. In marked contrast to the intergranular nature of sensitization, stress-corrosion cracking cuts right through the body of the grain; and although the austenitic steels are by far the most susceptible, with the martensitic and ferritic following in that order, it appears that any metal or alloy can probably be made to fail in this manner if subjected to the right chemical under appropriate conditions of stress. Present corrections or preventions lie principally in directions of reducing stress levels, both internal and external, and avoiding exposure to chemicals known to be particularly dangerous. A close approach to immunity is indicated through the use of highly purifying melting processes in steelmaking

Pit and crevice corrosion. These are forms specifically following from oxygen "starvation" in local areas. When the steel can no longer rust, the passive condition converts to an active one; and to worsen the situation, all of the remaining surface then takes on the role of cathode in a galvanic cell which concentrates all the anodic or dissolving activity right at the tiny point of breakdown. If this is just a local spot, the attack takes the form of pitting, but if two flat



Fig. 3. Three great corrosion effects on stainless steel. (a) Sensitization. (b) Stress-corrosion cracking. (c) Pit corrosion.

surfaces come close together as in bolt-and-washer assemblies, the thin film of entrapped liquid can run out of oxygen and cause a general breakdown called crevice corrosion. *See* CORROSION.

Production. Stainless steel has undergone revolutionary changes in its melting technology. There are two major problems in the melting of stainless steel which distinguish it markedly from melting ordinary steel: chromium must be brought to high levels, and carbon levels must be held low. With minor exceptions in the high-carbon martensitic grades, stainless cannot be produced in ordinary fuel-fired furnaces such as the open hearth. The technology arose solely because of the closer controls possible with electrically heated furnaces, specifically the arc furnace where tonnage production is concerned. Thus stainless melting arose historically out of ordinary steelmaking by graduating into specialty furnaces. Following World War II the technological demands upon stainless steel pressed the technology into still greater specialization, such that steelmaking in this field displayed the most brilliant technical achievements in the history of metallurgy.

Ordinary steelmaking is concerned only with removal of carbon from the liquid metal, and this is easily accomplished by the addition of iron ore-the carbon seizes the oxygen and escapes as a gas. But chromium also reacts strongly with oxygen, thereby removing itself from the metal to join the slag. Therefore, this branch of steelmaking had to develop a twostage process: melt down ordinary steel and drive off the carbon, and add the chromium after the oxidizing of carbon was completed. This in turn brought two severe handicaps: scrap stainless steel-the cheapest chromium carrier-could not be used in the firststage charge because the chromium would be lost; and an expensive chromium-rich alloy with low carbon content would have to be added in the second stage-so-called low-carbon ferrochrome. See STEEL MANUFACTURE.

In the 1930s Alexander Feild developed the rustless process in Baltimore, based upon the discovery that the competition between chromium and carbon for oxygen shifts strongly in favor of carbon at very high temperatures. By using chromium oxide itself for the hearth lining and then carefully pressing the arc heating some hundreds of degrees above normal operating temperatures, Feild found that the first-stage charge could be loaded not only with scrap but also with high-carbon ferrochrome, which is much cheaper than the low-carbon grade. Minimal chromium losses to the slag were then "kicked back" with a shower of ferrosilicon, which easily took the oxygen away from the chromium and left silicates in the slag, while returning metallic chromium to the bath.

In the 1940s an abundance of liquid oxygen became available from developments in war industries, and E. J. Chelius developed the idea of the "oxygen lance." A long, insulated steel pipe was plunged beneath the surface of the molten steel, churning the metal with pure oxygen. The difference between iron oxide and gaseous oxygen, with respect to oxidizing a steel bath, is that the latter is a gas, and so is carbon oxide. The bubbling of the one aids the escape of the other, while to chromium oxide it makes no difference. Furthermore, the exceedingly high temperatures developed locally around the lance brought the advantage of the Feild effect in removing carbon while conserving chromium. The 1950s then witnessed a virtual conversion of the entire industry to the oxygen lance, spreading into the open hearth as well as the furnaces used for making stainless steel. At last all stainless could be made with high scrap additions in the charge, right up to 100%. As for carbon contamination and the use of high-carbon ferrochrome, the higher the first-stage carbon the better. It served as a fuel, and a convenient source of heat during lancing.

Scarcely had these achievements come into widespread use when successful development of a porous refractory plug permitted blowing a cleansing gas through the bottom of a ladle. At first this technique involved only the use of an inert gas, such as argon, for removing damaging quantities of hydrogen often dissolved in steels of all kinds. But shortly the method evolved to the use of oxygen as a substitute for lancing. However, the violence of the reaction, from entry so deep in the liquid, made it necessary to dilute the oxygen with an inert gas. Again argon was chosen, and the argon-oxygen process was born. New achievements were electrifying: (1) Almost the entire electric furnace processing was dispensed with, the furnace being needed only to melt down the charge, with all further work done in the ladle. The expensive furnace time was cut by factors of three and four, or more. (2) The vastly greater facilities for quality control in a ladle, as compared to a furnace, shortly permitted the further bridging of this technology with the long-developed and highly sophisticated vacuum techniques of physics and laboratory furnaces. The argon-oxygen process then developed variations using vacuum chambers and reduced pressures instead of positive pressures, and in some cases substituting nitrogen or other gases for argon.

Concurrently two major developments entered the field, this time from directions of those highly specialized heat-resistant metals and alloys known as the superalloys and the refractory metals. The superalloys generally comprise systems based upon nickel, cobalt, or titanium, as well as highly complex stainless steels, while the term "refractory metals" typically refers to wolfram (tungsten), tantalum, molybdenum, and niobium (columbium). Frequently the superalloys have quality demands so high as to require remelting a cast form in order to purify it a second time. One process arose naturally out of the vacuum technology just mentioned, and the more critical grades of stainless steel are now often subjected to vacuum-arc remelting. The original melt is cast in the form of a huge electrode which, as with a consumable welding electrode, melts off at the tip and falls into a water-cooled mold in a vacuum chamber. The metal is thus purified, and its solidification structure is very favorable to good mechanical properties. The second process also arose out of a welding concept, namely the use of a slag cover to protect the metal from environmental reactions while molten. First conceived by R. E. Hopkins in America several decades ago, the process was developed in the Soviet Union, then captured worldwide interest in the 1960s. Again the metal in electrode form is remelted electrically, but under a carefully prepared slag blanket instead of a vacuum. Solidification is again well controlled, and the metal is highly purified. *See* HIGH-TEMPERATURE MATERIALS.

Technological advances continued at a rapid rate, with such sophisticated developments as melting with electron beams, plasma arcs, and lasers—even levitation melting, where crucibles or containers are no longer necessary. The metal just suspends itself in mid-air! While levitation melting remains in curiosity stages and lasers still serve principally for punching holes and welding, the plasma-arc technique did enter commercial usage in welding, with beginnings in steelmaking. The electron-beam furnace has led to remarkable accomplishments in producing commercial quantities of certain ferritic stainless steels whose properties were previously believed impossible.

Most stainless is cast as ingots, although continuous casting is rapidly becoming popular. Virtually all wrought grades have their casting counterparts in the foundry industry, separately standardized by the Alloy Casting Institute (ACI). The metallurgy of the cast alloys is essentially the same as that for the wrought, with the principal exception that certain liberties can be taken in alloy composition because the castings are not hot-worked. In designing casting alloys, one can forget about mechanical problems of forming the part and concentrate on the chemical challenges of corrosion resistance and heat resistance. *See* METAL CASTING. Carl A. Zapffe

Bibliography. American Society for Metals, *Metals Handbook, Properties and Selection: Irons and High Performance Alloys*, 10th ed., vol. 1, 1990; J. R. Davis (ed.), *Stainless Steels*, 1994; R. A. Lula, *Stainless Steel*, 1985; F. B. Pickering (ed.), *The Metallurgical Evolution of Stainless Steels*, 1979; L. L. Shreir (ed.), *Corrosion*, vols. 1 and 2, 3d ed., 1994.

Stalactites and stalagmites

Stalactites, stalagmites, dripstone, and flowstone are travertine deposits in limestone caverns (see **illus.**), formed by the evaporation of waters bearing calcium carbonate. Stalactites grow down from the roofs of caves and tend to be long and thin, with hollow cores. The water moves down the core and precipitates at the bottom, slowly extending the length while keeping the core open for more water to move down. Stalactites are banded concentrically to the center.

Stalagmites grow from the floor up and are commonly found beneath stalactites; they are formed from the evaporation of the same drip of water that forms the stalactite. Stalagmites are thicker and



Stalactite and stalagmite deposits. Along the central opening are deposits made by water as it dripped from the crack in the limestone. Outside layers were deposited by water flowing over the surface. Water may drip to the floor and create a stalagmite. (After E. B. Branson et al., Introduction to Geology, 3d ed., McGraw-Hill, 1952)

shorter than stalactites and have no central hollow core. The banding of stalagmites is parallel to the surface at the time of deposition of each band.

Dripstone and flowstone are travertine accumulations of much the same origin as stalactites and stalagmites (evaporation from cave waters) but are irregular. Their irregularity is due to deposition on irregular sloping walls or surfaces in the cave rather than drips from the roof. *See* CALCITE; CAVE; LIME-STONE; TRAVERTINE. Raymond Siever

Bibliography. D. Gillieson, *Caves: Processes, Development, Management*, 1997; C. A. Hill and P. Forti, *Cave Minerals of the World*, 2d ed., 1997.

Stall-warning indicator

A device that determines the critical angle of attack for a given aircraft, at which point the lift stops increasing and the aircraft will no longer sustain itself in steady-state condition (level flight or climb/descent). The indicator usually operates from vane sensors, airflow pressure sensors, tabs on the leading edge of the wings, and computing devices which include accelerometers, airspeed detectors, and vertical gyros. The indicator may be a pointer over a dial, lights, or an audible signal. In some cases, the computer may actuate a control-column shaker to alert the pilot. *See* AIRCRAFT INSTRUMENTATION; AIRPLANE; AUTOPILOT. James W. Angus

Standard (physical measurement)

An accepted reference sample which is used for establishing a unit for the measurement of physical quantities. A physical quantity is specified by a numerical factor and a unit; for example, a mass might be expressed as 8 g, a length as 6 cm, and a time interval as 2 min. Here the gram is a mass unit defined in terms of the international kilogram, which serves as the primary standard of mass. The centimeter is defined in terms of the international meter, which is the primary standard of length and is defined as the length of path traveled by light in a vacuum during a time interval of 1/299 792 458 of a second. In similar fashion, the minute is a time interval defined as 60 s, where the second is the international standard of time and is defined as the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine energy levels of the ground state of the cesium-133 atom.

The National Institute of Standards and Technology in the United States and comparable laboratories in other countries are responsible for maintaining accurate secondary standards for various physical quantities. *See* ELECTRICAL UNITS AND STAN-DARDS; LIGHT; METRIC SYSTEM; PHYSICAL MEASURE-MENT; TIME. Dudley Williams

Standard model

The theory that explains the three major interactions of elementary particle physics—the strong interaction responsible for nuclear forces, the weak interaction responsible for radioactive decay, and the electromagnetic interaction—in terms of a common physical picture. The model for this picture is quantum electrodynamics, the fundamental theory underlying electromagnetism. In that theory, electrons, viewed as structureless elementary constituents of matter, interact with photons, structureless elementary particles of light. The standard model extends quantum electrodynamics to explain all three interactions of subnuclear physics in terms of similar basic constituents.

Quantum electrodynamics (QED). In the 1940s, Richard P. Feynman, Julian S. Schwinger, and Shin-Ichiro Tomonaga constructed quantum electrodynamics, the quantum theory of the interactions of electrons with electromagnetic fields. They represented the electron as a pointlike particle, possessing no size but only mass and intrinsic angular momentum, or spin. Similarly, they represented electromagnetic waves as composed of pointlike quanta, photons, possessing spin but no mass. They assigned these particles the simplest equations of motion consistent with Einstein's special relativity, and postulated also the simplest relativistic coupling which allowed electrons to emit and absorb photons one at a time. This showed that these ingredients reproduce the observed properties of electromagnetism and its interaction with electrons. The experimental validation of this theory has only improved with time.

Computations in quantum electrodynamics of the intrinsic magnetic field of an electron agree with experiment to the level of parts per trillion. *See* ELECTRON; LIGHT; PHOTON; QUANTUM ELECTRODYNAMICS.

Spin. It may seem strange that a particle of zero size can have intrinsic spin. This is one of the unexpected and profound consequences of combining relativity theory and quantum mechanics. In quantum mechanics, angular momentum associated with motion takes only the discrete values which are multiples of \hbar , Planck's constant divided by 2π . Relativistic particles can also have intrinsic angular momentum, and this may come in either units or halfunits of \hbar . In quantum theory, a relativistic particle is viewed as a quantum of the quantized oscillations of fields extending throughout space. For each possible value of the spin, there is a unique simplest form for the equations of propagation of the field. The very simplest, the Klein-Gordon equation, is associated with a spinless (spin-0) particle. For a particle with intrinsic angular momentum of one unit of \hbar (spin 1), the required propagation equations are precisely Maxwell's equations for electromagnetic waves. For a particle with a spin of half a unit of \hbar (spin $\frac{1}{2}$), the characteristic equation is called the Dirac equation. Particles whose spin is an integer multiple of \hbar are called bosons; particles with an extra half-unit of spin are called fermions. It is a consequence of relativistic quantum mechanics that multiple bosons can occupy the same location. When present in great numbers, these particles can form macroscopic fields, such as the electric field. On the other hand, fermions in the same spin state are forbidden from occupying the same location in space. This property accounts for the Pauli exclusion principle, which leads to the rigidity of atoms and of atomic nuclei. It is thus natural to build up matter from elementary fermions, and to represent fields of force by elementary bosons. In quantum electrodynamics, the electrons have spin $\frac{1}{2}$ and the photons have spin 1. See BOSE-EINSTEIN STATISTICS; EXCLU-SION PRINCIPLE; FERMI-DIRAC STATISTICS; MAXWELL'S EQUATIONS.

The spin also affects the ability of a particle to obtain mass. For spin-0 particles, there is a zero-mass Klein-Gordon equation, but one may add mass simply by adding one extra term to this equation. For spin ¹/₂, the simplest Dirac equation involves two massless states; a particle which spins in the left-handed sense and an antiparticle which spins in the righthanded sense. The electron of quantum electrodynamics is described by putting together two of these pairs of states, to make up the left- and right-handed quantum states of the electron and the left- and righthanded states of the antielectron or positron. Only when the two halves are present is it possible to put a mass into the equation of motion. As an outcome, the electron and positron are predicted to have equal mass and opposite electric charge. (These relations are confirmed experimentally to an accuracy of 40 parts per billion.) The equation of motion for spin-1 particles does not allow a mass to be added at all, except under a special circumstance, explained later, in which the spin-1 particle couples to a spin-0 particle or field. This barrier can explain why the photon has zero mass. (Experimentally, the photon is observed to have mass less than 10^{-22} times the electron mass.) *See* RELATIVISTIC QUANTUM THEORY; SPIN (QUANTUM MECHANICS); WAVE EQUATION.

Quarks and leptons. The subnuclear world contains many more constituents than electrons and photons. First of all, two heavier spin-1/2 particles have been observed which have the same electromagnetic charge as the electron (*e*) but larger mass. These particles, the μ and τ , are unstable; they decay to the electron with the simultaneous emission of very light, uncharged spin-1/2 particles called neutrinos. Experiments have demonstrated that there is one neutrino species for each of the three charged particles *e*, μ , and τ . The three charged and three neutral fermions and their antiparticles are collectively called leptons. *See* LEPTON; NEUTRINO.

Next, experiments have observed a very large number of particles which experience the strong interactions. These particles, collectively called hadrons, include the proton and neutron, the π and K mesons, and many short-lived heavier species. It has been established that the hadrons are composite systems built from more elementary fermions, called quarks. The known list of hadrons can be built up from six species of quarks: u (up), d (down), s (strange), c (charm), b (bottom), and t (top). (The t quark was discovered in 1995.) For example, the proton is a bound state of three quarks (uud); the π^+ meson is a bound state of a quark and an antiquark (u-d). To give the correct electric charge assignments for the hadrons, the quarks must be assigned fractional electric charges, $^{2}/_{3}$ or $^{-1}/_{3}$. In addition, each species of quark must have three possible states of an invisible quantum number, called color. Hadrons built of three quarks, called baryons, must contain one quark of each color. With this rule, the bound states of three quarks consistent with the Pauli exclusion principle give precisely the observed set of baryons. See BARYON; COLOR (QUANTUM ME-CHANICS); HADRON; MESON; QUARKS.

Experiments at very high energy have confirmed the quark picture of hadrons. The pattern of scattering of high-energy electrons and neutrinos from protons confirms that quarks have fractional electric charge and spin $1/_2$. When electrons and positrons collide at high energy and annihilate, the resulting burst of energy produces hadrons at a rate three times larger than the simple prediction from electrodynamics, indicating the presence of the three color states. The hadrons form two back-to-back jets of particles, following the directions that would be preferred by spin- $1/_2$ quarks and antiquarks.

In the standard model, all matter is built from the quarks and the leptons. These 12 fermions organize themselves into three families, each containing the same set of electric charges. **Table 1** shows this organization and gives the mass of each particle. The masses of the u, d, and s quarks are quite uncertain; because most of the rest energy of a proton, for example, comes from the strong-interaction binding

TAE	3LE 1. Elementary matte	r particles of th	e standard model*			
		Leptons				
	Quarks			Charge - 0		
0	Charge = $\frac{2}{3}$	Char	$ge = -\frac{1}{3}$	(neutrinos)	Cha	rge = -1
и	3	d	5	ν_{e}	е	0.5
С	1200	S	100	$ u_{\mu}$	μ	105.7
t	174,000	b	4200	$\nu_{ au}$	au	1777.0
*T neut	he number next to each char rinos have very small masses	ged particle shows	s its observed mass, in	MeV. (The scale may be referred to the p	roton mass: $m_p = 93$	38 MeV.) The

rather than from the intrinsic masses of the quarks, these values must be inferred indirectly. Nevertheless, there is a clear hierarchical pattern to the spectrum of masses. The origin of this pattern is still a mystery.

Symmetry groups and quantum states. To discuss further the relations between the various quarks and leptons, it is necessary to review how symmetry operations are treated in quantum mechanics. Rotations in space provide a convenient example. A classical arrow can point in any direction; a classical rotation moves the arrow from one definite direction to another. In quantum mechanics, the Heisenberg uncertainty principle forbids one from specifying the orientation of an arrow. Instead, one starts with a fixed number of quantum states corresponding roughly to different orientations. The number of these states depends on the angular momentum. For example, a system with angular momentum 1, in units of \hbar , has three possible quantum states; a system with spin $1/_2$ has two. If the system starts in one of these states and is rotated, the result is a weighted sum of the possible states. By rotating far enough, it is possible to convert the original state completely into one of the others. Conversely, a rotation in space can be thought of as built up out of three basic operations: small rotations about the x, y, and z axes (Fig. 1). A continuous family of symmetry operations built up in this way is called a Lie group. The basic operations are called the generators of the group. The mathematical notion of a Lie group was actually invented in the nineteenth



Fig. 1. Basic rotations in three-dimensional space.

century. After the discovery of quantum mechanics, Hermann Weyl and Eugene P. Wigner found that this description of symmetries is the one that fits naturally into quantum theory. *See* GROUP THEORY; LIE GROUP; SYMMETRY LAWS (PHYSICS).

To understand the symmetry structure of the standard model, imagine setting all of the quark and lepton masses equal to zero. Then, as far as the information in Table 1 is concerned, the three families of quarks and leptons would be exactly equivalent. In this limit, there are four types of symmetry operations that could be applied to the states in the table. The most obvious, called a flavor rotation, transforms or rotates the states of one family into those of another. Another possible operation transforms the leptons into the quarks of one particular color. A third, called a color rotation, transforms the quarks of one color state to those of another. A fourth, called an isospin rotation, transforms the two quarks of each family into one another, and, similarly, the two leptons. These last two operations are especially important for the structure of the standard model. These operations have a mathematical structure equivalent to that of the group of rotations in a space with, respectively, three and two complex-valued coordinates. Mathematicians refer to these rotation groups as SU(3) and SU(2). The two groups are generated, respectively, from eight and three basic operations, in the sense discussed at the end of the previous paragraph.

Yang-Mills theory. In quantum electrodynamics, photons do not interact with one another, except indirectly through their coupling to matter. It can be shown that relativity and quantum mechanics put severe constraints on the possible interactions of spin-1 bosons. The unique consistent prescription was discovered by Chen Ning Yang and Robert L. Mills in 1954. To construct a Yang-Mills theory, one begins with a Lie group and assigns a spin-1 particle to each generator of the group. Matter particles are assigned to quantum states which are rotated into one another by the group operations. Then the spin-1 bosons are coupled to matter in such a way that the emission or absorption of such a boson converts one matter state into another according to the group transformation law. This prescription also causes the bosons to be coupled to one another: For rotations in space, as can be seen from Fig. 1, a rotation about x causes a rotation about y to be transformed into a rotation about z. Then the spin-1 boson associated

with x rotations can absorb a spin-1 boson associated with y and transform it into the spin-1 boson associated with z. The probability for boson absorption or emission is governed by a number g called the coupling constant, which generalizes the electric charge of electrodynamics. In fact, electrodynamics can be considered as a particularly simple case of Yang-Mills theory in which the Lie group is the group of rotations of a circle, which has only one generator; mathematicians call this group U(1). In the 1970s, Gerardus 't Hooft and Martinus J. G. Veltman, Andrei A. Slavnov, John C. Taylor, Benjamin Lee, and others showed that the Yang-Mills prescription gives a well-defined quantum theory of spin-1 bosons, and that prescriptions which lack some elements of this structure lead to inconsistencies. The spin-1 bosons which play the central role in Yang-Mills theory are called gauge bosons. See GAUGE THEORY.

The standard model can now be defined as the theory of elementary particles in which matter is constructed from the spin-1/2 fermions in Table 1 and the interactions of these particles are generated from a Yang-Mills theory based on the product of three independent groups of symmetry operations, $SU(3) \times$ $SU(2) \times U(1)$. The SU(3) group is taken to be the group of color rotations, as defined at the end of the previous section. The SU(2) group is taken to be the group of isospin rotations, but acting only on the left-handed quark and lepton states. The U(1) factor refers to one extra gauge boson similar to the photon of quantum electrodynamics. Using the mathematical structure described by Yang and Mills, this prescription leads to a compact set of equations describing massless fermions and massless spin-1 bosons. Table 2 lists, on the left, the three component symmetry groups of the standard model and a measure of their intrinsic coupling strengths; on the right, the three corresponding interactions of subnuclear physics, with their relative strengths and the spin-1 gauge bosons that mediate each.

Quantum chromodynamics (QCD). At this stage, it may seem that the standard model is an elegant physical theory, but it is not apparent that the model has anything to do with nature. The strong and weak interactions are both short-range; in quantum theory, this implies that they are mediated by massive, not massless, bosons. And a symmetry which acts only on left-handed and not right-handed fermions precludes combining these states to form the observed massive quarks and leptons. To show that this mathematical theory does describe nature requires new physical mechanisms.

The strong interactions will be considered first. In the standard model, these interactions are described by an SU(3) Yang-Mills theory acting on color, called quantum chromodynamics. The eight gauge bosons are called gluons. The most important qualitative feature of the strong interactions is that the observed physical particles, the hadrons, are bound states of quarks, while the individual quarks are not observed. Remarkably, this feature can be derived from quantum chromodynamics when its quantum-mechanical aspects are fully taken into account.

In quantum theory, the size of charges or coupling constants is not fixed but can vary as a function of the energy scale. In quantum electrodynamics, the electric charge e is fixed at macroscopic distances but becomes larger at very small distances. This is the result of an effect called vacuum polarization. Since particle-antiparticle pairs can be created by quantum processes, such pairs are constantly being created and destroyed, everywhere in space, by quantummechanical fluctuations. If there is an electric field present, the electrons and positrons created by these fluctuations will move to partially screen out the field. Then the field observed at larger distances will be smaller. An electron-positron pair in such a fluctuation can have any separation up to the electron Compton wavelength, 4×10^{-12} m. If two charged objects are separated by a distance smaller than this, they will be inside the fluctuating cloud of particles and antiparticles and will feel a stronger electric interaction. The strength of the electromagnetic force can be expressed by the dimensionless ratio called the fine structure constant, $\alpha = e^2/\hbar c$, where c is the speed of light. At macroscopic distances, $\alpha =$ 1/137, but in experiments in which high energies are used to bring electrons and positrons to within distances less than 10^{-18} m, the value $\alpha = 1/129$ is observed.

In Yang-Mills theories in which the gauge bosons interact directly, there is a compensatory quantum effect in the opposite direction. Through this effect, the interaction of a gauge electric field with a quantum fluctuation can reinforce its strength. This effect, called asymptotic freedom because it implies that the coupling constant is smaller at short distances, was discovered in 1973 by David Politzer, David J. Gross, and Frank Wilczek. In quantum chromodynamics, this effect overwhelms the vacuum polarization effect and makes the coupling grow dramatically as

TABLE 2. Elementary interactions of standard model					
Symmetry group	Coupling strength $(\alpha)^*$	Bosons	Interaction	Strength	
SU(3) SU(2) U(1)	1/8.4 1/29.6 1/98.4	Gluons Photon W ⁺ , W ^{-,} , Z ⁰	Strong Electromagnetic [†] Weak ^{†,‡}	1 10 ⁻² 10 ⁻⁵	
*Coupling strengths are $\alpha = g^2/c$. [†] Weak and electromagr [‡] The small strength of	*Coupling strengths are given at a high energy, equal to the rest energy of the Z^0 boson, 91,000 MeV, and are written as pure numbers by defining $x = g^2/c$. [†] Weak and electromagnetic interactions are combinations of the basic SU(2) and U(1) interactions, as explained in text. [‡] The small strength of the weak interactions is due to the large masses of the W and Z bosons.				

objects with color are separated. States with widely separated quarks or gluons acquire very high energy, so that it is always energetically favorable to create additional quark-antiquark pairs to neutralize the color charges. Then these configurations decay rapidly to quark and antiquark bound states. The separation at which the quantum chromodynamics coupling becomes strong, which sets the size of typical hadrons, is about 10^{-15} m. This corresponds to a binding energy of roughly m_pc^2 , where m_p is the proton mass. The massive bosons which mediate the nuclear force are then quark-antiquark bound states, mainly π and ρ mesons, which acquire their mass from their internal quantum chromodynamics energy.

Asymptotic freedom implies that experiments at energies well above the proton rest energy should be able to see quarks and gluons. At the end of the discussion of quarks, it was noted that the hadrons resulting from the annihilation of electrons and positrons at high energy form pairs of collimated jets that follow the directions expected for a quarkantiquark pair. The detailed properties of a jet indicate that it originates as a single quark or antiquark and acquires its complexity from the production of additional pairs along the gauge electric field direction. In 1979, experiments at the German accelerator PETRA observed events with three jets of hadrons, with the distribution of the third jet in energy and angle very similar to that expected for radiation of photons from the produced quark and antiquark. These additional jets originated as radiated gluons. See GLUONS; QUANTUM CHROMODYNAMICS.

Higgs mechanism. The massive particles which mediate the weak interactions appear experimentally to be spin-1 bosons with the simple couplings of Yang-Mills gauge bosons. There are three of these massive particles, called W^+ , W^- , and Z^0 . Given the fundamental difficulties noted above in writing a mass term for spin-1 particle, it is not clear how these states could arise from a Yang-Mills theory. There is one way known in which Yang-Mills bosons can acquire mass, discovered in the 1960s by Peter W. Higgs and a number of others. Though in a Yang-Mills theory the basic equations must be symmetric, it is possible that nature realizes these equations in an asymmetrical way. For example, with respect to rotations in a three-dimensional space (Fig. 1), if we lived in a medium which randomly chose a preferred direction in this space, say, the zdirection, only the rotation about z would be recognizable as a symmetry. The other two rotation symmetries are said to be spontaneously broken. Such spontaneous symmetry breaking occurs in magnetic materials when they choose an axis of magnetization, and in crystals when they choose axes of alignment.

If the symmetries that are spontaneously broken are those of a Yang-Mills theory, the corresponding gauge bosons become massive. For example, if some field took up an orientation that preferred the z axis in Fig. 1, the gauge bosons corresponding to x and yrotations would become massive, while the one associated with z rotations would remain massless. It is possible to turn this picture into a candidate for a unified model of weak and electromagnetic interactions, in which the two massive bosons are the W^+ and W^- , and the massless spin-1 boson is interpreted as the photon. Experiment chooses a different theory built up around the same idea, constructed in the 1960s by Sheldon L. Glashow, Steven Weinberg, and Abdus Salam (GWS). This theory makes use of the SU(2) and U(1) pieces of the standard model Yang-Mills symmetry. This symmetry group has four generators in all. Three of these are spontaneously broken and one remains unbroken. This gives the massive W^+ , W^- , and Z^0 , plus the massless photon. The field that is responsible for the symmetry breaking is still a complete mystery. In the simplest model, the symmetry breaking can be due to one spin-0 field, called the Higgs field; the quanta of this field are called Higgs bosons. In this model, the Higgs field also is responsible for coupling the left- and right-handed components of quarks and leptons and thus for generating the fermion masses. See HIGGS BOSON; SYM-METRY BREAKING.

Electroweak interactions. The Glashow-Weinberg-Salam theory describes electroweak interactions. This theory contains and relates three different types of interactions: electromagnetism, mediated by the photon; charge-changing processes, mediated by the *W* bosons; and the neutral-current interaction, mediated by the *Z* boson. The latter two processes involve couplings that are stronger than those of electromagnetism, but they appear weak in nuclear physics because the *W* and *Z* have large masses, 80,4000 and 91,188 MeV, respectively, in units in which the proton mass is 938 MeV.

The charge-changing processes include the radioactive beta-particle decay of nuclei, which corresponds in this picture to a *d* quark inside a neutron changing to a *u* quark and a quantum of the W^- field, which materializes as an electron and an electron antineutrino. The fundamental coupling involves only left-handed species, and this fact leads to predictions for the energy and angular distributions of the decay products which are in good accord with experiment. Similar predictions can be made for the products of more exotic reactions such as μ and τ decay and the scattering of high-energy neutrinos, and again experiments confirm the simple picture in which only the left-handed fermion states couple to the *W* bosons. *See* RADIOACTIVITY.

The neutral-current interaction, a short-ranged scattering interaction between fermions which supplements the electromagnetic interaction, is a characteristic feature of the Glashow-Weinberg-Salam theory. The charges by which the various fermions couple to this interaction, or to the *Z* boson, are given by the electric charges combined with an extra factor for left-handed species. The neutral-current interaction was first observed in the 1970s in experiments on the scattering of high-energy neutrinos. Accelerators at the Stanford Linear Accelerator Center (SLAC) in California and at CERN, in Geneva, Switzerland, have produced individual *Z* bosons by the annihilation of electrons and positrons and have observed



Fig. 2. Measurements of the rate for electron-positron annihilation into a Z^0 boson as a function of the collision energy, from the data of the OPAL experiment at CERN.

the decay of the Z to all of the fermion species of the standard model (except for the very heavy t quark). The results confirm the Glashow-Weinberg-Salam charge assignments in detail. One piece of this confirmation begins from the fact that in quantum mechanics an unstable state with lifetime τ has an uncertainty in energy given by \hbar/τ . Then a reaction which produces the unstable state proceeds over a range of energies, with the rate varying in a characteristic shape, called a resonance. Figure 2 shows the rate for electron-positron annihilation to form a Z boson, as measured by the OPAL experiment at CERN. The width of the resonance measures the lifetime of the Z. The agreement shown between theory and experiment actually tests all three of the interactions of the standard model-the weak interactions, because the decay rate of the Z depends on the coupling to each fermion species; the strong interactions, because the decay rate to quarks is enhanced by the color factor 3 and another 4% correction due to gluon effects; and the electromagnetic interactions, because the radiation of photons from the incident electrons and positrons produces the asymmetrical shape of the resonance curve. In this and other comparisons, the Z decay rate agrees with the prediction of the standard model to an accuracy of 0.1%. See ELECTROWEAK INTERACTION; INTERME-DIATE VECTOR BOSON; NEUTRAL CURRENTS; WEAK NU-CLEAR INTERACTIONS.

Open questions. The standard model brings together diverse aspects of the behavior of subnuclear particles in a common framework. We can truly claim to understand the nature of the strong, weak, and electromagnetic interactions. The explanation, though, sets up a number of new questions. The first among these is the nature of the agent that is responsible for spontaneously breaking SU(2) \times U(1) symmetry. Is it the Higgs field described above or a more complex set of interactions? Whatever this agent is, it is responsible for the masses of the heavy spin-1 bosons W and Z, and for all of the quark and lepton masses. A related question is the origin of the pattern of fermion masses shown in Table 1. The origin of the Yang-Mills symmetry group SU(3) \times $SU(2) \times U(1)$ is also a mystery. In 1973, Howard Georgi and Glashow and Jogesh C. Pati and Salam proposed that this group orginates as the set of unbroken symmetries that remain when a larger symmetry group, called the grand unification symmetry, is spontaneously broken at very short distances. The choice of SU(5) for the grand symmetry group explains the pattern of electric charge assignments of the quarks and leptons. New measurements of the standard model coupling constants have turned out to be inconsistent with the simplest form of the grand unification hypothesis, but they fit the expectation if the assumption is added of an enhanced spacetime symmetry called supersymmetry, which interconnects bosons and fermions. See GRAND UNIFICA-TION THEORIES; SUPERSYMMETRY.

At least a part of the answers to these questions should be found at the next generation of particle accelerators. If there is a Higgs boson, its mass should not be far above the masses of the *W* and *Z*. If grand unification is correct, new particles predicted by supersymmetry, or another principle which extends the standard model, should also appear. With luck, these experiments should lead to an even simpler and more all-encompassing mathematical description of nature. *See* PARTICLE ACCELER-ATOR. Michael E. Peskin

Bibliography. G. D. Coughlan, J. E. Dodd, and B. M. Gripaios, *The Ideas of Particle Physics*, 3d ed., 2006; R. P. Crease and C. C. Mann, *The Second Creation*, rev. ed., 1996; J. F. Donoghue, E. Golowich, and B. R. Holstein, *Dynamics of the Standard Model*, 1995; D. Griffths, *Introduction to Elementary Particles*, 1987; F. Halzen and A. D. Martin, *Quarks and Leptons*, 1984; G. Kane, *Modern Elementary Particle Physics*, 2d. ed., 1993; G. Kane, *The Particle Garden*, 1995; M. E. Peskin and D. V. Schroeder, *An Introduction to Quantum Field Theory*, 1995.

Staphylococcus

A genus of bacteria containing at least 28 species that are collectively referred to as staphylococci. Their usual habitat is animal skin and mucosal surfaces. Although the genus is known for the ability of some species to cause infectious diseases, many species rarely cause infections. Pathogenic staphylococci are usually opportunists and cause illness only in compromised hosts. *Staphylococcus epidermidis* is a major component of the human skin flora. *Staphylococcus aureus*, the most pathogenic species, is usually identified by its ability to produce coagulases (proteins that affect fibrinogen of the blood-clotting cascade). Since most other species of staphylococci do not produce coagulase, it is useful to divide staphylococci into coagulase-positive and coagulasenegative species. Coagulase-negative staphylococci are not highly virulent but are an important cause of infections in certain high-risk groups. Although *Staphylococcus* infections were once readily treatable with antibiotics, some strains have acquired genes making them resistant to multiple antimicrobial agents. *See* BACTERIA.

Cell structure, physiology, and genetics. *Staphylococcus* cells are spherical with a diameter of 0.5-1.5 micrometers. Clumps of staphylococci resemble bunches of grapes when viewed with a microscope, owing to cell division in multiple planes. The staphylococci have a gram-positive (retain violet color of Gram stain after destaining with alcohol) cell composition, with a unique peptidoglycan structure that is highly cross-linked with bridges of amino acids. *See* STAIN (MICROBIOLOGY).

Most species are facultative anaerobes (grow in the presence or absence of oxygen). Within a single species, there is a high degree of strain variation in nutritional requirements. Staphylococci are quite resistant to desiccation and high-osmotic conditions and can grow in the presence of 10% sodium chloride. These properties facilitate their survival in the environment, growth in food, and communicability.

In addition to genetic information on the chromosome, pathogenic staphylococci often contain accessory elements such as plasmids, bacteriophages, pathogenicity islands (DNA clusters containing genes associated with pathogenesis), and transposons. These elements harbor genes that encode toxins or resistance to antimicrobial agents and may be transferred to other strains. Genes involved in virulence, especially those coding for exotoxins and surface-binding proteins, are coordinately or simultaneously regulated by loci on the chromosome such as the accessory gene regulator, *agr*, and the *S. aureus* regulator, *sar. See* BACTERIAL GENETICS; BACTE-RIOPHAGE; PLASMID; TRANSPOSONS.

Resistance to antimicrobial agents. Staphylococci have a remarkable ability to acquire resistance to antimicrobial agents. Presently, most S. aureus isolates are resistant to penicillin and many other antibiotics. The most common resistance mechanism is mediated by production of beta-lactamase enzymes, which inactivate antibiotics by preventing their inhibition of bacterial cell wall synthesis. Methicillin is the drug of choice for strains that are resistant to multiple antimicrobial agents. Unfortunately, methicillinresistant S. aureus is being isolated with increasing frequency, especially from patients with nosocomial (hospital-acquired) infections. For most of these isolates, vancomycin is the drug used as a last resort. In the past few years, a few isolates with partial resistance to vancomycin have been found, causing concern that this organism may become untreatable unless new drugs become available. See DRUG RESIS-TANCE.

Strain typing. Infections caused by staphylococci, especially S. aureus, may be communicable, and it is important to trace the source of organism in outbreaks. Several typing methods have been used for epidemiologic studies. The classical phenotypic method, phage typing, was employed successfully for decades and is based on the fact that S. aureus may carry bacteriophages that can infect some less-related isolates. Based on susceptibility profiles, strains are grouped into five types. Because of technical difficulty and the emergence of nontypeable isolates, phage typing is being replaced by molecular techniques. One technique, pulsed-field gel electrophoresis, analyzes chromosomal DNA fragments generated by the restriction enzyme SmaI. Since only a few sites are recognized by SmaI in the S. aureus chromosome, a limited but variable number of fragments are generated. These fragments can be separated by pulsed-gel electrophoresis, and the profile generated represents a "fingerprint" for the strain in question. See GENETIC MAPPING.

Coagulase-positive (S. aureus) infections. Staphylococcus aureus causes approximately 8.9 million cases of human disease annually in the United States. Many people carry S. aureus and can develop endogenous infections if they become immunocompromised or if the organism enters the body through a break in the skin or mucous membrane. Infections acquired from exogenous sources are also common. Exogenous sources can be other people, fomites (inanimate objects, such as dishes, clothing, towels), the environment, or occasionally animals. Staphylococcus aureus strains are variable in expression of multiple-potential virulence factors; therefore, an encounter with this organism can lead to three types of illness: pyogenic infections, toxigenic infections, and food-borne intoxications.

Pyogenic infections. Staphylococcus aureus infections are common, and most develop into a pyogenic (pus-forming) lesion caused by acute inflammation. Inflammation helps eliminate the bacteria but also damages tissue at the site of infection. Typical pyogenic lesions are abscesses with purulent centers containing leukocytes, fluid, and bacteria. The abscess is enclosed by fibrin and cells. Many S. aureus products trigger inflammation, but the most important are cell wall components and several staphylococcal exotoxins (proteins that are secreted and not associated with the cell). Exotoxins with hemolytic or cytotoxic activity destroy tissue by damaging host cell membranes. Some also act on leukocyte membranes and prolong survival of S. aureus in the abscess

Pyogenic infections can occur anywhere in the body. Skin infections are common and may be communicable. Styes and boils are uncomplicated infections of hair follicles in the eyelid or sweat or sebaceous glands in the skin. More extensive skin infections are termed carbuncles or furuncles. Infections may also develop internally if staphylococci enter through a wound or escape from an abscess. Invasion may occur through a direct extension of an abscess or via the bloodstream. Blood infections (septicemia) can disseminate the organism throughout the body and abscesses can form internally. Infections in certain parts of the body, such as in the bone (osteomyelitis), respond poorly to antimicrobial therapy.

Toxigenic infections. Certain strains of *S. aureus* produce exotoxins that mediate two illnesses, toxic shock syndrome and staphylococcal scalded skin syndrome. In both diseases, exotoxins are produced during an infection, diffuse from the site of infection, and are carried by the blood (toxemia) to other sites of the body, causing symptoms to develop at sites distant from the infection.

Toxic shock syndrome is an acute life-threatening illness mediated by staphylococcal superantigen exotoxins. Toxic shock syndrome patients develop fever, hypotension, and multiorgan involvement and subsequent desquamation (peeling of the skin). The occurrence of staphylococcal toxic shock syndrome in the United States is approximately 0.3 to 0.8 case per 100,000 people. The S. aureus superantigen family contains nine antigenic variants of staphylococcal enterotoxins and toxic shock syndrome toxin-1 (TSST-1). Superantigens are unique molecules that bind simultaneously to major histocompatibility complex (MHC) Class II molecules on antigen-presenting cells and to receptors on T cells. Unlike conventional antigens, superantigens are neither processed by antigen-presenting cells nor expressed within the MHC Class II peptide groove. Instead, superantigens bind to peripheral surfaces of MHC Class II molecules and to the T-cell receptor at semiconserved elements of the β -chain variable region (V β). Binding in this manner stimulates a high percentage of T cells. The activated immune cells express excessive quantities of cytokines, which generate fever and adverse effects on the cardiovascular system, resulting in shock. Menstrual toxic shock syndrome is caused by TSST-1 produced by S. aureus colonizing the vagina or cervix. This illness is linked to tampon use during menstruation, with absorbency being a risk factor. Nonmenstrual toxic shock syndrome occurs when S. aureus causes an infection elsewhere in the body and produces TSST-1 or enterotoxins. See CELLULAR IMMUNOLOGY; TOXIC SHOCK SYNDROME.

Staphylococcal scalded skin syndrome, also known as Ritter's disease, refers to several staphylococcal toxigenic infections. It is characterized by dermatologic abnormalities caused by two related exotoxins, the type A and B exfoliative (epidermolytic) toxins. Exfoliative toxins are superantigens, but they are not typically associated with toxic shock syndrome. Instead, they are proteases that cause the separation of granular cells within the epidermis and skin peeling in a single plane at the stratum corneum. Bullous impetigo is a localized form of staphylococcal scalded skin syndrome in which the toxin acts only at the site of infection to form a liquid-filled lesion termed a bulla. Systemic staphylococcal scalded skin syndrome occurs when the exfoliative toxin disseminates via toxemia. One form of systemic illness, generalized exfoliative disease, is characterized by skin peeling in large sheets over most parts of the body. In an alternate form, staphylococcal scarlantiniform eruption, patients develop a generalized roughened sandpaper appearance of the skin without bulla formation or generalized exfoliation. Since nearly all people develop antibodies early in life, systemic staphylococcal scalded skin syndrome is rare after the age of 5.

Standard therapy for toxigenic infections is to identify and control the infection to minimize the amount of toxin produced. Toxic shock syndrome and severe staphylococcal scalded skin syndrome are typically treated with intravenous administration of beta-lactamase-resistant, cell-wall-active agents such as methicillin. In toxic shock syndrome, it is also important to provide supportive treatment for shock and counteract serum chemistry abnormalities. Intravenous immunoglobulin has proven successful in reducing severity of toxic shock syndrome; however, there is currently no vaccine for toxic shock syndrome or staphylococcal scalded skin syndrome.

Food-borne intoxications. At least 1 million cases of staphylococcal food poisoning occur annually in the United States. Staphylococcal food poisoning is not an infection, but an intoxication that results from ingestion of staphylococcal enterotoxins in food. The enterotoxins are produced when food contaminated with *S. aureus* is improperly stored under conditions that allow the bacteria to grow. Although contamination can originate from animals or the environment, food preparers with poor hygiene are the usual source.

The staphylococcal enterotoxins are stable and survive moderate heating. If ingested, they act in the gastrointestinal tract to stimulate the vagus nerve and signal the emetic center in the medulla. Vomiting, diarrhea, and abdominal cramps are the usual symptoms of staphylococcal food poisoning. Symptoms begin 1-6 hours following ingestion and usually resolve within 24 hours. Most cases do not require treatment. Antimicrobial agents have no effect. Severe cases accompanied by dehydration should be treated with intravenous fluids and antiemetic agents to control vomiting. Fatalities are rare and usually occur only in the very young or elderly. Because of toxin stability, the illness is not routinely preventable by heating foods. Pasteurization is effective in eliminating organisms in dairy products, although postpasteurization contamination by human sources is possible. See FOOD POISONING.

Effective methods for preventing staphylococcal food poisoning are aimed at eliminating contamination through common hygiene practices, such as wearing gloves, and proper food storage to minimize toxin production.

Coagulase-negative diseases. Because they frequently colonize human skin and mucosal surfaces, *Staphylococcus* species other than *S. aureus* are important opportunistic pathogens for humans. Most of these have low virulence and do not produce coagulase. Although several coagulase-negative staphylococci occasionally cause infections, *S. epidermidis* and *S. saprophyticus* are frequently associated with

certain human diseases. Staphylococcus epidermidis, the most virulent, coagulase-negative streptococcus, is a major cause of infected prosthetic devices, shunts, and catheters. The more frequent use of these devices has led to an increased occurrence of infections. Strains that cause these infections typically produce a variety of extracellular substances that facilitate their adherence to, and subsequent accumulation on, artificial materials. Staphylococcus saprophyticus causes at least 10% of urinary tract infections in nonhospitalized women aged 16-25 years. The source of this pathogen is unclear, but recent sexual intercourse is the leading risk factor. Staphylococcus saprophyticus is not highly virulent otherwise. Its surface components that bind to the urinary epithelium are important for disease. See MEDICAL BACTERIOLOGY.

Animal diseases. Coagulase-positive staphylococci are the most important *Staphylococcus* pathogens for animals. Certain diseases of pets and farm animals are very prominent. *Staphylococcus aureus* is the leading cause of infectious mastitis in dairy animals. This udder infection causes nearly \$2 billion in lost revenue to the United States dairy industry per year. *Staphylococcus intermedius* causes skin pyoderma in dogs and cats. *Staphylococcus aureus* and other coagulase-positive and coagulase-negative *Streptococcus* species are also important pathogens for goats, sheep, horses, pigs, poultry, and dolphins. Greg Bohach

Bibliography. J. E. Alouf and J. H. Freer (eds.), The Comprehensive Sourcebook of Bacterial Protein Toxins, 1999; J. C. Buzby et al., Bacterial Food-Borne Disease, Medical Costs and Productivity Losses, 1996; K. B. Crosby and G. L. Archer (eds.), The Staphylococci in Human Disease, 1997; M. P. Doyle, L. R. Beuchat, and T. J. Montville (eds.), Food Microbiology: Fundamentals and Frontiers, 1997; V. R. Fischetti et al., Gram Positive Pathogens, ASM Press, 2000; J. G. Holt et al. (eds.), Bergey's Manual of Systematic Bacteriology, vol. 9, 1994; S. Ladhani et al., Clinical, microbiological, and biochemical aspects of the exfoliative toxins causing staphylococcal scalded skin syndrome, Clin. Microbiol. Rev., 12:224-242, 1999; R. P. Novick (ed.), Molecular Biology of the Staphylococci, 1990; J. F. Timoney et al., Microbiology and Infectious Diseases of Domestic Animals, 8th ed., 1988.

Star

A self-luminous body that during its life generates (or will generate) energy and support by thermonuclear fusion.

Sun as reference. The fundamental reference for properties of stars is the Sun. The Sun's distance of 1.496×10^8 km (9.30×10^7 mi) from the Earth defines the astronomical unit (AU) and gives a solar radius (*R*) of 6.96×10^5 km (4.32×10^5 mi). From the energy received on Earth (1367 W m⁻²), the solar luminosity (*L*) is 3.85×10^{26} W. Now, $L = 4\pi R^2$ (surface area) $\times \sigma T^4$ (luminosity per unit area ac-

cording to the Stefan-Boltzmann law, where σ is the Stefan-Boltzmann constant); therefore, the effective blackbody "surface" temperature (*T*) is 5780 K (9950°F). (The "surface" is where the solar gases become opaque thanks to absorption by the negative hydrogen ion.) The Earth's orbital parameters and Kepler's third law give a solar mass of 1.99 × 10^{30} kg, the solar volume leading to an average density of 1.4 g cm⁻³ (1.4 times that of water). *See* ASTRONOMICAL UNIT; HEAT RADIATION; SUN.

The Sun's temperature and density rise inwardly as a result of gravitational compression, and at the still-gaseous solar center reach values of 15.7×10^6 K (28.3×10^{6} °F) and 150 g cm⁻³. Application of atomic theory to the solar spectrum and analysis of the solar wind and oscillations show the Sun to be 92% hydrogen (by number of atoms), 8% helium, and about 0.015% everything else (oxygen dominating, then carbon, neon, and nitrogen). The solar chemical composition provides a standard for other stars as well as an observational benchmark against which to compare theories of the chemical evolution of the Milky Way Galaxy.

Above about 8×10^6 K ($14 \times 10^{6^\circ}$ F), within the inner quarter of the radius and about half the mass, hydrogen (the proton, ¹H) fuses to helium primarily via the proton-proton chain:

 $^{1}\text{H} + ^{1}\text{H} \rightarrow ^{2}\text{H} + \text{positron (positive electron)} + \text{neutrino}$ positron + electron \rightarrow two gamma rays

$$^{2}\text{H} + ^{1}\text{H} \rightarrow ^{3}\text{He} + \text{gamma ray}$$

 $^{3}\text{He} + ^{3}\text{He} \rightarrow ^{4}\text{He} + 2(^{1}\text{H})$

The resulting release of energy supports the Sun against gravitational contraction. The radiation works its way out, degrading in energy with decreasing temperature, until it is released largely at optical wavelengths. The neutrinos, nearly massless particles, exit the Sun at close to the speed of light. Though nearly everything is transparent to them, enough are captured by terrestrial neutrino telescopes to confirm the theory of the Sun's energy production. There was initially enough hydrogen to run the chain for 10 billion years. With the Sun 4.6 billion years old (from radioactive dating of meteorites), it has over 5 billion years of core hydrogen fusion left. *See* NEUTRINO; POSITRON; PROTON-PROTON CHAIN; SOLAR NEUTRINOS.

Names. Over 6000 stars can be seen with the unaided eye. The brightest carry proper names from ancient times, most of Arabic origin that reflect the position of the star within its constellation. A more general system (from Johannes Bayer in the early seventeenth century) names stars within constellations by Greek letters roughly in accord with apparent brightness, followed by the Latin genitive of the constellation name (for example, Vega, in Lyra, is also Alpha Lyrae). More generally yet, brighter stars carry numbers, derived from the work of the astronomer John Flamsteed (1646–1719), in easterly order within a constellation (Vega also 3 Lyrae). All naked-eye stars also have HR (Harvard Revised)

Proper name	Greek-letter name	Apparent visual magnitude (V)	Distance, light-years	Absolute visual magnitude (M_v)	Spectral class
Sirius	α Canis Majoris	-1.46	8.6	1.43	A1 V
Canopus	α Carinae	-0.72	313	-5.63	F0 II
Rigil Kentaurus	α Centauri	-0.29	4.4	4.06	G2 V + K1 V
Arcturus	α Boötis	-0.04	37	-0.30	K1.5 III
Vega	α Lyrae	0.03	25	0.58	A0 V
Capella	α Aurigae	0.08	42	-0.48	G8 III + G0 III
Rigel	β Orionis	0.12	775	-6.8	B8 lab
Procyon	α Canis Minoris	0.34	11	2.62	F5 IV-V
Achernar	α Eridani	0.46	145	-2.76	B3 V
Hadar	β Centauri	0.61	335	-4.44	B1 III
Betelgeuse	α Orionis	0.70	425	-4.9	M2 lab
Acrux	α Crucis	0.76	320	-4.2	B0.5 IV + B1 V
Altair	α Aquilae	0.77	17	2.21	A7 V
Aldebaran	α Tauri	0.85	65	-0.65	K5 III
Antares	α Scorpii	0.96	600	-5.9	M1.5 lb
Spica	α Virginis	1.04	260	-3.49	B1 III-IV + B2 \
Pollux	β Geminorum	1.14	34	1.07	K0 III
Fomalhaut	α Piscis Austrini	1.16	25	1.73	A3 V
Deneb	α Cygni	1.25	2600	-8	A2 la
Mimosa	β Crucis	1.25	350	-3.9	B0.5 III
Regulus	α Leonis	1.35	78	-0.53	B7 V
Adhara	ϵ Canis Majoris	1.50	430	-4.2	B2 II

numbers assigned in order east of the vernal equinox. A variety of catalogs list millions of telescopic stars. *See* ASTRONOMICAL CATALOGS; CON-STELLATION.

Magnitudes and colors. About 130 B.C., Hipparchus assigned naked-eye stars to six brightness groups or "apparent magnitudes" (*m*), with first magnitude the brightest (**Table 1**). This scheme is now quantified as a logarithmic system such that five magnitudes correspond to a factor of 100 in brightness, rendering first magnitude 2.512... times brighter than second, and so on; calibration extends the brightest stars into negative numbers. The faintest stars now observed are around magnitude 30, 10¹² times fainter than magnitude zero.

Stars assume subtle colors from red to blue-white, reflecting different spectral energy distributions that result from temperatures ranging from under 2000 K ($(3100^\circ F)$ to over 100,000 K ($(180,000^\circ F)$). The magnitude of a star therefore depends on the detector's color sensitivity. A blue-sensitive detector "sees" blue stars brighter than does the eye, which is most sensi-

tive to yellow light, and will assign them lower magnitudes. Numerous magnitude systems range from the ultraviolet into the infrared, though the apparent visual magnitude ($m_v = V$) is still standard. The differences among the systems allow measures of stellar color and temperature. The total brightness at the Earth across the spectrum is given by the apparent bolometric magnitude (m_{bol}), which is dependent on the visual magnitude and temperature. *See* COLOR INDEX; MAGNITUDE (ASTRONOMY).

Distances. The fundamental means of finding stellar distances is parallax. As the Earth moves in orbit around the Sun, a nearby star will appear to shift its location against the background. The parallax (*p* in arcseconds) is defined as one half the total shift, and is the angle subtended by the Earth's orbital radius as seen from the star. Distance in parsecs (pc) is 1/pc, where 1 pc = 206,265 AU = 3.26 light-years. (A light-year is the distance that a ray of light will travel in a year.) The nearest star, a telescopic companion to Alpha Centauri, is 1.34 pc = 4.39 light-years away (**Table 2**). Parallaxes of hundreds of thousands of

Star	Distance, light-years	Spectral class	Apparent visual magnitude (V)	Absolute visua magnitude (M _v
Proxima Centauri	4.21	M5.5 V	11.05	15.49
Alpha Centauri A	4.40	G2 V	-0.01	4.34
Alpha Centauri B	4.40	K1 V	1.33	5.68
Barnard's Star	5.94	M4 V	9.54	13.24
Wolf 359	7.66	M6 V	13.54	16.68
BD+36°2147	8.32	M2 V	7.49	10.47
Sirius A	8.61	A1 V	-1.47	1.43
Sirius B	8.61	DA [†]	8.44	11.33
_726-8A	8.64	M5.5 V	12.57	15.45
_726-8B	8.64	M5.5 V	12.52	15.40
Ross 154	9.69	M3.5 V	10.95	13.58



Fig. 1. The Milky Way in Sagittarius, composed of millions of stars in the disk of the Galaxy. (© Anglo-Australian Observatory; photograph by David Malin)

stars found with the *Hipparcos* satellite allow fair accuracy to distances of over 1000 light-years. Parallax measures then allow the calibration of other distance methods (such as spectroscopic distance, main-sequence cluster fitting, and use of Cepheid variables) that extend much farther. *See* LIGHT-YEAR; PARALLAX (ASTRONOMY); PARSEC.

Distribution and motions. All the unaided-eye stars and 200 billion more are collected into the Milky Way Galaxy (or simply, the Galaxy), 98% concentrated into a thin disk over 100,000 light-years across. From the Sun, inside the disk and 27,000 light-years from the Galaxy's center in Sagittarius, the disk appears as the Milky Way. The appearance of the Milky Way varies considerably. It is brightest and widest toward the galactic center, and faintest in the opposite direction toward the anitcenter in Taurus and Auriga (**Fig. 1**). Surrounding the disk is a vast but sparsely populated halo.

Since stars orbit the Galaxy's center, they are seen to move relative to the Sun. Angular proper motions across the line of sight range from 10 arcseconds per year to below a thousandth of an arcsecond. Proper motions (μ , in arcseconds per year) depend on velocities across the line of sight (v_t , in kilometers per second) and distances (D, in parsecs) according to Eq. (1), which allows v_t to be found. Velocities along

$$\mu = \frac{v_t}{4.74D} \tag{1}$$

the line of sight (v_r) are measured from Doppler shifts of stellar spectrum lines. The combination yields space velocities (v_s) relative to the Sun according to Eq. (2). Statistical analysis of these motions shows

$$v_s^2 = v_t^2 + v_r^2$$
 (2)

the Sun to be moving through the local stars at a speed of 15-20 km/s (9-12 mi/s), roughly toward Vega. From radial velocities of sources outside the Galaxy, it is found that the Sun moves in a roughly circular orbit at 220 km/s (137 mi/s), which when combined with the space motions of other stars allows their galactic orbits to be determined. Stars in the disk have closely circular orbits; those in the halo have elliptical ones. *See* DOPPLER EFFECT; MILKY WAY GALAXY.

Absolute magnitudes. The apparent visual magnitude of a star depends on its intrinsic visual luminosity and on the inverse square of the distance. Knowledge of the distance allows the determination of the true visual luminosity, expressed as the absolute visual magnitude, M_v . This quantity is defined as the apparent visual magnitude that the star would have at a distance of 10 pc, and is found from the magnitude equation (3). Absolute visual magnitudes

$$M = m + 5 - 5\log d \text{ (pc)} \tag{3}$$

from class O to mid L range from $M_v = -10$ to +23 (a factor of 10^{13}). The Sun's absolute visual magnitude is in the middle of this range, +4.83. When the invisible infrared radiation from faint cool stars and the equally invisible ultraviolet from the ultraluminous hot stars are taken into account, the range in absolute bolometric luminosities (expressed in magnitudes as $M_{\rm bol}$) is 30 times lower. The quantity $M_{\rm bol}$ can be calibrated in watts; for the Sun, $M_{\rm bol} = +4.75$.

Spectral classes. Stars exhibit a variety of absorption-line spectra. The absorptions, narrow cuts in the spectra, are produced by atoms, ions, and molecules in the stars' thin, semitransparent outer layers, or atmospheres. Some stars show powerful hydrogen absorptions, others weak hydrogen and metals, yet others molecules but no hydrogen. Over a century ago, Edward C. Pickering lettered them according to the strengths of their hydrogen lines. After he and his assistants dropped some letters and rearranged others for greater continuity, they arrived at the standard spectral sequence, OBAFGKM, which was later decimalized (the Sun is in class G2). Classes L and T were added in 1999 (Fig. 2; Table 3). Since the sequence correlates with color, it must also correlate with temperature, which ranges from near 50,000 K (90,000°F) for hot class O, through about 6000 K (10,000°F) for solar-type class G, to under 2000 K (3100°F) for class L, and below 1000 K (1300°F) for class T (all of which are brown dwarf substars)

The majority of stars in the galactic disk have chemical compositions like that of the Sun. Differences in spectra result primarily from changes in molecular and ionic composition and in the efficiencies of absorption, all of which correlate with temperature. Molecules exist only at low temperatures. At higher temperatures they break up, and neutral atoms become ionized, so the absorptions of the neutral state eventually weaken with increasing temperature. When the temperature becomes high enough, single ionization is replaced by double, and so on. Hydrogen reaches maximum strength in class A. Hotter than class A, hydrogen becomes ionized and the neutral hydrogen lines weaken in classes B and O. Even the absorptions of individual ions are affected, as absorbing electrons are pumped toward higher-energy states at higher temperatures.

Most stars have more oxygen than carbon. In some cool stars, however, the ratio is reversed, carbon molecules replacing those of metallic oxides. Stars whose temperatures track class M are assigned class N, warmer ones class R; both are now called carbon stars and lumped into class C. In between, where oxygen and carbon are in equal abundance, leftover oxygen attaches preferentially to zirconium (whose abundance is also raised above normal); the zirconium oxide stars are called class S. *See* ASTRO-NOMICAL SPECTROSCOPY; CARBON STAR; SPECTRAL TYPE.

Hertzsprung-Russell diagram. Shortly after the invention of the spectral sequence, H. N. Russell and E. Hertzsprung showed that luminosity correlates with spectral class. A graph of absolute visual (or bolometric) magnitude (luminosity increasing upward) plotted against spectral class (or temperature, decreasing toward the right) is called the Hertzsprung-Russell (HR) diagram (Fig. 3). The majority of stars lie in a band in which luminosity climbs up and to the left with temperature (with the Sun in the middle). But another band begins near the location of the Sun and proceeds up and to the right (to lower temperature), luminosities increasing to thousands solar as temperature drops to class M. To be bright and cool, such stars must have large radiating areas and radii. To distinguish between the bands, the larger stars are called giants, while those of the main band are termed dwarfs (or main-sequence stars). Most carbon stars are giants, and stars of classes R, N, and S are mixed in with those of class M. Yet brighter stars to the cool side of the main sequence, with luminosities approaching 10⁶ solar, are called supergiants. In between the giants and the dwarfs lie a few subgiants. At the top, superior to the supergiants, are



Fig. 2. Spectral sequence ranging from hot class O to cool class M. The hydrogen lines increase in strength from O to A and then weaken, disappearing in class M, where molecules take over. (From N. Ginestet et al., Atlas de Spectres Stellaires, Observatoire Midi-Pyrénées and Observatoire de Strasbourg, 1992; Y. Yamashita et al., An Atlas of Representative Stellar Spectra, University of Tokyo Press, 1978; J. B. Kaler, Extreme Stars, Cambridge University, Press, 2001)

the very rare hypergiants. *See* DWARF STAR; GIANT STAR; SUBGIANT STAR; SUPERGIANT STAR.

White dwarf stars. In the lower left corner of the diagram, beneath the main sequence, are stars so dim that they must be very small. Since the first ones found were hot and white, they became known as white dwarfs in spite of their actual temperatures or colors. White dwarfs cannot be classified on the standard system. Most have almost pure hydrogen atmospheres caused by the downward settling of heavier helium atoms. Others have no hydrogen at all and display only helium lines. White dwarfs must therefore be positioned on the HR diagram according to

Class	Characteristic absorption lines	Color ^a	Temperature, K ^t
0	H, He ⁺ , He	Blue	31,000–49,000
В	H, He	Blue-white	9800-31,000
А	H	White	7500-9800
F	Metals, H	Yellow-white	6000-7500
G	Ca ⁺ , metals	Yellow	5300-6000
К	Ca ⁺ , Ca	Orange	3900–5300
Μ	TiO, other molecules, Ca	Orange-red	2200-3900
L°	Metal hydrides, alkalai metals	Red	<2200
T ^d	Methane, water, ammonia bands	Infrared	1000
R ^e	Carbon	Orange	4000-5800
N ^e	Carbon molecules	Red	2000-4000
Se	ZrO and other molecules	Orange-red	2000-4000



Fig. 3. Hertzsprung-Russell (HR) diagram, showing the positions of the major kinds of stars.

their temperatures (rather than their spectral classes). *See* WHITE DWARF STAR.

Subdwarfs. Stars in the galactic halo are deficient in heavy elements. Typical iron contents are a hundredth that of the Sun, and in the extreme a hundredthousandth. Low metal content makes halo dwarfs bluer than those of the standard main sequence, shifting them to the left and seemingly downward on the HR diagram, where they are known as subdwarfs. *See* STELLAR POPULATION.

Stellar sizes. Since, for any spherical star, luminosity $L = 4\pi R^2 \sigma T^4$, the luminosity in solar units is given by Eq. (4). Temperature (*T*) from spectral class

$$(L/L_{\rm sun}) = (R/R_{\rm sun})^2 (T/T_{\rm sun})^4$$
(4)

or spectral analysis plus the luminosity from absolute bolometric magnitude $M_{\rm bol}$ gives the stellar radius. Dwarf radii increase from about 10% that of the Sun in class M to nearly 15 solar in class O. Giant radii range past 10 times solar to over 1 AU, supergiants to nearly 10 AU. White dwarfs are about the size of Earth. The radii of stars are confirmed by direct measurement of angular size through interferometry and analysis of eclipsing double (binary) stars. *See* ECLIPSING VARIABLE STARS.

Spectral lines are sensitive to stellar density, so stars can be classified by size and luminosity on the basis of their spectra alone. In the standard Morgan-Keenan-Kellman (MKK) system, roman numerals I through V stand for supergiant, bright giant, giant, subgiant, and dwarf (VI is sometimes used for subdwarf). Once luminosity and absolute visual magnitude are known from a star's spectrum, its spectroscopic distance can be calculated from the magnitude equation (which must contain a correction for dimming by interstellar dust).

Stellar frequency distribution. The unaided-eye sky is filled with class A and B dwarfs, and with giants and supergiants that range from class O to M. Per unit volume of space, however, dim M and L dwarfs (none visible to the unaided eye and not counting dimmer brown dwarfs) dominate, constituting over 80% of all stars (Table 2). Of the standard sequence (discounting L and T), some 10% of stars are G dwarfs like the Sun and only 1% are class A stars. Rarest, making up only 0.0001%, are the O stars, which are so bright that they can be seen for great distances. There are about as many giants as A dwarfs, and supergiants are as rare as O stars. *See* HERTZSPRUNG-RUSSELL DIAGRAM.

Double and multiple stars. Most stars are members of some sort of community, from doubles through multiples (double-doubles, and so forth) to clusters, which themselves contain doubles. Separations between components of double stars range from thousands of astronomical units (with orbital periods of a million years) to stars that touch each other (and orbit in hours). Orbits of the more widely spaced binaries (periods of months to centuries) can be determined directly (or through interferometry). Limited orbital parameters of the more closely spaced ones are found by means of Doppler shifts caused by high orbital velocities. Some of the latter eclipse each other, the eclipses establishing orbital tilts, true orbital speeds, and stellar radii. The sum of the masses of a binary is found from the period and orbital dimension through Kepler's third law; the ratio of the masses is found by locating the center of mass of the system; and the combination of these two leads to determination of the individual masses. See BINARY STAR; KEPLER'S LAWS; ORBITAL MOTION.

Masses and main-sequence properties. Observations of hundreds of binary stars show that mass (M) increases upward along the main sequence from about 7.5% solar at cool class M to over 20 solar in the cooler end of class O. Extrapolation by theory, as well as through observations of clusters, suggests masses of 120 solar at the extreme hot end (class O3) of the main sequence. Fainter than the Sun, luminosity is proportional to M^3 ; from the Sun and brighter, luminosity goes as M^4 and then back to M^3 . This mass-luminosity relation is the result of higher internal temperatures and pressures in more massive stars caused by gravitational compression. Given the Sun's energy source and that luminosity varies smoothly with mass, the main sequence must be a hydrogen-fusing mass sequence. Giants, supergiants, and white dwarfs must therefore have some other energy source. See MASS-LUMINOSITY RELATION.

Carbon cycle. As internal temperature climbs above the 8×10^6 K limit, hydrogen fuses to helium via



Fig. 4. Images of the brown dwarf Gliese 229B, a companion to a much brighter M dwarf. (a) Discovery image, taken at Palomar Observatory in California (*T. Nakajima and S. Kulkarni*). (b) Hubble Space Telescope image (*S. Durrance and T. Golimowski; NASA*).

the proton-proton cycle at an ever-increasing rate. Above about 15×10^6 K ($27 \times 10^{6^\circ}$ F), so does fusion by the carbon cycle, in which carbon acts as a nuclear catalyst:

$$^{12}C + ^{1}H \rightarrow ^{13}N + gamma ray$$

But ¹³N is radioactive and decays, so the cycle proceeds:

 $^{13}\mathrm{N} \rightarrow \,^{13}\mathrm{C}$ + positron + neutrino

positron + electron \rightarrow two gamma rays $^{13}C + {}^{1}H \rightarrow {}^{14}N +$ gamma ray

$$^{14}N + {}^{1}H \rightarrow {}^{15}O + gamma rav$$

Again, ¹⁵O is unstable, so the cycle concludes:

$$^{15}\text{O} \rightarrow ^{15}\text{N}$$
 + positron + neutrino

positron + electron \rightarrow two gamma rays ^{15}N + $^{1}H \rightarrow ^{12}C$ + ^{4}He

Four protons are therefore converted to an atom of helium. Together with side chains that create heavier oxygen and fluorine, the carbon cycle overtakes the proton-proton chain around class F5 (close to 1.3 solar masses), greatly shortening the star's life. *See* CARBON-NITROGEN-OXYGEN CYCLES.

Stellar convection. The onset of carbon-cycle dominance coincides with a change in stellar structure. The Sun has a radiatively stable core surrounded by an envelope whose outer parts are in a state of convection that helps produce a magnetic field and magnetic sunspots. Hotter dwarfs have shallower convection layers, and above class F, envelope convection disappears, the cores becoming convective. The convective layers of cooler dwarfs deepen, until below about 0.3 solar mass (class M3), convection takes over completely and the stars are thoroughly mixed. *See* CONVECTION (HEAT).

Brown dwarfs. Below a lower mass limit of 7.5% solar, internal temperatures and densities are not great enough to allow the proton-proton chain to operate. Such stars, called brown dwarfs, glow dimly and redly from gravitational contraction and from the fusion of natural deuterium (²H) into helium (**Fig. 4**). Long sought, brown dwarfs are now being found in abundance with infrared technologies. These and the stars at the bottom of the main sequence together inhabit class L. All class T stars are brown dwarfs. The lower mass limit to brown dwarfs is unknown; they may even overlap the masses of planets. *See* BROWN DWARF; EXTRASOLAR PLANETS; INFRARED AS-TRONOMY.

Stellar rotation. The Sun rotates with an equatorial velocity of 2.0 km/s (1.25 mi/s) and a period of 25 days. Rotation broadens stellar spectrum lines, allowing the observation of projected velocities which, since axial inclinations are not generally known, are lower limits to the true velocities. The assumption of random inclinations allows rotation speeds of stars in different spectral classes to be found. Main-sequence stars also exhibit magnetically induced sunspots, whose rotation in and out of sight allows accurate knowledge of rotation periods. Cooler dwarfs rotate slowly because the magnetic fields they generate couple with stellar winds and slow them down. Above the "rotation break" in class F, rotation speeds increase, reaching maximum values of 300 km/s (190 mi/s) or more in class B. Class B stars are commonly surrounded by rotation-related disks, the "Be" stars exhibiting emission lines in their spectra. Because of their large sizes, giants and supergiants rotate slowly. See STELLAR ROTATION.

Clusters. Doubles and multiples are highly structured. Clusters are not, the member stars orbiting a common center of mass. Open clusters are



Fig. 5. The Jewel Box open cluster (Kappa Crucis), which lies next to the Southern Cross. (© Anglo-Australian Observatory: photograph by David Malin)

fairly small collections in which a few hundred or a thousand stars are scattered across a few tens of light-years. Examples are the Pleiades and Hyades in Taurus, which are among thousands of such clusters occupying the Galaxy's disk (and thus the Milky Way) [**Fig. 5**]. About 150 globular clusters occupy the Galaxy's halo, the poorest about as good as a rich open cluster, the best containing over a million stars within a volume 100 light-years across (**Fig. 6**). The main-sequence stars of globular clusters are subdwarfs, and like other halo stars are deficient in metals.

The HR diagrams of clusters are radically different from the HR diagrams of the general stellar field. Those of open clusters differ among themselves in having various portions of the upper main sequence removed. Some main sequences go all the way from class M to O, while others extend only from M through G. The effect is related to the cluster's age, since high-mass stars die first. Globular clusters, which lack an upper main sequence and are therefore all old (up to about 12×10^9 years), contain a distinctive "horizontal branch" composed of modest giants.

All clusters disintegrate with time as their stars escape through a form of evaporation aided by tides raised by the Galaxy. Old open clusters (more than 10^9 years old) are rare, while the compact globulars can survive for the age of the Galaxy.

If the HR diagram of a distant cluster is constructed using apparent magnitude rather than absolute, then comparison of the cluster's main sequence with that found from local stars using parallax allows the determination of the difference between absolute and apparent magnitude for the cluster, and therefore of the cluster's distance (after correction for interstellar absorption). Knowing the cluster's distance, in turn, makes it possible to learn the absolute magnitudes of the kinds of stars that are not within parallax distance. By locating the center of the distribution of globular clusters, most of which lie outside the obscuring dust of the galactic disk, it is possible to determine the distance to the center of the Milky Way Galaxy. *See* STAR CLUSTERS.

Variable stars. Dwarfs are generally stable. Giants and supergiants, however, can have structures that allow them to pulsate. Cepheids (named after the star Delta Cephei) are F and G supergiants and bright giants that occupy a somewhat vertical instability strip in the middle of the HR diagram. They vary regularly by about a magnitude over periods ranging from 1 to 50 days. The pulsation is driven by a deep layer of gas in which helium is becoming ionized. This layer alternately traps and releases heat from the outgoing radiation. Larger and more luminous Cepheids take longer to pulsate. Once this period-luminosity relationship is calibrated through parallax and mainsequence cluster fitting, the period of a Cepheid gives its absolute magnitude, which in turn gives its distance. Cepheids are vital in finding distances to other galaxies (Fig. 7). See CEPHEIDS; HUBBLE CON-STANT.

As the instability strip crosses the class F and A giants found in globular clusters, it produces the lowmetal RR Lyrae stars, which have periods of a day or less. Since they all have absolute magnitudes around 0.7, they too are valuable distance indicators. Farther down the HR diagram, the instability strip crosses the class A and F main sequence. Here and in the giants just above it are the subtly varying Delta Scuti



Fig. 6. Globular cluster 47 Tucanae, one of the grandest in the Milky Way Galaxy, containing over 1 million stars. (© Anglo-Australian Observatory; photograph by David Malin)



Fig. 7. A Cepheid 51 million light-years away in the spiral galaxy M100. (a–c) Hubble Space Telescope images of the star (in center of each image) on three different nights, showing changes in brightness. (d) Location of the star in one of the galaxy's spiral arms. (Wendy L. Freedman, Observatories of the Carnegie Institution of Washington; NASA)

stars. Where the instability strip hits the hydrogenrich white dwarfs, at temperatures between 10,500 and 13,000 K (18,400 and 22,900°F), it produces the ZZ Ceti stars, which pulsate nonradially with periods of minutes. (Hydrogen-poor white dwarfs have their own instability zone around 22,000 K or $39,100^{\circ}$ E)

Miras (after the star Mira, Omicron Ceti), or longperiod variables, are luminous M, C, or S giants that can vary visually by more than 10 magnitudes over periods that range from 50 to 1000 days, the pulsations again driven by deep ionization layers. Much of the visual variation is caused by small changes in temperature that shift the starlight from the visual to the infrared and that also alter the formation of obscuring titanium oxide molecules. The infrared (and bolometric) variation amounts to only two magnitudes or so. Other giants, and especially supergiants, are semiregular or even irregular variables with no periodicities at all. Still other stars, such as the class B subgiant Beta Cephei stars, are multiply periodic. *See* MIRA.

Duplicity produces its own set of intrinsic variations. If the members of a binary system are close enough together, one of them can transfer mass to the other, and instabilities in the transfer process cause the binary to flicker. If one companion is a white dwarf, infalling compressed hydrogen can erupt in a thermonuclear runaway, producing a sudden nova that can reach absolute magnitude -10. If the white dwarf gains enough matter such that it exceeds its allowed limit (the Chandrasekhar limit) of 1.4 solar masses, it can even explode as a supernova (discussed below). *See* CATACLYSMIC VARIABLE; NOVA; VARIABLE STAR.

Off the HR diagram. Various kinds of stars are not placeable on the classical HR diagram. The most common examples are the central stars of planetary nebulae, which are complex shells and rings of ionized gas that surround hot blue stars that range in temperature from 25,000 to over 200,000 K (45,000 to over 360,000°F) [Fig. 8]. While they can have luminosities as high as 10,000 solar, these central stars appear relatively dim to the eye, as they produce most of their radiation in the high-energy ultraviolet portion of the electromagnetic spectrum. This radiation ionizes the surrounding nebulae and causes them to glow. On the theoretician's extended HR diagram, in which luminosity is plotted directly against temperature, the planetary nebula central stars clearly connect the red giants with the white dwarfs. See PLANETARY NEBULA.



Fig. 8. Hubble Space Telescope image of the Ring Nebula in Lyra, a planetary rebula made of glowing ionized gas that surrounds a hot blue star. (Hubble Heritage Team; Associated Universities for Research in Astronomy; Space Telescope Science Institute; NASA)

Even hotter neutron stars (10^6 K, $1.8 \times 10^{6^\circ}$ F) are found associated with the exploded remains of supernovae (supernova remnants). Only 25 or so kilometers (15 mi) across, the visible ones spin rapidly, and are highly magnetic, with fields 10^{12} times the strength of Earth's field. Radiation beamed from tilted, wobbling magnetic axes can strike the Earth to produce seeming pulses of radiation (**Fig. 9**). Over time, the spin rates slow and these pulsars disappear. The remains of exploded stars, quiet pulsars (and active ones whose radiation does not hit Earth) litter the Galaxy. *See* NEUTRON STAR; PULSAR.

Evolution. The different kinds of stars can be linked through theories of stellar evolution. Stars are born by the gravitational collapse of dense knots within cold dusty molecular clouds found only in the Galaxy's disk. When the new stars are hot enough inside to initiate fusion, their contraction halts and they settle onto the main sequence. The higher the stellar mass, the greater the internal compression and temperature, and the more luminous the star. But the higher the internal temperature, the greater the rate at which hydrogen is fused, and the shorter the star's life. The fuel supply is proportional to mass, M, and the rate of fusion is proportional to the luminosity, L. Lifetime t is proportional to (fuel supply)/(rate of use) and thus to M/L. But L is on average proportional to $M^{3.5}$, so t is proportional to $1/M^{2.5}$. For example, a 100-solar-mass star will exhaust its core hydrogen in only 2.5 million years. See MOLECULAR CLOUD; PRO-TOSTAR.

Main-sequence lifetime is so long for stars under 0.8 solar mass that none has ever evolved in the history of the Galaxy. When a star between 0.8 and about 10 solar masses uses all its core hydrogen, its outer layers expand and cool to class K (Fig. 10). The star then brightens as it becomes a giant, lower masses brightening by larger factors and first becoming subgiants. The ascent to gianthood is terminated when the central temperature is high enough (10^8 K) or 1.8×10^{8} °F) to initiate the fusion of core helium via the triple-alpha process, wherein three helium nuclei (alpha particles) combine to make one carbon atom. This process temporarily stabilizes the star with a helium-burning core and a hydrogen-burning shell. (Fusion with another ⁴He nucleus makes oxygen.)

When the helium is fused to carbon and oxygen, the core contracts, helium fusion spreads outward



Fig. 9. Hubble Space Telescope image of the inner parts of the Crab Nebula supernova remnant. A spinning neutron star (pulsar) is the left of the two stars near the center of the image. Part *b* is blowup of boxed area in *a*. (*J. Hester and P. Scowen; NASA*)

into a shell, and the star again climbs the giant branch, the hydrogen- and helium-fusing shells sequentially turning on and off. Such asymptotic giant branch (AGB) stars become larger and brighter than before, passing into class M, where they eventually become unstable enough to pulsate as Miras. Within a certain mass range, convection in giants can bring carbon from helium fusion (as well as other new elements) to the surface, and the star can become a class S and then a carbon star. Powerful winds strip the star nearly to its fusion zone, which is protected from the outside by a low-mass hydrogen envelope. As the inner region becomes exposed, the star heats and eventually illuminates the surrounding outflowing matter to produce a planetary nebula, which dissipates into the interstellar medium, leaving a carbonoxygen white dwarf behind.

When high-mass stars, those above about 10 solar masses, use up their core hydrogen, they too migrate to the right on the HR diagram, becoming not so much brighter but larger, cooling at their surfaces and turning into supergiants. Below about 40 solar masses they become class M red supergiants, losing huge amounts of matter through immense winds. Some stabilize there under the action of helium fusion; others loop back to become blue supergiants. Above 60 solar masses, so much mass is lost that the supergiants do not make it much past class B, stalling there as helium fusion begins. Supergiants are so massive that, whatever their outward disposition, when the internal helium is fused to carbon and oxygen, the core shrinks and heats to the point that carbon fusion can begin. The mixture converts to oxygen, neon, and magnesium, which then burns to a mixture of silicon and sulfur, and finally to iron, each fusion stage moving outward to surround the core in successive shells.

Since iron cannot fuse and produce energy, the core suddenly collapses and somehow "bounces" to explode away all the outer layers. From the outside, the observer sees the explosion as a type II supernova that can reach absolute magnitude –18. Temperatures within the exploding layers are high enough to create all the chemical elements, including iron, nickel, and uranium. The debris—the supernova remnant—expands for centuries into space (Fig. 9), its shock wave helping to compress the interstellar matter and create new stars. The remaining core is so dense that it converts to a neutron star or, if the mass is great enough, even to a black hole. *See* BLACK HOLE; CRAB NEBULA; NEBULA.

Other routes to stellar explosion derive from double stars. Mass transfer from a low-mass mainsequence star to a white dwarf companion can cause the white dwarf to exceed the Chandrasekhar limit of 1.4 solar masses (beyond which it cannot support itself by way of electron degeneracy). The resultant collapse of the white dwarf produces a hydrogenless type Ia supernova that can reach $M_v = -19$ with such reliability that such supernovae are used to measure distances to distant galaxies. Stellar evolution feeds huge quantities of enriched matter back into the star-generating clouds of interstellar space, caus-



Fig. 10. Evolution of stars on the HR diagram. Solar masses of stars are the indicated numbers along the main sequence. Lower main-sequence stars have never had time to evolve. Intermediate-mass stars evolve from the main sequence to become giants that stabilize during helium fusion in a "clump." (Subsequent evolution, when helium fusion is done, is not shown here.) High-mass stars evolve to the right as supergiants and then explode. (After J. B. Kaler, Astronomy!, Addison-Welsey-Longman, 1994, from work of *I. lben*, *Jr.*)

ing the metal content of the Galaxy to increase and providing the raw materials from which the Earth is made. *See* NUCLEOSYNTHESIS; STELLAR EVOLUTION; SUPERNOVA. James B. Kaler

Bibliography. L. H. Aller, *Atoms, Stars, and Nebulae*, 3d ed., Cambridge University Press, 1991; E. Böhm-Vitense, *Introduction to Stellar Astrophysics*, 3 vols. Cambridge University Press, 1989, 1992; J. B. Kaler, *Extreme Stars*, Cambridge University Press, 2001; J. B. Kaler, *Stars*, Freeman, New York, 1992, paperback, revised 1998; J. B. Kaler, *Stars and Their Spectra*, Cambridge University Press, 1989; J. B. Kaler, *The Cambridge Encyclopedia of Stars*, Cambridge University Press, 2006; J. B. Kaler, *The Hundred Greatest Stars*, Copernicus, New York, 2002; R. J. Tayler, *The Stars: Their Structure and Evolution*, 2d ed., Cambridge University Press, 1994.

Star clouds

Large groupings of stars with dimensions of 1000 to several thousand light-years. Star clouds are not stable clusterings of stars or even groups of a common origin, such as stellar associations. Instead, they are large areas where the stellar density is higher than average in a galaxy. They are nevertheless physical entities in that they result from large-scale star formation events or series of events that are to some extent limited in size by the characteristics of the galactic environment.

Large star clouds are found commonly in the spiral arms of galaxies such as the Milky Way Galaxy and the Andromeda Galaxy (M31), but are not found in very old galaxies such as elliptical galaxies M32 or NGC 3379. Conspicuous star clouds in the Milky Way Galaxy are found in the constellations Carina, Cygnus, Sagittarius, and Scutum. They are particularly spectacular when photographed with a widefield telescope, when tens of thousands of stars crowd the photo. The densest star clouds, such as those in the constellation Sagittarius, are also spectacular when photographed with the Hubble Space *Telescope* (see **illustration**). In some galaxies, such as the giant spiral galaxy M101, they are immense, ranging up to 6000 light-years $(5.7 \times 10^{16} \text{ km or})$ 3.6×10^{16} mi) in diameter. See ANDROMEDA GALAXY; MILKY WAY GALAXY.

Some astronomers believe that the evidence suggests that star groups are hierarchical, with open clusters the smallest units; they are grouped together



Hubble Space Telescope image of the center of the Sagittarius star cloud. (Space Telescope Science Institute; NASA)

in OB associations, and the associations are grouped in star clouds, which in turn are grouped in galaxies. This is sometimes interpreted to indicate that star formation proceeds from the large to the small scale, with the interstellar gas first breaking up into large clouds that gravitationally separate into protostar clouds, and then episodes of star formation occur within these clouds to form OB associations, the nuclei of which then can become stable star clusters. *See* GALAXY, EXTERNAL; INTERSTELLAR MAT-TER; MOLECULAR CLOUD; STAR; STAR CLUSTERS; STEL-LAR EVOLUTION. Paul Hodge

Bibliography. A. E. Dolphin and D. A. Hunter, Recent star formation in Shapley Constellation III in the Large Magellanic Cloud, *Astron. J.*, 116:1275-1285, 1998; Y. Efremov, Star complexes and associations, *Astron. J.*, 110:2757-2770, 1995; D. A. Gouliermis et al., The star-forming region NGC 346 in the Small Magellanic Cloud with *Hubble Space Telescope* ACS observations, I. Photometry, *Astrophys. J. Suppl.*, 166:549-556, 2006; R. C. Kennicutt, F. Bresolin, and A. E. Dolphin, Systematic properties of young clusters and associations in galaxies, in D. Geisler, E. K. Grebel, and D. Minniti (eds.), *Extragalactic Star Clusters*, IAU Symp. 207, pp. 410-420, Astronomical Society of the Pacific, San Francisco, 2002.

Star clusters

Groups of stars held together by mutual gravitational attraction. There are two basic morphological types: open clusters and globular clusters. Typical densities of field stars near the Sun are 0.1 star per cubic parsec. (One parsec equals 1.9 \times 10^{13} mi or 3.1 \times 10¹³ km.) Open clusters, with dozens to thousands of stars, have central densities of 0.3 to 10 stars per cubic parsec, and are often elongated or amorphous in shape (Fig. 1). Globular clusters, with a thousand to several million stars, and central densities of a few hundred to over 100,000 stars per cubic parsec, are generally spherical (Fig. 2). Associations are even looser assemblages, with a few hundred stars and central densities lower than the field. They are often recognized because of unusually large numbers of special types of stars. OB associations are dominated by hot, luminous, young O and B stars. T associations are dominated by young T Tauri variable stars. See SPECTRAL TYPE; T TAURI STAR.

Star clusters are important because, within each cluster, member stars probably are at the same distance from the Sun, and have the same age and the same initial chemical composition. Stars with larger masses begin life hotter and brighter than lower-mass stars and end their lives earlier. Most stars within clusters lie along a band in the luminosity-temperature plane (where temperature is measured by spectral type or color). In older clusters, this main-sequence band terminates at fainter and cooler levels. Ages derived from such observations provide a chronology of which clusters formed first. Differences observed among stars away from the main sequence, such as in luminosity and temperature or in chemical composition, reflect changes to the stars during the ends of their lives and provide laboratories for the study of stellar evolution. Apparent luminosities of main-sequence stars may be converted into intrinsic luminosities by distance estimates obtained via geometric methods, including stellar parallaxes or stellar motions, or by photometric methods. Such distance determinations map out the locations of clusters in the Milky Way Galaxy. *See* HERTZSPRUNG-RUSSELL DIAGRAM; MILKY WAY GALAXY; PARALLAX (ASTRONOMY); STELLAR EVOLUTION.

Open clusters. Open clusters, once called galactic clusters, are found mostly along the band of Milky Way, and are rarely found more than 1000 parsecs above or below the plane. A dozen are visible to the naked eye, such as Ursa Major, the Hyades (the horns of Taurus), the Pleiades (also in Taurus), and Praesepe ("the Beehive," in Cancer). Over a 1100 are cataloged, and the Milky Way probably contains tens of thousands, most hidden by interstellar obscuration. *See* HYADES; PLEIADES.

Distances and dimensions. The nearest, the Ursa Major cluster, lies about 25 parsecs away. The Hyades and Pleiades cluster lie about 45 and 125 parsecs away, respectively. Open clusters have been found at distances as large as 15,000 parsecs (Berkeley 29). Open clusters have diameters of a few to about 20 parsecs.

Spectral characteristics. The brightest main-sequence stars in the youngest clusters are luminous, violet O stars. In slightly older clusters, such as the Pleiades, the brightest main-sequence stars are blue B stars. In the oldest clusters, the brightest main-sequence stars are yellow G stars. Post-main-sequence stars include luminous red giants. About a dozen clusters contain Cepheid variables. Measuring distances to such clusters provide the intrinsic luminosities of the Cepheids, and their detection and measurement in distant galaxies provides excellent distance estimates to those galaxies. This is a key component in measuring the rate of the universal expansion. The Pleiades and other clusters of the same age or older often contain white dwarf stellar remnants. See CEPHEIDS; COSMOLOGY; HUBBLE CON-STANT; WHITE DWARF STAR.

Embedded clusters. The youngest clusters may be hidden within dark dust clouds opaque to optical light. Infrared methods can reveal such young clusters, including T Tauri variables, Herbig-Haro objects, and bipolar outflows associated with them. *See* HERBIG-HARO OBJECTS; INFRARED ASTRONOMY; PROTOSTAR.

Motions. Proper motions, distances, and radial velocities provide a snapshot of the motion of a cluster. Open-cluster motions are not very dissimilar from that of the Sun and appear to confine the clusters to the plane of the Milky Way.

Ages and dissolution. Comparison of main-sequence data with stellar evolution models yields age estimates. Embedded clusters have ages of less than 10^6 years. The Pleiades is about 3×10^7 years old, and the oldest open clusters, NGC 6791 and Berkeley 17, may reach 10^{10} years. Few old clusters are known, partly because of the dissolution



Fig. 1. Double cluster in Perseus. (Yerkes Observatory)

of clusters with time. The lower-mass stars can be said to float to the outer boundaries of a cluster and may be stripped away by passing stars, clusters, or giant molecular clouds. The oldest clusters are farther from the galactic plane, probably since they are more likely to survive in the lower densities there. *See* MOLECULAR CLOUD.

Chemistry. Most open-cluster stars have chemical compositions similar to that of the Sun, beginning life with 88% hydrogen, 10% helium, and 2% in so-called heavy elements such as oxygen, carbon, and iron. Open-cluster heavy-element abundances range from a factor of about 3 below that of the Sun to



Fig. 2. Great globular cluster in Hercules, Messier 13, photographed with 200-in. (500-cm) telescope. (*California Institute of Technology, Palomar Observatory*)

about 2 times above. Open clusters in the outer part of the plane of the Milky Way Galaxy tend to have lower heavy-element abundances.

Stellar associations. These are very low density stellar ensembles of young stars, first named by V. A. Ambartsumian in 1947. About 80 associations are cataloged, 8 of which contain Cepheid variables. Associations are no longer gravitationally bound, probably because of expulsion by strong stellar winds and supernovae of large amounts (over 50%) of gas and dust out of which the stars formed. Open clusters often appear as subconcentrations within some associations. The double cluster h and χ Persei (Fig. 1) lies within the Perseus OB1 association. *See* SUPERNOVA.

Globular clusters. Generally more distant, more massive, older, and more deficient in heavy elements than open clusters, globular clusters are dispersed all over the sky, but with a strong concentration in the direction of the galactic center (Sagittarius). In 1917, H. Shapley estimated distances to globulars to infer the distance to the galactic center. A total of 153 globular clusters are known, some of which are visible to the unaided eye, such as M13 in Hercules and especially ω Centauri. One to three dozen more clusters may be awaiting discovery behind interstellar obscuration. A significant fraction of globular clusters may have been formed in dwarf galaxies that have been absorbed by the Milky Way Galaxy. The Sagittarius dwarf galaxy is in the process of being absorbed, and several globular clusters are or have been part of it. The globular cluster M54 may be the Sagittarius dwarf galaxy's nucleus. See GALAXY, EXTERNAL.

Distances and dimensions. The nearest globular (M4 in Scorpius) lies about 2000 parsecs away, while the most distant ones in the Milky Way Galaxy, AM-1 and Palomar 3, lie over 100,000 parsecs from the Sun. The distances are estimated by photometric means since they are too distant for geometric methods. RR Lyrae variable stars, often found in globulars, are a frequently used "standard candle" to estimate their distances. Globular-cluster diameters range from about 10 to about 100 parsecs. The typical globular cluster contains about 100,000 stars, but the range is from 1000 (AM-4) to over 3 million (ω Centauri).

Spectral characteristics. The brightest main-sequence stars in globulars are yellow G stars, signifying that all such clusters in the Galaxy are old. Red giants are the brightest stars. Some clusters contain RR Lyrae variables, and some contain a few blue stragglers, stars hotter and more luminous than the brightest and hottest normal main-sequence stars. Some globular clusters contain a type of cepheid variable, but they differ from those found in open clusters and are not frequently used as distance indicators. *See* BLUE STRAGGLER STAR; VARIABLE STAR.

Structure. Some globular clusters have low central concentrations; other have very high values, approaching 100,000 stars per cubic parsec. Some show signs of having experienced what is known as core collapse. As the stars at the outer edges evaporate from the cluster, the central regions contract, densities become extremely high, and binary star sys-

tems form. About 20% of the globular clusters have experienced core collapse. The high frequencies of blue stragglers in some cluster cores, and the presence of millisecond pulsars (17 known in nine clusters) may be related to this dynamical evolution of the cluster. Heavy stellar remnants such as neutron stars are probably also present nearer the center. Bright x-ray sources may result from binary star transfer of mass onto a neutron star. Globular clusters are unusually rich in such sources, and all are near the cluster centers. *See* ASTROPHYSICS, HIGH-ENERGY; BINARY STAR; NEUTRON STAR; PULSAR; X-RAY ASTRONOMY.

Chemistry. Heavy-element abundances in globular clusters cover a wide range, from 50% above solar to 300 times below solar. There seem to be two chemically distinct classes of globular clusters. The metal-deficient class is typically 30 times poorer in heavy elements than the Sun, while the metal-rich class is typically about 3 times more deficient. Those associated with the Sagittarius dwarf galaxy have different element-to-iron abundance patterns than most other globular clusters.

Motions. The metal-deficient globulars have very high velocities relative to the Sun, and so do not belong to the disk of the Milky Way; they are found in a spherical distribution around the galactic center. The more metal-rich globulars have motions characteristic of disk stars and clusters, and are found within a few kiloparsecs of the plane. The metal-rich clusters are all within 10,000 parsecs of the galactic center.

Ages. All globular clusters in the Milky Way are very old, typically about 1.2×10^{10} years. There may be a modest age spread of 3×10^9 years. The globular clusters are a primary method of estimating the lower limit to the age of the galaxy and of the universe. This may be compared to the age derived by the *Wilkinson Microwave Anisotropy Probe* (*WMAP*) mission and the expansion rate derived from studies of extragalactic Cepheids. *See* COSMIC BACKGROUND RADIA-TION; WILKINSON MICROWAVE ANISOTROPY PROBE.

Clusters in other galaxies. Giant associations are known in spiral galaxies (such as M31 and M33) and irregular galaxies (such as Magellanic Clouds). Large open clusters are also known in spiral and irregular galaxies throughout the Local Group. Globular clusters are known in most galaxies, including ellipticals, throughout the Local Group and in galaxies out to disances of 10⁸ parsecs. The galaxy M87 has the largest number (6000 counted, 16,000 estimated). Generally all extragalactic globular clusters are old, like those in the Milky Way, although there is a class of blue populous clusters that are young, roundish, and fairly massive (10,000 stars) in some star-forming galaxies. See ANDROMEDA GALAXY; LOCAL GROUP; MAGELLANIC CLOUDS; STAR. Bruce W. Carney

Bibliography. *Astronomical Almanac*, annually; W. E. Harris, Globular cluster systems in galaxies beyond the Local Group, *Annu. Rev. Astron. Astrophys.*, 29:543–579, 1991; S. R. Majewski, Galactic structure surveys and the evolution of the Milky Way, *Annu. Rev. Astron. Astrophys.*, 31:575–639, 1993; C. Payne-Gaposchkin, *Stars and Clusters*, 1979.

Star tracker

A device that automatically measures the angular separation of stellar observations with respect to a reference platform. It is also referred to as an as-trotracker.

By utilizing the star tracker in conjunction with a precise time reference (chronometer) and a deadreckoning device consisting of gyroscopes and accelerometers (inertial navigator), a digital computer can correct many of the inertial navigator errors so that precise, autonomous (free from any radio position aids), terrestrial navigation can be achieved. The major errors corrected by the star tracker are introduced by the inertial navigator's gyroscopes that result in attitude deviations. In this configuration, called a stellar inertial system, very high precision aircraft autonomous navigation is achieved. Since the availability of radio position aids poses no problem for commercial aircraft, stellar inertial navigation technology is applied only on military vehicles. These navigation devices are used when radio position aids, such as Loran and the Global Positioning System (GPS), may be unavailable. See CHRONOME-TER; ELECTRONIC WARFARE; INERTIAL GUIDANCE SYS-TEM.

Star trackers are also utilized extensively for both military and nonmilitary applications on space probes, space-based interceptors, and satellites. In these applications the precise attitude capabilities of these devices provide the fiducial precision reference for pointing of the vehicle and Earth or planet sensors. On space missions, star trackers are the only sensor presently available that can provide arc-second attitude accuracy. *See* SPACE NAVIGATION AND GUIDANCE.

Stellar visibility. Terrestrial star trackers must be capable of making star measurements at sea level and in daylight. Cloud cover represents an impediment to terrestrial stellar observations. The probability of seeing a star at any given time on the Earth at sea level is 50%. At altitude greater than 45,000 ft (13.7 km), which is above the clouds, stellar observations are always available. This guaranteed availability of stellar observations at high altitudes has led to these devices being used on high-altitude aircraft (such as reconnaissance vehicles), ballistic missiles, satellites, and space probes.

An example of a terrestrial star tracker application is the stellar inertial navigation system of the U.S. Air Force's B-2 stealth bomber. When the B-2 is flying at altitudes above 45,000 ft, the inertial navigation system is continually updated by stellar observations. When it is flying a terrain-following, low-altitude mission, the star tracker is capable of making stellar observations 50% of the time and using these observations to correct the inertial navigation system's gyroscopic errors. Thus, the system is not dependent on any radio navigation aids which may be rendered unavailable in a wartime scenario; in particular, it is not susceptible to radio jamming of the Global Positioning System (GPS) or natural electromagnetic disturbances causing GPS outages. See MILITARY AIRCRAFT; SATELLITE NAVIGATION SYSTEMS.

Mounting. A gimbaled stellar inertial system is mounted on the stable element of an inertial navigator. The tracking system measures the telescope azimuth rotation angle and elevation angle to the acquired star with respect to a stable inertial reference. The tracker is programmed to observe different, widely angular separated stars in order to achieve accurate operation. The deviation of the measured stellar observations from their ideal stellar positions is utilized to enhance the performance of the inertial navigator.

The preponderance of modern systems do not contain gimbals. These systems are completely strapped down (that is, with no moving parts) and observe stars at random as the stars come into the rigid, fixed star tracker's field of view. Modern star trackers can operate in two modes. The first (lost in space) uses sophisticated star pattern matching algorithms to determine initial position of the vehicle. The second uses its knowledge of the telescope's location and direction to know which stars should be in the field of view and where their images are located on the star tracker's optical detector (Fig. 1). In order to make a three-axis attitude correction, there should be a least two stellar observations ideally separated by 90°. Smaller angular separations lead to a dilution of attitude correction precision. The multiple measurements need not be performed simultaneously since the star tracker corrects the attitude of each axis based on the stellar observations available. Thus, these strapdown startracker systems consist of either multiple telescopes with moderate fields of view or a single telescope with a very wide field of view for proper operation. See CELESTIAL NAVIGA-TION.

Processor functions. Digital computers, sometimes referred to as processors, perform many functions in a stellar inertial system. The computer has stored in its memory the ideal locations of the stars in a catalog called the ephemeris. Depending on the ephemeris used for the star tables, it usually stores the colorcorrected star magnitude, declination, and right ascension. Since the declination and right ascension change slightly during a year, these parameters are periodically altered. In some systems the proper motions of the stars are stored in the star catalog in order to minimize the catalog updates. The catalog deliberately culls out stars that are located close to each other, such as binary stars, in order to eliminate stellar observation ambiguity. Using accurate time supplied to the computer from a chronometer, the processor indexes the ephemeris with the apparent position of the platform so that it can determine which stellar observations will be processed. See AL-MANAC; DIGITAL COMPUTER; EPHEMERIS.

For a gimbaled star tracker, pointing and tracking commands are issued to precisely direct the telescope at a star. Since the strapdown star tracker has no gimbals to point the telescopes, the computer indicates the ideal locations of the stars that should be in the field of view of the fixed telescopes, along with their color-corrected brightness. The color correction in the ephemeris modifies the brightness


Fig. 1. Strapdown satellite star tracker. (a) Mounting of system on the satellite. (b) Star tracker in housing. (Northrop Grumman Corp.)

according to the responsivity of the telescope and its optical detector. Because of the lack of a pointing capability, the strapdown star-tracker ephemeris must contain as many as 10,000 or more stars, whereas the star tracker in the gimbaled, or pointing, system may contain only a table of less than 70 of the brightest stars. Thus, the strapdown star-tracker ephemeris is far more extensive than the gimbaled one, but the complexity and cost are more than offset by the reduction in cost, parts, size, and weight as well as enhanced reliability gained by the elimination of the gimbaled system's moving parts and gimbal shaft angular encoder readouts.

The computer processes the stellar data and chronometer data to arrive at the measured attitude. It then processes any known misalignments and compares the corrected measurements with the inertial estimate of attitude to provide enhanced attitude performance. Care is taken so that no star tracking is performed on stars in the vicinity of the Sun, Moon, or zodiacal light. This is done to prevent their high irradiance from corrupting the detection of the much weaker stellar irradiance signal. *See* ZODIACAL LIGHT.

Detectors. The preponderance of star-tracker photodetectors used to measure the stellar irradiance are solid-state, silicon semiconductor photosensors. These devices have their peak responsivity in the near-infrared region (0.6-1.0 micrometer). Imaging focal-plane arrays can have several million photosensors on a single semiconductor silicon chip arranged in a rectangular grid matrix (Fig. 2). Each of the individual photosensors on the focal-plane array is called a pixel. The devices are similar to the photosensors that have found wide utilization in home video-camera recorders. Two types of focalplane arrays used for star trackers are charge-coupled devices (CCD) and charge-injection devices (CID). These rectangular devices are less than 0.125 in. (3 mm) thick and are approximately 0.25 in. (6 mm) per side for a 65,000-pixel sensor to 1 in. (25 mm) per side for a megapixel sensor. The advantages of the imaging photosensor can be illustrated by applying it to a standard 4° field-of-view star-tracker telescope. For a star tracker with a star magnitude 6 capability (some strapdown star trackers have the capability of measuring stars as dim as magnitude 10), on the average, there will be over three stars on the photosensor image plane. Thus, these imaging focalplane arrays can process multiple stars with each stellar observation. The silicon focal-plane arrays are also more reliable than earlier photodetectors based on photomultipliers and image dissector electron tubes, and occupy only a small fraction of their volume. See CHARGE-COUPLED DEVICES; IMAGE TUBE (AS-TRONOMY); MAGNITUDE (ASTRONOMY); PHOTOMUL-TIPLIER; TELEVISION CAMERA.

Accuracy. Star-tracker accuracy ranges from subarcsecond to 30 arcseconds. For terrestrial astroinertial navigation, which must track stars in daylight at sea level, the predominant noise source is the sky background. The sky background noise is usually mitigated by large telescope apertures and by minimizing the field of view. In the aircraft applications of astro-inertial navigation systems, there are number of other error sources besides the star tracker accuracy, such as accelerometer noise, that degrade the system's performance. For these applications a star-tracker accuracy of 1 arcsecond is compatible with providing a latitude and longitude navigation accuracy of less than 600 ft (180 m) for mission flight times of up to 12 hours, although 1 arcsecond of the Earth's circumfereance is only about 100 ft (30 m). On the opposite end of the spectrum are the space trackers used for satellites and deep-space probes. These space-probe trackers do not contend with sky background noise when the stellar observation is sufficiently displaced from looking through a planet's atmosphere. Also, no star tracking is performed with the telescope looking in the vicinity of the Sun or Moon. This is done by selecting the location of the tracker with respect to its orbit or by providing multiple trackers to assure that the darkness of space provides the background illumination



Fig. 2. Silicon focal-plane arrays. (Northrop Grumman Corp.)

of the telescope that is tracking stars. The accuracy of these devices ranges from 30 arcseconds for communications satellites to sub-arcsecond trackers that are used on reconnaissance satellites and space probes which track scientific targets. The high precision is required on reconnaissance satellites operating in high Earth orbits (altitudes of thousands of miles) to steer Earth-observation sensors and to report the location of terrestrial regions of interest. In the reconnaissance and mapping satellite applications, the angular accuracy of the satellite's star tracker represents a pointing location error for objects located on the surface of the Earth. For a 1-arcsecond star tracker the terrestrial location reporting errors of objects range from 30 ft (9 m) for a 1000-nautical-milealtitude (1850-km) low-Earth-orbit (LEO) satellite to 600 ft (180 m) for a 19,325-nautical-mile-altitude (35,800-km) geosynchronous- or geostationary-orbit (GEO) satellite. See MILITARY SATELLITES.

Seymour Levine

Bibliography. K. Bachman, R. Dennis, and S. Levine, Strapdown astro-inertial navigation utilizing the optical wide-angle lens startracker, *Navig. J. Inst. Navig.*, 37(4):347-362, Winter 1990-1991; International Society of Optical Engineering (SPIE), *Space Guidance, Control and Tracking Conference* (vol. 1949), Orlando, Florida, April 12-14, 1993; M. Kuritsky and M. Goldstein, Inertial navigation, *Proc. IEEE*, 71:1156-1176, 1993; G. Quasius and F. McCanless, *Star Trackers and System Design*, 1966.

Starburst galaxy

A galaxy that is observed to be undergoing an unusually high rate of formation of stars. It is often defined as a galaxy that, if it continues to form stars at the currently observed rate, will exhaust its entire supply of star-forming material, the interstellar gas and dust, in a time period that is very short compared to the age of the universe. For a typical starburst galaxy, this gas and dust exhaustion time scale is less than 10⁸ years, that is, less than 1% of the age of the universe. Since such a galaxy must shortly run out of star-forming material, the high star formation rate currently observed not only must end soon but also must have started relatively recently or the gas supply would have run out long ago. It follows that such galaxies must be undergoing a passing burst of star formation. The rates of star formation that are estimated to occur in starbursts can exceed 100 times the mass of the Sun per year for the most energetic bursts observed in the nearby universe. This may be compared to about 2-3 solar masses per year in the Milky Way Galaxy. However, there are large uncertainties in such estimates because there is only weak evidence on the range of masses of stars that form in starburst events. There are theoretical limits to the total rate of star formation that can occur in a starburst. One limit is derived from the mass of all the available interstellar material turning into stars during the time taken for that material to fall inward under gravity to form a typically sized galaxy. Another limit is related to the expectation that when large numbers of newly formed massive stars reach supernova stage, they will blow away much of the remaining star-forming material, thus inhibiting the burst.

The term "starburst," coined in the 1981 by Daniel W. Weedman and his collaborators, usually implies that the burst of star formation is occurring in the nuclear regions of the galaxy, the term first being used to describe a sample of luminous spiral galaxies with bright pointlike nuclei. The prototype starburst nucleus is NGC 7714. Indeed, it has been confirmed by extensive studies that starbursts are very frequently confined to the inner approximately 1000-light-year-diameter regions of massive galaxies. However, there exist galaxies that meet the definition of short gas-exhaustion time scale but exhibit more galaxy-wide star formation. These include extragalactic H II regions, clumpy irregular galaxies, blue compact dwarf galaxies, and the nearest, best-studied starburst galaxy, M82 (**Colorplate 1**). The related term "starburst region" is also now often used to refer to large regions within galaxies where exceptionally energetic star formation is occurring.

Radiative emission. Starburst galaxies often appear blue because of the dominance of their energy output by hot massive stars. Many blue starburst galaxies were discovered in the extensive ultraviolet-sensitive objective prism surveys of B. E. Markarian. Most starburst galaxies are also strong emitters of mid- and far-infrared radiation, because the newly formed stars illuminate their natal clouds of dust and gas with ultraviolet and optical radiation, and as the dust warms, it reradiates this energy at far-infrared wavelengths. The Infrared Astronomical Satellite (IRAS), which surveyed the entire sky at far-infrared wavelengths in 1983, discovered over 100,000 galaxies that emit strongly at 60 micrometers, the majority of which are likely to be starburst galaxies. The most luminous of these galaxies are commonly termed ultraluminous infrared galaxies (ULIRGs). More detailed and more sensitive mapping in the mid-infrared with the Infrared Space Observatory (ISO), which operated from 1995 to 1998, helped to confirm that a large fraction of the newest star formation activity in starburst galaxies is hidden from view in the optical waveband by the dense, dusty molecular clouds which spawn the new stars. The Spitzer Space Telescope has been able to build on the foundations laid by the IRAS and ISO observatories, imaging starburst galaxies in the mid- and far-infrared with unprecedented sensitivity and detail (Colorplate 2). Only the slightly older young star complexes (a few million years) are clearly visible in the optical as bright blue complexes of stars and hot gas; these stars have had time to move away from their nurseries, or expel their natal clouds by stellar wind or supernova wind pressure. One of the most spectacular results from Spitzer is the multicolor imaging of starbursts, which show regions with intense emission from large molecules, probably polycyclic aromatic hydrocarbons (PAHs). This emission, usually colored red in the Spitzer false color images, dramatically highlights regions where the molecules are excited to shine by strong shocks occurring in star-forming regions. See INFRARED ASTRONOMY; INTERSTELLAR MATTER; MOLECULAR CLOUD; PROTOSTAR; SPITZER SPACE TELE-SCOPE.

Strong millimeter-wavelength emission lines of carbon monoxide, often highly concentrated to the nuclear regions, have also shown that starburst galaxies are rich in molecular gas, the dense, relatively cool material in which star formation takes place. In several galaxies there is evidence that these nuclear molecular clouds are disklike in shape, rotating around the galactic nucleus. At radio wavelengths, starburst galaxies show emission that is strongly correlated in strength with the far-infrared emission. Supernovae are common in starbursts because the most massive young stars have very short lives that end spectacularly by supernova. It is thought that the radio emission from starbursts arises from the effects of the supernova explosions on the surrounding interstellar medium. Extremely high resolution observations at radio wavelengths using the techniques of very long baseline interferometry (VLBI) have demonstrated the detection of individual young and highly luminous radio supernovae in a small number of galaxies, including one of the most energetic local ultraluminous infrared galaxies, Arp 220. In the submillimeter wavelength range, polarized emission-emission which has a preferred orientation for the light waves to oscillate in-has been found in Messier 82, thought to be caused by strong magnetic fields within the galaxy. The revealed orientation of the magnetic field lines indicates that the fields may help drag interstellar material into the center of the system. In the x-rays, starburst galaxies are weak compared to active galactic nuclei, and the xray emission that they do radiate is thought to come from superwinds and from binary-star radiation processes such as seen in the Milky Way Galaxy. See BINARY STAR; POLARIZATION OF WAVES; RADIO AS-TRONOMY; STELLAR EVOLUTION; SUPERNOVA; X-RAY ASTRONOMY.

Spectrum. The existence of a starburst in a galaxy is often deduced from the presence of strong, narrow emission lines in the optical spectrum. These lines originate in hot gaseous nebulae, called H II regions, which form around the hot young stars. Stronger proof of the existence of a large population of hot young stars in a galaxy comes from the detection in the optical spectrum of broad emission lines of helium, characteristic of very massive and shortlived Wolf-Rayet stars. Over 130 Wolf-Rayet galaxies are known. In a small number of galaxies, the ultraviolet absorption features caused by hot young stars have been detected by the International Ultraviolet Explorer (IUE) satellite. Starbursts can also be recognized spectroscopically in the mid-infrared by the detection of broad emission features from PAH molecules, using the Infrared Spectrograph on board Spitzer. See ASTRONOMICAL SPECTROSCOPY; NEBULA; ULTRAVIOLET ASTRONOMY; WOLF-RAYET STAR

Superwinds and stellar smoke. Energetic outflows of hot material have been discovered along the minor axes of many nearby starburst galaxies. Revealed in x-rays, emission lines of hydrogen, and interstellar absorption lines, these so-called superwinds are believed to be ejected by the combined energy of many supernova explosions and stellar winds in the starburst. Outflow wind velocities seem to be in the range 400-800 km/s (250-500 mi/s), and the rate of gas expulsion can be tens of solar masses per year, perhaps even exceeding the rates at which

gas is being converted into stars inside the host galaxies. These results have implications for the development over time of the heavy-element enrichment of the gas, both within galaxies and in the intergalactic medium. For example, they may help to explain the observation that galaxy spheroids of larger mass tend to have stars that are more enriched in heavy elements than the stars in lower-mass spheroids. Accompanying the hot x-ray wind in the nearest starburst galaxy, M82, *Spitzer* has discovered great plumes of cool "stellar smoke," revealed in midinfrared images sensitive to dust emission and to the PAH molecule (Colorplate 1).

Super star clusters. In some starburst galaxies, images from the *Hubble Space Telescope* indicate the existence of super star clusters. These are knots of emission thought to be extraordinarily luminous and compact clusters of young stars, possibly young analogs of globular clusters. Some of the most beautiful examples lie along the extended tails of interacting galaxy systems, like "strings of pearls." *See* HUBBLE SPACE TELESCOPE; STAR CLUSTERS..

Causes of starburst episodes. There are several theories as to why starbursts occur. One likely cause is the interaction between two galaxies as they pass close to or collide with one another. The tidal forces generated result in shock-wave compression of the interstellar material, loss of angular momentum and infall of material into the central regions of the galaxy, and star formation in the compressed clouds. Many starburst galaxies show evidence of interactions, including distorted appearance and long wispy tails of material. Not all starbursts can be due to interactions, however, since some display no evidence for any recent disturbance. Other mechanisms that are thought responsible for high star-formation rates in galaxies are very strong spiral density waves and central bar instabilities. See GALAXY FORMATION AND **EVOLUTION**

Starbursts in the early universe. In the local universe, starbursts are rare events that contribute less than 1% of the total amount of energy radiated by all galaxies, so they may seem exotic and interesting but not very important. However, dramatic observational breakthroughs made possible by the Hubble Space Telescope, the Keck Telescopes on Mauna Kea in Hawaii, the Cosmic Background Explorer (COBE) satellite, the Submillimeter Common User Array (SCUBA) on the James Clerk Maxwell Telescope in Hawaii, and the Spitzer Space Telescope have demonstrated that star formation rates in galaxies were probably much higher at very early times in the universe than they are now. In particular, evidence has been found that ultraluminous infrared galaxies were much more common and much more important at earlier times in the history of the universe than now. It is not known whether starbursts are responsible for the tremendous energy of these distant, young, highly infrared-luminous galaxies; if they are, then, starburst events may even have dominated the processes of galaxy building long ago, and much of the total energy radiated by early galaxies may be detectable only at infrared or submillimeter wavelengths because ultraviolet and optical light would have been very strongly obscured by the dense clouds of gas and dust involved in the starburst. *See* SUBMILLIMETER ASTRONOMY; TELESCOPE; UNIVERSE.

Relationship to other objects. Nuclei of starburst galaxies resemble active galactic nuclei in being bright and pointlike. However, active galactic nuclei are believed to differ from starbursts in harboring a central massive black hole, since they display, among other phenomena, an emission-line spectrum that is excited by a more powerful source of ultraviolet energy than can be provided by young stars. Many galaxies display the characteristics of both starbursts and active galactic nuclei, with a ring of star formation surrounding an active galactic nucleus. It is very likely that there exists a physical relationship between starbursts and the events leading to active galactic nuclei because both are frequently seen in the same galaxy, but the details of the connection remain to be fully explained. Possibilities include the following: (1) a galaxy-galaxy interaction triggers both phenomena by channeling gaseous material into the central regions of the galaxy; (2) the starburst event promotes the formation of a nuclear massive black hole; (3) the starburst event promotes the fueling of an existing, but previously quiescent, nuclear massive black hole by providing it with fuel; and (4) an active galactic nucleus turns on first, and then triggers a starburst in the surrounding interstellar medium. See ASTROPHYSICS, HIGH-ENERGY; BLACK Carol J. Lonsdale HOLE; GALAXY, EXTERNAL.

Bibliography. B. Brandl et al., The mid-infrared properties of starburst galaxies from *Spitzer*-IRS spectroscopy *Astrophys. J.*, 653:1129–1144, 2006; R. Genzel, D. Lutz, and L. Tacconi, Star formation triggered by galaxy collisions, *Nature*, 395:859–862, 1998; J. S. Greaves, Magnetic field surrounding the starburst nucleus of the galaxy M82 from polarized dust emission, *Nature*, 404:732–733, 2000; T. Jarrett et al., Remarkable disk and off-nuclear starburst activity in the Tadpole Galaxy as revealed by the *Spitzer Space Telescope, Astron. J.*, 131:261–281, 2006; D. W. Weedman et al, NGC 7714—The prototype star-burst galactic nucleus, *Astrophys. J.*, 248:105–112, 1981.

Starch

A carbohydrate that occurs as discrete, partially crystalline granules in the seeds, roots (tubers), stems (pith), leaves, fruits, and pollen grains of higher plants. Starch functions as the main storage or reserve form of carbohydrate; it is second in abundance only to cellulose, a major structural component of plants. Cereal grains, tuber and root crops, and legumes (seeds) have long been used as major sources of carbohydrate in human diets. *See* CELLU-LOSE.

Starch is isolated commercially from the following sources: cereal grain seeds [maize (corn), wheat, rice, sorghum], roots and tubers [potato, sweet



Fig. 1. Cornstarch as seen by (a) scanning electron microscope, (b) light microscope, and (c) polarized light. (From R. L. Whistler, E. F. Paschall and J. N. BeMiller, eds., Starch: Chemistry and Technology, 2d ed., Academic, 1984)

potato, tapioca (cassava), arrowroot], and stems and pith (sago). Cereal grains are often steeped first to loosen the starch granules in the endosperm matrix, followed by wet grinding or milling. Roots and tubers are ground to give a suspension containing starch granules. Then follows screening or sieving, washing, centrifuging, dewatering, and drying.

Characteristics. Starch, normally a white powder of 98–99.5% purity, is insoluble in cold water, ethanol, and most common solvents. Starch is stained a blue to reddish-purple color, depending on the amylose content, by dilute aqueous iodine.

The size and shape of starch granules depend on the botanical source of the starch. Rice starch, a small-granule starch, has polygonal granules ranging in size from 3 to 5 micrometers, while potato starch, a large-granule starch, has ellipsoidal or oval granules from 15 to 120 μ m. Normal cornstarch has spherical or polygonal granules from 5 to 30 μ m (**Fig. 1**). When viewed through a microscope under polarized light, starch granules are birefringent and show a characteristic "maltese cross" (Fig. 1*c*). Starch granules have characteristic x-ray diffraction patterns and have been estimated to be 20–50% crystalline.

Structure and architecture. Starch, a polymer of glucose, is an alpha-glucan, predominantly containing alpha-1,4-glucosidic linkages with a relatively small amount of alpha-1,6-glucosidic linkages forming branch points (**Fig. 2**). Two major polymeric components are present: amylose and amylopectin. *See* GLUCOSE.

Amylose, long thought to be a linear polymer, contains a limited amount of long-chain branching. Normal corn amylose contains about 900–1000 glucose units and appears to be a mixture of nearly equal numbers of large branched (5.3 chains on average) molecules and small unbranched molecules. In many respects amylose behaves like a linear polymer. In aqueous solution, it complexes with iodine to yield a deep blue color and associates with itself (retrogrades) to form precipitates or gels.

Amylopectin is a high-molecular-weight (10⁷-10⁸) branched polymer with an average chain length of 20-26 glucose units. The structure is composed of clusters of branch points with individual chains in each cluster containing 12-16 glucose units. Some longer chains extend through more than one cluster. Association of chains in amylopectin clusters forms the crystalline regions of the granule. The branch points are located in the less organized, amorphous regions of the granule.

The amounts of amylose and amylopectin vary with the source of the starch. Common starches



Fig. 2. Structures involved in starch. (a) Linear chain of alpha-1,4-glucan. (b) Alpha-1,6 branch point. (After Y. Pomerzanz, ed., Wheat: Chemistry and Technology, 2d ed., American Association of Cereal Chemists, 1971)

contain 15–30% amylose, the remainder being amylopectin. Special hybrids, with altered ratios of amylose to amylopectin, have been produced. Waxy starches from maize (commercially available), barley, rice, and sorghum contain only amylopectin. Highamylose maize starches containing 50–70% amylose are available.

Gelatinization. As a dilute aqueous suspension of starch is heated above 50°C (122°F), swelling of the granule begins in the less organized, amorphous regions with loss of birefringence and crystallinity. This is known as gelatinization and occurs over an interval of about 10°C (18°F) within the range of 55-75°C $(131-167^{\circ}F)$ for most common starches. As the temperature increases, irreversible swelling occurs, the crystalline areas undergo disorganization, and amylose leaches from the granule. When a sufficiently high temperature is reached (cooking, pasting), granule structure is destroyed, with apparent dissolution of the amylose and amylopectin components. The increase in viscosity occurring during gelatinization is an important property of starch, resulting in its common use as a thickener in foods.

Use in food systems. Starches are involved in important roles in foods, either naturally occurring in an ingredient or added to achieve a desired functional characteristic. Often the desired functional characteristic (thickening; gelling; adhesive; binding; improving acid, heat, and shear stability) cannot be achieved by using a native starch. Starches may be altered physically, chemically, or enzymatically to produce modified starches with improved functional properties.

Modified starches, commonly known as starch derivatives, are controlled in the United States by Food and Drug Administration (FDA) regulations. Modification of starch may involve a change in physical form (pregelatinization), controlled degradation (dextrinization, oxidation, acid or enzyme modification) or chemical reaction (oxidation, substitution, cross-linking). Most modifications are accomplished in aqueous suspensions under conditions in which the granule is not destroyed. Often, only minor modification is required to generate significant changes in starch properties. Modified starches can be produced from any starch, but the common commercial sources in the United States are normal maize, waxy maize, tapioca, and potato starches. *See* FOOD.

Starches are modified for the following reasons: to improve the resistance of swollen granules to acid, heat, and mechanical processing conditions; to eliminate unwanted gelling characteristics; to reduce or increase starch gelatinization temperature; to generate freeze-thaw stability; and to maximize, reduce, or inhibit viscosity.

Modified starches that produce solutions of lower viscosity are obtained by (1) treating aqueous suspensions of starch with dilute acid below the gelatinization temperature (acid-thinned starches); (2) treating a gelatinized starch slurry with an alpha-amylase (enzymatically modified starch), commercially accomplished in a continuous process; (3) heating (roasting) dry starch with small amounts of acid (dextrinization) to produce dextrins (pyrodextrins, British gums, white dextrins, yellow dextrins) that often have a pale yellow to tan color; or (4) oxidizing starch in an aqueous suspension, which results in the introduction of carboxyl and carbonyl groups and partial depolymerization of the starch. Each of these processes produces starch derivatives with the desired altered properties, for example, viscosity, paste clarity, or gel-forming ability, for different uses.

Cross-linked starch derivatives are produced by treating an aqueous suspension of starch granules with cross-linking reagents. The resulting modified starches have improved cooking characteristics (higher gelatinization temperatures); increased stability to acid, heat, or shear, with resulting control of viscosity during processing; and improved textural characteristics of starch pastes. Starch ethers or esters (substituted starches) are formed by treating an aqueous suspension of starch with appropriate reagents in the presence of an alkaline catalyst. These stabilized starch derivatives have freeze-thaw stability and a reduced tendency to undergo syneresis (formation of water on the surface of a gel or paste due to retrogradation).

Glucose (dextrose), glucose syrups, maltodextrins, maltose syrups, and high-fructose corn syrup are obtained from starch by using acid or enzymes. Enzymes are used industrially with increasing frequency because of the mild conditions required for processing and the range of products that can be obtained. High-fructose corn syrup finds extensive use as sweetener, particularly in soft drinks. *See* ENZYME; FOOD MANUFACTURING; FRUCTOSE; MALTOSE.

Nonfood uses. The paper, textile, adhesive, chemical, pharmaceutical, and polymer industries use starch and starch derivatives. Organic acids and organic solvents for use as chemical intermediates, enzymes, hormones, antibiotics, and vaccines are industrially produced from starch. *See* CARBOHY-DRATE. David R. Lineback

Bibliography. T. L. Barsby, A. M. Donald, and P. J. Frazier, *Starch: Advances in Structure and Function*, 2002; A.-C. Eliasson (ed.), *Starch in Food*, 2004; M. N. Sivak, J. Preiss, and S. Taylor (eds.), *Advances in Food and Nutrition Research*, vol. 14: *Starch—Basic Science to Biotechnology*, 1998; R. L. Whistler, E. F. Paschall, and J. N. BeMiller (eds.), *Starch: Chemistry and Technology*, 3d ed., 2001; O. B. Wurzburg, *Modified Starches: Properties and Uses*, 1986.

Stark effect

The effect of an electric field on spectrum lines. The electric field may be externally applied, but in many cases it is an internal field caused by the presence of neighboring ions or atoms in a gas, liquid, or solid. Discovered in 1913 by J. Stark, the effect is most easily studied in the spectra of hydrogen and helium, by observing the light from the cathode dark space of an electric discharge. Because of the large potential



Fig. 1. Stark effect for helium lines. (a) The 414.4- and 416.9-nm lines at field strengths 0–85,000 V/cm. Note the disappearance of $^{1}P^{-1}H$ line at a certain field strength, and symmetry of the s components (those with polarization perpendicular to the field), indicated by dots. (b) The 492.2-nm line at 0–40,000 V/cm. Note the crossing-over of two $^{1}P^{-1}F$ components. (J. S. Foster and C. Foster, McGill University)

drop across the region, the lines are split into several components. For observation perpendicular to the field, the light of these components is linearly polarized. The splitting can be easily resolved with a spectrograph of moderate dispersion, amounting in the case of the H_{γ} line at 434 nanometers to a total spread of 3.2 nm for a field of 10,400 kV/m.

Linear Stark effect. This effect exhibits large, nearly symmetrical patterns. Examples are shown in Fig. 1, where π and σ refer to the two states of polarization. The interpretation of the linear Stark effect was one of the first successes of the quantum theory. According to this theory, the effect of the electric field on the electron orbit is to split each energy level of the principal quantum number n into 2n - 1 equidistant levels of separation proportional to the field strength. Thus the higher members of a series show a larger number of components and greater overall splittings. The criterion for the occurrence of the linear Stark effect is that the splitting of the levels shall be large compared to the natural separation of levels of the same *n* but different *L*, where L is the quantum number of orbital angular momentum. See QUANTUM NUMBERS; ZEEMAN EFFECT.

Quadratic Stark effect. This occurs in lines resulting from the lower energy states of many-electron atoms. Here the large separation of states of different *L* results from the penetration of the valence electrons into the core of other electrons, with the result that the permanent dipole moment associated with hydrogenlike orbits no longer exists. There is, however, a small induced dipole moment due to polarization of the atom. This moment is proportional to the electric field strength, and since the energy change is proportional to the product of the dipole moment and the field strength, the energy levels shift by an

amount depending on the square of the field. Thus all levels have shifts of the same sign and therefore are displaced to lower energies. Each field-free level is also split, as a result of the space quantization of the angular momentum vector J, so that there are J + 1 components if J is integral, or $J + \frac{1}{2}$ components if it is half-integral. Because the lower levels are usually less displaced than the higher ones, the quadratic Stark effect ordinarily shows itself as a shift of the lines toward the red end of the spectrum, with an accompanying separation into several components. The quadratic Stark effect is basic to the explanation of the formation of molecules from atoms, of dielectric constants, and of the broadening of spectral lines.

Intermolecular Stark effect. Produced by the action of the electric field from surrounding atoms or ions on the emitting atom, the intermolecular effect causes a shifting and broadening of spectrum lines. The molecules being in motion, these fields are inhomogeneous in space and also in time. Hence the line is not split into resolved components but is merely widened. Particularly in the electric discharge through gases with high currents, the large ion density may cause very wide lines. The amount of the broadening is found to run parallel to the sensitivity of the line to the Stark effect and thus is greatest for those lines susceptible to the linear effect.

Inverse Stark effect. This is the effect as observed with absorption lines. It has been detected, for example, by applying an electric field to potassium vapor and measuring a small displacement of the absorption lines toward the red. The displacements are found to be proportional to the square of the field strength, as in the quadratic Stark effect. In addition, certain lines of large *n*, which are normally forbidden by the selection rule for *L*, are observed to appear in the presence of the field. This type of transition is said to be produced by forced dipole radiation. Such forbidden lines also may be obtained in emission. In Fig. 1, the lines 2^1P-6^1S and 2^1P-4^1F are examples of forced dipole transitions. F. A Jenkins; W. W. Watson

Quantum-confined Stark effect. This is the Stark effect observed in structures in which the hydrogenic system is confined in a layer of thickness much less than its normal diameter. This is not practical with atoms, but the effect is observed with excitons in semiconductor quantum-well heterostructures. It is important that quantum-confined Stark shifts can be much larger than the binding energy of the hydrogenic system. The resulting shifts of the exciton optical absorption lines can be used to make optical beam modulators and self-electrooptic-effect optical switching devices.

Wannier excitons behave like very large hydrogen atoms; for example, in gallium arsenide (GaAs) their diameter is approximately 30 nm, and their binding energy is approximately 4 meV. In a quantum well, consisting, for example, of a 10-nm gallium arsenide well layer between aluminum gallium arsenide (Al-GaAs) barrier layers, the exciton is confined primarily in the gallium arsenide layer, becoming overall smaller. This smaller size, incidentally, gives the exciton a larger binding energy (for example, 10 meV), and in contrast to most semiconductors, exciton absorption resonances can be seen clearly even at room temperature in many quantum-well structures. When an electric field is applied perpendicular to the layer planes, the electron and hole in the exciton are polarized, being displaced toward opposite sides of the well. This polarization by the field gives essentially a quadratic Stark shift of the ground-state exciton in a symmetric quantum well. Such Stark shifts in quantum wells can be much larger than the binding energy (typically 40 meV).

In normal semiconductors with unconfined excitons, some quadratic Stark shifts can also be seen, but these are limited to a small fraction (10%) of the binding energy because of line broadening resulting from field ionization. Fields so large that they correspond to a potential drop of more than one binding energy over the exciton diameter (for example, about 100 kV/m) field-ionize the unconfined exciton very rapidly, pulling the electron and hole away from one another in much less than a classical orbit time. Because the particle lifetime is so short, the optical absorption spectrum shows primarily strong broadening of the exciton lines, according to the Heisenberg uncertainty principle. With the confined exciton, however, the barriers prevent the exciton from being rapidly field-ionized, even at fields of about 10 MV/m. Hence the exciton absorption lines in the quantum-confined Stark effect are not strongly broadened, even though the Stark shifts in such large fields can be many times the binding energy of the exciton (Fig. 2). The exciton absorption lines in Fig. 2 correspond to the creation of excitons in their ground state, and the shifts of the absorption line correspond to the (quadratic) quantum-confined Stark shift of this exciton ground state. The absorption peak height is reduced with field because there is less spatial overlap of the electron and hole as they are pulled apart, lowering the absorption strength of the transition.



Fig. 2. Optical absorption spectra of 9.4-nm gallium arsenide (GaAs) quantum wells for different fields perpendicular to the layers, showing the quantum-confined Stark effect shift of the exciton absorption peak. (After D. A. B. Miller, J. S. Weiner, and D. S. Chemla, Electric-field dependence of linear optical properties in quantum will structures: Waveguide electroabsorption and sum rules, IEEE J. Quant. Electron., QE-22:1816–1830, 1986)

A typical structure for observing the quantumconfined Stark effect would contain about 50 gallium arsenide quantum-well layers of total thickness of 1 micrometer within a semiconductor diode structure. Reverse-biasing the diode with about 10 V applies a 10-MV/m electric field; this can change the transmitted power of a light beam of appropriate wavelength (for example, 850 nm) by a factor of 2 to make an optical modulator. This offers a method for modulating light beams with very low operating energy in small devices at room temperature. *See* ARTIFICIALLY LAYERED STRUCTURES; ELECTROOP-TICS; EXCITON; OPTICAL MODULATORS; SEMICONDUC-TOR HETEROSTRUCTURES. David A. B. Miller

Bibliography. B. H. Bransden and C. J. Joachain, *Physics of Atoms and Molecules*, 2d ed., 2003; E. U. Condon and H. Odabasi, *Atomic Structure*, 1980; R. D. Cowan, *The Theory of Atomic Structure and Spectra*, 1981; H. Haug (ed.), *Nonlinear Optical Properties of Semiconductors*, 1988; P. H. Heckman and E. Trabert, *Introduction to the Spectroscopy of Atoms*, 1989.

Static

A hissing, crackling, or other sudden sharp noise that tends to interfere with the reception, utilization, or enjoyment of desired signals or sounds. Perhaps the commonest form of static is that heard in ordinary broadcast receivers during electrical storms. Interference in radio receivers caused by improperly operating electric devices in the vicinity is sometimes also called static. *See* SFERICS.

The crackling sounds heard when long-playing plastic phonograph records are played are also called static. These sounds are caused by sudden deflection of the phonograph needle by dust particles, which are attracted to the grooves of the record by surface electric charges that are caused by friction on dry days. Static appears as momentary white specks in a television picture. *See* ELECTRICAL INTERFERENCE; ELECTRICAL NOISE. John Markus

Static electricity

Electric charge at rest, generally produced by friction or electrostatic induction. In the sixth century B.C., Thales of Miletus, who is usually credited with the discovery of static electricity, described experiments in which rubbed amber induced movement in nearby particles. It was not until the eighteenth century, when Benjamin Franklin carried out his kite-flying experiments in which he "captured" some thunderstorm electricity in a Leyden jar, that it was verified that thunderstorm electricity and static electricity were one and the same. *See* ELECTRIC CHARGE.

Triboelectrification is the process whereby charge transfer between dissimilar materials, at least one of which must have a high electrical resistivity, occurs due to rubbing or mere contact. Contact and separation of materials occur during industrial operations such as powder processing and the manufacture of plastic and other materials in sheet form. *See* ELECTRICAL RESISTIVITY.

Most organic and polymeric materials have volume resistivities greater than 10⁸ ohms per meter and retain charge for periods of many hours. Triboelectric charging is not fully understood, and the magnitude of charging may be strongly influenced by surface contamination of the contacting materials, relative humidity of the atmosphere, and the energy of rubbing. For example, during the grinding of material during powder manufacture, charge levels of about a microcoulomb per kilogram are typical. This is a relatively low level of charge. The more energetic process of pneumatically conveying a powder along a pipe may give rise to charge levels a thousand times greater and an electrostatic hazard may arise. Hazards are due to sparks to ground after a relatively large amount of static charge has accumulated on a body that is improperly grounded. See ELECTRIC IN-SULATOR; ELECTRIC SPARK.

In modern industry, highly insulating synthetic materials, such as plastic powders and insulating liquids, are used in large quantities in an ever increasing number of applications. Such materials charge up readily, and large quantities of electrical energy may develop with an attendant risk of incendiary discharges. When, for example, powder is pneumatically transported along pipes, charge levels of up to about 100 microcoulombs per kilogram can develop and potentials of thousands of volts are generated within powder layers and the powder cloud. Energetic sparking from charged powder may initiate an explosion of the powder cloud. Similar problems occur when insulating liquids, such as certain fuels, are pumped along pipes, and it is essential that strict grounding procedures are followed during the refueling of aircraft, ships, and other large vehicles.

The capacity of a person for retaining charge depends upon stature, but is typically about 150 picofarads. Even the simple operations of removing items of clothing or sliding off a chair can lead to body discharges to ground of about 0.1 μ C, which are energetic enough to ignite a mixture of natural gas and air. Human body capacitance is sufficiently high that, if poorly conducting shoes are worn, body potential may rise to 15,000 V or so above ground during industrial operations such as emptying bags of powder. Sparking may then occur with energy exceeding the minimum ignition energy of powder or fumes, so initiating a fire or explosion. Conducting footware should be used to prevent charge accumulation on personnel in industrial situations where triboelectrification may occur. See CAPACI-TANCE.

In the microelectronics industry, extremely lowenergy discharges, arising from body potentials of only a few tens of volts, can damage microelectronics systems or corrupt computer data. During the handling of some sensitive semiconductor devices, it is imperative that operators work on metallic grounded surfaces and are themselves permanently attached to ground by conducting wrist straps. Whenever films of insulating material are wound over rollers, such as in the photographic industry, surface charging and subsequent discharging occur. The discharges observed from charged insulator surfaces are brushlike and are known as brush discharges. *See* ELECTRO-STATICS. A. G. Bailey

Bibliography. M. Glor, *Electrostatic Hazards in Powder Handling*, 1988; D. M. Taylor and P. E. Secker, *Industrial Electrostatics: Fundamentals and Measurements*, 1994.

Static var compensator

A thyristor-controlled (hence static) generator of reactive power, either lagging or leading, or both. The word var stands for volt ampere reactive, or reactive power. The device is also called a static reactive compensator.

Need for reactive compensation. Reactive power is the product of voltage times current where the voltage and current are 90° out of phase with one another. Thus, reactive power flows one way for one-quarter of a cycle, the other way for the next quarter of a cycle, and so on (in contrast to the real power, or active power, which flows in one direction only). This back-and-forth flow results in no net power being delivered by the generator to the load. However, current associated with reactive power does flow through the conductor and creates extra losses. *See* ALTERNATING-CURRENT CIRCUIT THEORY; ELECTRIC POWER MEASUREMENT; VOLT-AMPERE.

Most loads, particularly motor loads, draw lagging reactive power along with active power. Increased lagging reactive power causes electric power system voltage to sag, and under heavy load conditions the voltage can decrease below desirable levels. (Utilities usually hold system voltage to within $\pm 5\%$ of the nominal value.) On the other hand, under light loads; the capacitance of high-voltage lines can create excessive leading reactive power, causing the voltage at some locations to rise above the nominal value.

Finally, it is prudent to keep reactive power flows to a minimum in order to allow the lines to carry more active power. For example, cables have very high capacitance to ground (compared to overhead lines), and the resulting leading reactive power loads a 10–20-mi (15–30-km) transmission cable to its thermal limit when voltage is applied, leaving it with no ability to carry real power.

Mechanical versus static compensation. Utilities frequently install capacitors connected from line to ground to compensate for lagging reactive power and reactors connected from line to ground to compensate for leading reactive power. These reactors and capacitors are switched in and out with mechanical switches based on the level of line loading as it varies throughout the day. However, frequent operation of these mechanical switches may reduce their reliability.



Fig. 1. One phase of a static var compensator (thyristorcontrolled reactor). (a) Circuit diagram of thyristorcontrolled reactor in the electric power system. (b) Voltage and current waveforms with partial conduction. (c) Waveforms with greater, but still partial, conduction. (d) Waveforms with full conduction.

It is desirable to have a controllable source of reactive power (leading or lagging); and the static var compensator, controlled with static switches, called thyristors, for higher reliability, fulfills this function. It is more expensive than mechanically switched capacitors and reactors (due to the cost of thyristor valves and associated equipment), and hence its use is based on an economic trade-off of benefits versus cost. *See* SEMICONDUCTOR RECTIFIER.

Operation. Figure 1*a* shows one phase of a typical static var compensator (thyristor-controlled reactor). Lagging reactive power is produced by controlling the current flow through a reactor with thyristors.

By controlling the start of conduction of thyristor 1, current is allowed to flow at the desired moment, say x_1 (Fig. 1*b*). The current rises, reaches its peak at the voltage zero, and then falls back to zero in about the same time period that it took to reach its peak. Thyristor 2 similarly controls the current flow in the opposite direction. More current and hence an increasing amount of lagging reactive power can be allowed to flow by firing the thyristors at an earlier time (Fig. 1*c*). Maximum current will flow when the thyristors are fired 90° before the respective voltage zero (Fig. 1*d*). Thus, lagging reactive power can be rapidly controlled by the appropriate firing of thyristors. Control circuits, including firing circuits and sophisticated controls, vary the firing time. Since power systems are three-phase, three single-phase compensators must be constructed (**Fig. 2**).

Voltage. Since thyristors for the high voltages of transmission lines can be expensive, it is customary to provide a step-down transformer to lower the voltage to about 10–20 kV for connection of a static var compensator. Often a power transformer connected to a transmission circuit is provided with a tertiary winding of low voltage, to which the static var compensator can be connected. The relatively low voltage capability of thyristors requires connecting several of them in series. *See* TRANSFORMER.

Use of filters. Except when the thyristors are fired at 90° for maximum current, the current is not a pure 60-Hz sine wave, but rather contains undesirable harmonics of the 60-Hz component. Filters must be provided to eliminate most of these harmonics (Fig. 2). *See* ELECTRIC FILTER.

Since these filters have large capacitors, they also become a source of leading reactive power. Thus, the required rating of the controlled reactor is correspondingly increased to compensate for them. If smooth reactive power control over a wide range of leading and lagging reactive power is desirable, then additional mechanically switched capacitors (Fig. 1) and a larger controlled reactor are provided.

If mechanically controlled capacitors are not desirable because of the need for frequent operation, the capacitors can be controlled with thyristors (Fig. 2). However, because the current in the



Fig. 2. Complete three-phase static var compensator, constructed from three singlephase compensators.

capacitors cannot be controlled as conveniently as in the reactors, several small capacitor banks are provided. Narain G. Hingorani

Bibliography. R. Hauth and R. Moran, *Static VAR Systems for Voltage and VAR Control*, IEEE Tutorial 78EH0135-4-PWR, IEEE Summer PAS Meeting, Los Angeles, July 1978; Institute of Electrical and Electronics Engineers, Power Systems Engineering Committee, *Application of Static VAR Systems for System Dynamic Performance* (symp.), IEEE Spec. Doc. 87TH0187-5-PWR, 1987; *Proceedings of the Transmission Static VAR Systems Seminar*, Elec. Power Res. Inst. Spec. Rep. EL-1047-SR, April 1979; *Transmission System Static Var Con trol*, Elect. Power Res. Inst. EL-2754 (final report), Dec. 1982.

Statics

The branch of mechanics that describes bodies which are acted upon by balanced forces and torques so that they remain at rest or in uniform motion. This includes point particles, rigid bodies, fluids, and deformable solids in general. Static point particles, however, are not very interesting, and special branches of mechanics are devoted to fluids and deformable solids. For example, hydrostatics is the study of static fluids, and elasticity and plasticity are two branches devoted to deformable bodies. Therefore this article will be limited to the discussion of the statics of rigid bodies in two and three spatial dimensions. *See* BUOYANCY; ELASTICITY; HYDROSTAT-ICS; MECHANICS.

Mechanics is the study of motions in terms of mass, length, time, and forces. In statics the bodies being studied are in equilibrium. Positions of points in space, velocities, forces, and torques are all vector quantities, since each has direction, magnitude, and units. Beams, bridges, machine parts, and so on are not really rigid, but whenever the largest change in length of a portion of a body Δl is much smaller than the length *l*, that is, $l > \Delta l$, then it is a satisfactory approximation to treat that body as a rigid body. Many of the objects used in architecture, engineering, and physics can be satisfactorily idealized as rigid bodies, and much of building, machine, bridge, and dam design is based upon the study of statics.

The equilibrium conditions are very similar in the planar, or two-dimensional, and the threedimensional rigid-body statics. These conditions are that the vector sum of all forces acting upon the body must be zero, and the resultant of all torques about any point must be zero. Thus it is necessary to understand the vector sums of forces and torques.

Force as a vector. The physical effect of a force is a push or pull on an object at its point of application. The effect of a force is to change the velocity of the body. The SI units of force are newtons (1 N = 1 kilogram-meter per second squared). The vector nature of forces is expressed by their direction in space. This direction is the direction of the push or pull exerted by the force. The notation for a



Fig. 1. Directed line \overrightarrow{AB} or \overrightarrow{P} is a vector quantity in the plane.

vector quantity is an overarrow, for example, \vec{F} or \vec{P} . In **Fig. 1** a two-dimensional vector \vec{P} is shown as a directed line segment \vec{AB} in the plane of the paper. A three-dimensional vector \vec{P} is shown in **Fig. 2** as the directed line segment \vec{OD} . A complete description of a vector is obtained by specifying a coordinate system and giving its components in that coordinate system. The sum of any two vectors $\vec{A} + \vec{B}$ is the sum of their components, so that \vec{A} and \vec{B} must have the same number of components.

Superposition and transmissibility. In studying statics problems (and general mechanics problems), two principles, superposition and transmissibility, are used repeatedly on force vectors. They are applicable to all vectors, but specifically to forces and torques (first moments of forces).

1 The principle of superposition of *d*-dimensional vectors is that the sum of any two *d*-vectors is another *d*-vector. Of course, some or all components can be zero. This principle is illustrated in **Fig. 3**. The two vectors labeled Q, which can be shown applied to a rigid structural member D, add up to zero, and so do the two labeled R. Superposition applies to all sums of vectors, not just sums which vanish.

2 The principle of transmissibility of a force applied to a rigid body is that the same mechanical effect is produced by any shift of the application of the force. The principle of transmissibility is illustrated in Fig. 3. The solid rectangle C represents a solid structural member, and the line AB is the line



Fig. 2. Rectangular components of a force vector \vec{P} in three-dimensional space.



Fig. 3. Illustration of two theorems of statics.



Fig. 4. Moments about the axis in rectangular coordinates.

of action for the equivalent forces *P*. Since these two forces have the same magnitudes and direction and are applied along the same line, by transmissibility, all of their mechanical effects are equivalent. To use the superposition principle to add two vectors, the principle of transmissibility is used to move some vectors along their line of action in order to add to their components.

Components of a force. Figure 1, the twodimensional example, shows the construction by which orthogonal force components are defined. In this figure \vec{P} or \vec{AB} is a force vector, and *L* a directed line whose positive sense is toward its labeled end. Construction lines *AC* and *BD* are in planes (not shown) normal to *L*, and θ is the direction angle of \vec{P} relative to *L*; it is a plane angle between the directions in the positive senses of \vec{P} and *L*; further, $0 < \theta < 180^{\circ}$.

The orthogonal vector component of force \vec{P} on directed line *L* is a force of direction and magnitude given by P_L or *CD*, where P_L is in the direction of *L*. Its magnitude is given by Eq. (1).

$$P_L = \overline{CD} = \overline{AE} = \overline{AB} |\cos\theta| = P |\cos\theta| \qquad (1)$$

The component of \vec{P} on *L* is $P_L = P \cos \theta$, where P_L is positive if $0^\circ \ge 90^\circ < 90^\circ$ and negative if $90^\circ < \theta \le 180^\circ$. The absolute magnitude of P_L is designated $|P_L|$.

The rectangular components of a force are its components on mutually perpendicular lines. In Fig. 2, \vec{P} or \vec{OA} , \vec{P} or \vec{OB} , and \vec{P} or \vec{OC} are the rectangular vector components of \vec{P} or \vec{OD} in the directions of lines (axes) X, Y, and Z, respectively.

The corresponding components have the magnitudes $P_X = P \cos \theta_X$, $P_Y = P \cos \theta_Y$, and $P_Z =$ $P \cos \theta_Z$, and are in the *X*, *Y*, *Z* directions, respectively. The magnitude of the three-dimensional vector \vec{P} is given by Eq. (2).

$$P = \sqrt{P_X^2 + P_Y^2 + P_Z^2}$$
(2)

Moment of a force. The moment of a force about a directed line is a signed number whose value can be obtained by applying these two rules:

1 The moment of a force about a line parallel to the force is zero.

2 The moment of a force about a line normal to a plane containing the force is the product of the magnitude of the force and the least distance from the line to the line of the force. Conventionally, the moment is positive if the force points counterclockwise about the line as viewed from the positive end of the line.

In **Fig.** 4 the moments of \vec{P} about the *X*, *Y*, and *Z* coordinate axes are $M_X = bP_Z - cP_Y$ about the *X* axis; $M_Y = cP_X - aP_Z$ about the *Y* axis; and $M_Z = aP_Y - bP_X$ about the *Z* axis. The resultant of the two torques or moments of force, \vec{T}_1 and \vec{T}_2 about a point *P*, is shown as \vec{T}_r in **Fig. 5**. The components of these torques are due to forces \vec{F}_1 and \vec{F}_2 oriented with respect to \vec{T} as shown.

Figure 6 illustrates cases where the forces in a structure may or may not depend on the relative strength of its parts. Case 6a is called statically determinate, and case 6b statically indeterminate. The terms are actually misnomers, for the forces can be found in both cases. Better terminology would be rigidly determinable and indeterminable. In case 6b, if each member were rigid (infinitely strong), the forces in each could not be determined.

In summary, the statics of rigid bodies in statically determinate structures is carried out by summing vector forces and vector torques and setting the



Fig. 5. Resultant of two torques.



Fig. 6. Statically determinate and indeterminate structures. (a) The forces in members A and B are independent of their strengths (for small deformations). (b) The relative strength of the added member D will determine how much of the load W it supports.

resultants equal to zero. *See* COUPLE; EQUILIBRIUM OF FORCES; FORCE; TORQUE. Brian De Facio

Bibliography. F. P. Beer and E. Russell Johnston, Vector Mechanics for Engineers: Statics, 6th ed., 1996; W. L. Fowler and A. Bedford, Engineering Materials: Statics, 2d ed., 1998; R. C. Hibbeler, Engineering Mechanics: Statics, 8th ed., 2000; D. J. McGill and W. W. King, Engineering Mechanics: An Introduction to Statics, 3d ed., 1995; J. L. Meriam and L. G. Kraige, Engineering Mechanics: SI-Englisb Version, vol. 1: Statics, 3d ed., 1992; A. Pytel and K. Kiusalaas, Engineering Mechanics: Statics, 1994.

Statistical mechanics

That branch of physics which endeavors to explain the macroscopic properties of a system on the basis of the properties of the microscopic constituents of the system. Usually the number of constituents is very large. All the characteristics of the constituents and their interactions are presumed known; it is the task of statistical mechanics (often called statistical physics) to deduce from this information the behavior of the system as a whole.

Scope of Statistical Mechanics

Elements of statistical mechanical methods are present in many widely separated areas in physics. For instance, the classical Boltzmann problem is an attempt to explain the thermodynamic behavior of gases on the basis of classical mechanics applied to the system of molecules. Historically, this was the first systematic investigation of a statistical problem, and many of the procedures and methods originate from this investigation. It is important to realize that statistical mechanics gives more than just an explanation of already known phenomena. By using statistical methods, it often becomes possible to obtain expressions for empirically observed parameters, such as viscosity coefficients, heat conduction coefficients, and virial coefficients, in terms of the forces between molecules. This kind of result, a direct relation between an observed macroscopic entity and an intermolecular potential, constitutes one of the main achievements of statistical physics. See BOLTZMANN

STATISTICS; INTERMOLECULAR FORCES; KINETIC THE-ORY OF MATTER.

Statistical considerations also play a significant role in the description of the electric and magnetic properties of materials. In this case the objective is to deduce such characteristics as the permittivity, electrical conductivity, and magnetic permeability from the known properties of the atom. To obtain the Maxwell equations, which are macroscopic equations for matter in the bulk, from the Lorentz (electron) equations is typically a statistical question; new information which can be obtained from this approach includes, for example, the temperature dependence of the permittivity. *See* ELECTRICAL CON-DUCTIVITY OF METALS; MAGNETIC SUSCEPTIBILITY; MAXWELL'S EQUATIONS; PERMITTIVITY; RELATIVISTIC ELECTRODYNAMICS.

If the problem of molecular structure is attacked by statistical methods, the contributions of internal rotation and vibration to thermodynamic properties, such as heat capacity and entropy, can be calculated for models of various proposed structures. Comparison with the known properties often permits the selection of the correct molecular structure. The statistical description of the activated complex in reaction mechanisms has enabled chemists to formulate a theory of absolute reaction rates. *See* MOLECULAR STRUCTURE AND SPECTRA.

Perhaps the most dramatic examples of phenomena requiring statistical treatment are the cooperative phenomena or phase transitions. In these processes, such as the condensation of a gas, the transition from a paramagnetic to a ferromagnetic state, or the change from one crystallographic form to another, a sudden and marked change of the whole system takes place. The appropriate description of such processes is one of the most difficult but most interesting problems of statistical physics. *See* PHASE TRANSITIONS.

Statistical considerations of quite a different kind occur in the discussion of problems such as the diffusion of neutrons through matter. In this case, the probability of the various events which affect the neutron are known, such as the capture probability and scattering cross section. The problem here is to describe the physical situation after a large number of these individual events. The procedures used in the solution of these problems are very similar to, and in some instances taken over from, kinetic considerations. Similar problems occur in the theory of cosmic-ray showers. Here the probabilities of the basic processes-pair production, annihilation, ionization, meson production-are all presumed known. The major task is the computation of the combined effect of a large number of these individual events. Special techniques of statistical physics, the use of distribution functions and transport-type equations, find extensive applications in these areas. See BOLTZMANN TRANSPORT EQUA-TION; COSMIC RAYS; REACTOR PHYSICS.

It happens in both low-energy and high-energy nuclear physics that a considerable amount of energy is suddenly liberated. An incident particle may

be captured by a nucleus, or a high-energy proton may collide with another proton. In either case, there is a large number of ways (a large number of degrees of freedom) in which this energy may be utilized. In the nuclear case, there are usually many decay and excitation modes; in the case of the proton-proton collision, there is enough energy available for the creation of a number of mesons. To survey the resulting processes, one can again invoke statistical considerations. The statistical problem here is to calculate the probability that, given a total amount of energy, a particular process will occur. Of course, the statistical method cannot help in the investigation of the actual mechanisms of these processes. However, the statistical factors must be considered in the interpretation of experiments. See NUCLEAR REACTION; SCATTERING EXPERIMENTS (NUCLEI).

Of considerable importance in statistical physics are the random processes, also called stochastic processes or sometimes fluctuation phenomena. The brownian motion, the motion of a particle moving in an irregular manner under the influence of molecular bombardment, affords a typical example. The process may be described in terms of a fluctuating force acting on a particle, perhaps in addition to a systematic force. The interest in this problem centers around a calculation of the position and velocity of a particle after it has experienced many collisions or after the fluctuating force has acted for a long time. The stochastic processes are in a sense intermediate between purely statistical processes, where the existence of fluctuations may safely be neglected, and the purely atomistic phenomena, where each particle requires its individual description. All statistical considerations involve, directly or indirectly, ideas from the theory of probability of widely different levels of sophistication. The use of probability notions is, in fact, the distinguishing feature of all statistical considerations. See BROWNIAN MOVE-MENT; PROBABILITY; PROBABILITY (PHYSICS); STATIS-TICS; STOCHASTIC PROCESS.

Methods of Statistical Mechanics

In the following sections procedures used in statistical mechanics are discussed.

Phase space. For a system of N particles, each of mass m, contained in a volume V, the positions of the particles may be labeled $x_1, y_1, z_1, \ldots, x_N$, y_N , z_N , their cartesian velocities v_{x1} , ..., v_{zN} , and their momenta p_{x_1}, \ldots, p_{z_N} . The simplest statistical description concentrates on a discussion of the distribution function $f(x,y,z;v_x,v_y,v_z;t)$. The quantity $f(x,y,z;\upsilon_x,\upsilon_y,\upsilon_z;t)$ ($dxdydzd\upsilon_xd\upsilon_yd\upsilon_z$) gives the (probable) number of particles of the system in those positional and velocity ranges where x lies between x and x + dx; v_x between v_x and $v_x + dv_x$, and so on. These ranges are finite. The six-dimensional space defined by x, y, z, v_x , v_y , v_z is called the μ space, and the behavior of the gas is represented geometrically by the motion of N points in the μ space. It was shown by L. Boltzmann that in the course of time f approaches an equilibrium distribution f^0 as in Eq. (1). Here $d^3x = dxdydz$, A and β are parameters,

$$f^{0}(\mathbf{x}, \mathbf{v}) d^{3}x d^{3}v$$

= $A \exp\left[-\frac{1}{2}\beta m \mathbf{v}^{2} - \beta U(x, y, z)\right] d^{3}x d^{3}v$ (1)

and U(x,y,z) is the potential energy at the point x, y, z. Boltzmann also could interpret Eq. (1) as the most probable distribution.

Actually, important as the use of the distribution function is in practice, there are in principle serious limitations and difficulties associated with it. The limitations refer to the fact that the description by means of a single distribution function is correct only for noninteracting particles; also, the neglect of triple and higher collisions restricts the applicability to very dilute systems. The indiscriminate use of the ideas of Boltzmann leads to paradoxical results.

In the further study of these questions, the phase space of a dynamical system plays an important role. This is the 6N-dimensional euclidean space, whose 6N axes are given in notation (2). The state of the

$$x_1, y_1, z_1, \dots, x_N, y_N, z_N; p_{x_1}, \dots, p_{z_N}$$
 (2)

system at a given time is completely specified by these 6N numbers. These numbers define a point (the representative point, or phase point) in the 6Ndimensional phase space. In the course of time the variables x_1, \ldots, p_{z_N} change, hence the phase point moves. If the system is conservative, Eq. (3) defines, for the energy *E*, a (6N - 1)-dimensional surface in

$$\sum_{i} \frac{\mathbf{p}_i^2}{2m_i} + U(x_1, \dots, z_N) = E$$
(3)

the phase space called the energy surface. For a conservative system, the phase point in the course of time wanders over the energy surface. Intuitive as this geometrical representation of an involved mechanical system might seem, the actual trajectory of the phase point is extremely complicated. Also, the use of the phase space does not by itself imply any statistical considerations. A knowledge of the trajectory is precisely equivalent to a solution of the complete mechanical problem, for it determines each pand x as a function of time. However, many different microscopic states all correspond to the same macroscopic situation.

Observations made on a system always require a finite time; during this time the microscopic details of the system will generally change considerably as the phase point moves. The result of a measurement of a quantity Q will therefore yield the time average, as in Eq. (4). The integral is along the tra-

$$\overline{Q}_t = \frac{1}{t} \int_0^t Q \, dt \tag{4}$$

jectory in phase space; Q depends on the variables x_1, \ldots, p_{zN} and t. To evaluate the integral, the trajectory must be known, and as mentioned above, this requires the solution of the complete mechanical problem.

For many years one of the major problems in statistical mechanics was precisely to recast the expression for a time average in a more tractable form. Generally attempts were made to replace time averages by phase space averages, by trying to show that the time spent in a region was proportional to the size of the region. The legitimacy of this procedure was studied extensively by mathematicians, but although these studies opened up significant areas in mathematics, they did not really help elucidate problems of physical interest.

The subtlety of the problems involved is illustrated by the so-called ergodic theorem of G. D. Birkhoff, which asserts that $\lim_{t\to\infty} \overline{Q}_t$ exists for almost all trajectories, although this limit varies discontinuously from trajectory to trajectory. The Poincaré recurrence theorem, which states that a system within a finite time will return as closely as desired to its initial state, is of special interest in statistical mechanics. Apart from providing the mathematical framework for these theorems, the phase space provides also the appropriate framework for the formulation of ensemble theory, which is the basis of modern statistical mechanics.

Ensembles; Liouville's theorem. J. Willard Gibbs first suggested that instead of calculating a time average for a single dynamical system, a collection of systems, all similar to the original one, should instead be considered. Such an ensemble of systems is to be constructed in harmony with the available knowledge of the single system, and may be represented by an assembly of points in the phase space, each point representing a single system. If, for example, the energy of a system is precisely known, but nothing else, the appropriate representative example would be a uniform distribution of ensemble points over the energy surface, and no ensemble points elsewhere. An ensemble is characterized by a density function $\rho(x_1, \ldots, z_N; p_{x_1}, \ldots, p_{z_N}; t) \equiv \rho(x, p, t)$. The significance of this function is that the number of ensemble systems dN_e contained in the volume element $dx_1 \dots dz_N$, $dp_x \dots dp_{z_N}$ of the phase space (this volume element will be called $d\Gamma$), at time *t* is as given in Eq. (5).

$$\rho(x, p, t)d\Gamma = dN_e \tag{5}$$

The ensemble average of any quantity Q is given by Eq. (6). The basic idea now is to replace the time

$$\overline{Q}_{\rm ens} = \frac{\int Q\rho \ d\Gamma}{\int \rho \ d\Gamma} \tag{6}$$

average of an individual system by the ensemble average, at a fixed time, of the representative ensemble. Stated formally, the quantity \overline{Q}_t defined by Eq. (4), in which no statistics is involved, is identified with \overline{Q}_{ens} defined by Eq. (6), in which probability assumptions are explicitly made. Another form of this same connection between the behavior of the individual system and the ensemble is that the probability that the individual system at time *t* will be in a region *R* of the phase space is given by the fraction of ensemble systems contained in R at time t, Eq. (7).

$$P(R,t) = \frac{\int_{R} \rho \, d\Gamma}{\int \rho \, d\Gamma} \tag{7}$$

It is clear, on the basis of these relations, that the complete statistical behavior of a system is known once its representative ensemble has been obtained. In the construction of such ensembles, it is always assumed that accessible parts of the phase space should be weighted equally. There is one other general requirement that the density function must satisfy. Inasmuch as the number of ensemble members remains constant (no ensemble members are created or destroyed in the course of time), there is a continuity equation, Eq. (8), for the density

$$\frac{\partial \rho}{\partial t} + \sum_{i=1}^{3N} \left[\frac{\partial}{\partial x_i} \left(\rho \frac{dx_i}{dt} \right) + \frac{\partial}{\partial p_i} \left(\rho \frac{dp_i}{dt} \right) \right] = 0 \quad (8)$$

function ρ which simply states that the change in the number of systems in a volume of phase space per unit time equals the difference of the number of systems flowing in and flowing out of that volume. If the time derivatives dx_i/dt and dp_i/dt are now expressed through the hamiltonian equations, Eq. (9) is obtained. Here $d\rho/dt$ is the substantial,

$$\frac{\partial \rho}{\partial t} + \sum_{i=1}^{3N} \left(\frac{\partial \rho}{\partial x_i} \frac{dx_i}{dt} + \frac{\partial \rho}{\partial p_i} \frac{dp_i}{dt} \right) \equiv \frac{d\rho}{dt} = 0 \qquad (9)$$

or total derivative. *See* EQUATION OF CONTINUITY; HAMILTON'S EQUATIONS OF MOTION.

Equation (9), called the Liouville theorem, asserts that the time rate of change of ρ is zero along a streamline in phase space. It can also be deduced from Eq. (9) that the ensemble systems move in the course of time in a volume-preserving fashion through the phase space. If at some initial time a set of ensemble points occupy a volume V_0 , then in the course of time each representative point will move in a complicated fashion. However, at time t the volume V_t made up by the ensemble points which initially were in V_0 is still equal to V_0 . The shape of the volume changes, but the actual volume is constant. In the attempts which have been made to justify the postulate that $\overline{Q}_t = \overline{Q}_{ens}$, this volume-preserving property plays a dominant role. The proof of the Poincaré recurrence theorem also depends crucially on this property. Yet in spite of the general significance of the Liouville theorem, Eq. (9) generally does not help in the task of setting up an appropriate representative ensemble. (It is relevant, however, to the subsequent discussion on nonequilibrium theory of liquids.) The choice of the ensemble is usually determined by physical considerations and by the already mentioned postulate of equal a priori probabilities in phase space. Perhaps most important is the a posteriori justification of the statistical procedures, obtained through a comparison with the experimental facts. It is important that inasmuch as the methods are statistical, not only average values but also fluctuations around these average values can be computed. It turns out that usually these fluctuations are

extremely small; however, in those instances in which they are appreciable, they can be compared with experiments, providing additional support for the statistical procedures. Hence, despite the fact that no completely rigorous mathematical justification of the ensemble methods exists, these methods are physically so plausible and lead to such agreement with experiments for a variety of systems that they reasonably must be considered as a wellestablished part of statistical physics.

Relation to thermodynamics. It is certainly reasonable to assume that the appropriate ensemble for a thermodynamic equilibrium state must be described by a density function which is independent of the time, since all the macroscopic averages which are to be computed as ensemble averages are time-independent. Thus an equilibrium ensemble is realized when $\partial \rho / \partial t = 0$. In that case it follows from Liouville's theorem that Eq. (10) holds. It is easy to

$$\sum_{i=1}^{3N} \left(\frac{\partial \rho}{\partial x_i} \frac{dx_i}{dt} + \frac{\partial \rho}{\partial p_i} \frac{dp_i}{dt} \right) = 0$$
(10)

see that if ρ is a function of a quantity α (which is a function of *x* and *p* and which is a constant of the motion, such as the energy, so that $\partial \alpha / \partial t = 0$), then any $\rho(\alpha)$ satisfies Eq. (10). Hence any density function ρ which is a function of *x* and *p* through the dependence of the energy *E* on *p* and *x* satisfies the Liouville equation (9). The functional form of $\rho(E)$ is left completely unspecified.

Microcanonical ensemble. In the case of an isolated system, where the energy is specified within a well-defined range, the most obvious example to use would be the so-called microcanonical ensemble defined by Eq. (11a), where *c* is a constant, for the en-

$$\rho(p, x) = c \tag{11a}$$

$$\rho(p, x) = 0 \tag{11b}$$

ergy *E* between E_0 and $E_0 + \Delta E$; for other energies Eq. (11*b*) holds. By using Eq. (6), any microcanonical average may be calculated. The calculations, which involve integrations over volumes bounded by two energy surfaces, are not trivial. Still, many of the results of classical Boltzmann statistics may be obtained in this way.

Canonical ensemble. For applications and for the interpretation of thermodynamics, the canonical ensemble is much more preferable. This ensemble describes a system which is not isolated but which is in thermal contact with a heat reservoir. By describing the complete system (system plus reservoir) by a microcanonical ensemble, it may be shown that the system itself may be represented by Eq. (12*a*). Here

$$\rho(x, p) = N_e \exp\left[\frac{\psi - E(x, p)}{\theta}\right]$$
(12*a*)

$$\int \rho(x,p) \, d\Gamma = N_e \tag{12b}$$

 N_e is the total number of ensemble systems, E(x,p) is

the mechanical energy of the system, and ψ and θ are parameters independent of x and p characterizing the canonical ensemble. From the fact that Eq. (12b) is valid, it follows that the two parameters ψ and θ are not independent, but that Eq. (13) applies.

$$e^{-(\psi/\theta)} = \int d\Gamma \exp\left[-\frac{E(x,p)}{\theta}\right] \equiv Z$$
 (13)

The quantity Z is usually called the partition function. It is customary to define Z for a system of N identical particles with a different multiplicative constant by Eq. (14). Here b is the Planck constant. It is now

$$Z' = \frac{1}{N! b^{3N}} \int \exp\left[-\frac{E(p, x)}{\theta}\right] d\Gamma \qquad (14)$$

very important that the parameters θ and ψ be identified with definite thermodynamic functions. This may be done by considering some of the thermodynamic relations. If η is the entropy, ϵ the thermodynamic energy (internal energy), and ψ' the free energy (Helmholtz free energy), then some examples of thermodynamic relations are, for gases, described by pressure, volume, and temperature (*P*, *V*, *T*) as in Eqs. (15).

1

$$\epsilon' = \epsilon - T\eta$$
 (15*a*)

$$d\psi' = -\eta \, dT - P \, dV \tag{15b}$$

$$\frac{\partial \psi'}{\partial T} = -\eta \qquad \frac{\partial \psi'}{\partial V} = -P \qquad (15c)$$

$$\epsilon = \psi' - T \frac{\partial \psi'}{\partial T} \tag{15d}$$

See ENTROPY; FREE ENERGY; INTERNAL ENERGY; THERMODYNAMIC PRINCIPLES.

From Eqs. (12*a*), (6), and (13) the average energy \overline{E} may be calculated, Eq. (16). For an ideal gas, in which the energy of a molecule is $\mathbf{p}^2_i/2m$, Eq. (17)

 \overline{E}

$$=\theta^2 \frac{\partial}{\partial \theta} \log Z \tag{16}$$

$$Z = \int \cdots \int dx_1 \cdots dp_{z_N} \exp\left(-\frac{1}{2m\theta} \sum_{i=1}^N \mathbf{p}_i^2\right)$$
$$= V^N (2\pi \, m\theta)^{3N/2}$$
(17)

holds. (The evaluation of the integrals is elementary; the space integrations merely contribute V^N , and the momentum integrals are gaussian.) Equation (18*a*) is

$$\overline{E} = \frac{3N}{2}\theta \tag{18a}$$

$$\epsilon = {}^3/{}_2NkT \equiv {}^3/{}_2RT \tag{18b}$$

then obtained as the energy of an ideal gas, having N molecules in a vessel of any volume V. Now it is known by experiment that the thermodynamic energy of 1 mole of an ideal gas is given by Eq. (18*b*). Here R is the ideal gas constant, T is the absolute temperature, N is Avogadro's number, and k is the Boltzmann constant. Comparison of Eqs. (18*a*) and (18*b*) leads to the identification of $\theta = kT$. It is also

easy to deduce the Maxwell-Boltzmann distribution. Applying Eq. (7) yields Eq. (19) for the probability

$$P(\mathbf{p}_{1}\mathbf{x}_{1}) d^{3}p_{1}d^{3}x_{1}$$

$$= \frac{\int \cdots \int d_{3}x_{1} \ d_{3}p_{1} \ d^{3}x_{2} \cdots d^{3}p_{N}e^{-E/\theta}}{\int \cdots \int d^{3}x_{1} \cdots d^{3}p_{N}e^{-E/\theta}}$$

$$= \frac{1}{V}e^{-\mathbf{p}^{2}/(2m\theta)}(2m\theta)^{-3/2} \ d^{3}p_{1}d^{3}x_{1}$$
(19)

that molecule 1 will have a given range of positions and momenta. The canceled symbols in the numerator mean that the coordinates and momenta of the molecule 1 are not integrated over. The result given by Eq. (19) is again true for an ideal gas. The identification of $\theta = kT$, however, is made in general. By differentiating Eq. (13) with respect to θ , it is possible to establish from Eq. (18a) the general connection (using $\theta = kT$) written as Eq. (20). When this general

$$\overline{E} = \psi - T \frac{\partial \psi}{T} = \epsilon \tag{20}$$

relation, deduced from the canonical ensemble, is compared with the general thermodynamic relation (15*d*), the parameter θ and the free energy θ' are indeed seen to play identical roles. Again, for an ideal gas, ψ can be computed, and it is the same as the ideal gas free energy. Hence ψ and ψ' are identified in general. Equation (13) relates the thermodynamic free energy to an integral containing the mechanics of the microscopic problem. It is also interesting to observe that a quantity η' defined by Eq. (21) plays in

$$\eta' = \frac{\int \rho \log \rho \, d\Gamma}{\int \rho \, d\Gamma} \tag{21}$$

all instances the role of the entropy. In fact, it can be shown that the equation $\psi = \epsilon - T\eta'$ is in harmony with Eq. (15a). These thermodynamic analogies can be followed through in all detail to see that the canonical ensemble combined with the relevant definitions does indeed reproduce the formal aspects of thermodynamics. The average energy \overline{E} for a canonical ensemble has already been calculated. It is important to know just how often appreciable deviations from this average energy can be expected. For this, the fractional fluctuations in energy given by notation (22) must be calculated. The definition of \overline{E}^2 , in Eq. (23), follows again from Eq. (6). From Eq. (23)

$$\frac{\overline{E}^2 - (\overline{E})^2}{(\overline{E})^2} \tag{22}$$

4

$$\overline{E}^{2} = \frac{\int E^{2} e^{-E/\theta} d\Gamma}{\int e^{-E/\theta} d\Gamma}$$
(23)

the quantity \overline{E}^2 may be expressed directly in terms of Z, the partition function. For an ideal gas it is then possible to obtain, using Eq. (17), an explicit expression, Eq. (24), for the fluctuations in energy. Since the

$$\left(\frac{\overline{E}^2 - (\overline{E})^2}{(\overline{E})^2}\right)^{1/2} = \sqrt{\frac{2}{3N}}$$
(24)

number of particles N is about 10^{22} in systems dealt

with in practice, the chance of observing a sizable deviation from the average energy is extremely small. Using these same methods, defining in addition the specific heat $C_v = \partial \overline{E} / \partial T$, it may be shown generally that the fractional fluctuation in energy is given by Eq. (25). Even though the energy fluctuations are

$$\left(\frac{\overline{E}^2 - (\overline{E})^2}{(\overline{E})^2}\right)^{1/2} = \left(\frac{kT^2C_v}{(\overline{E})^2}\right)^{1/2}$$
(25)

negligible for an ideal gas, this is not always so for other systems. For solids at low temperatures, the fluctuations may be appreciable. When phase transitions (of the first order) take place, C_{ν} becomes infinite, indicating via Eq. (25) that the fluctuations become very large. It is clear that the main problem of the applications is reduced to the calculation of the partition function Z; all thermodynamic entities follow from Eqs. (16) and (13).

Grand canonical ensemble. There is yet another ensemble which is extremely useful and which is particularly suitable for quantum-mechanical applications. Much work in statistical mechanics is now based on the use of this so-called grand canonical ensemble. The grand ensemble describes a collection of systems; the number of particles in each system is no longer the same, but varies from system to system. The density function $\rho(N,p,x) d\Gamma_N$ gives the probability that there will be in the ensemble a system having N particles, and that this system, in its 6N-dimensional phase space Γ_N , will be in the region of phase space $d\Gamma_N$. The function ρ is given by Eq. (26). Here θ , Ω , and μ are the parameters char-

$$\rho(N, p, x) = \exp\left(\frac{\Omega + \mu N - E(p, x)}{\theta}\right)$$
(26)

acterizing the ensemble, just as ψ and θ characterize the canonical ensemble. It is again true that these parameters are directly related to thermodynamic state functions. The detailed argument follows the identical pattern indicated in the discussion of the canonical ensemble. The normalization condition is now given by Eq. (27). The grand canonical average of

$$\sum_{N} \int d\Gamma_{N} \rho(N, p, x)$$
$$= \sum_{N} \int d\Gamma_{N} \exp\left(\frac{\Omega + \mu N - E(p, x)}{\theta}\right) = 1 \quad (27)$$

any quantity Q is given by Eq. (28) and the grand partition function by Eq. (29a). Again Eq. (29b) is

$$\overline{Q}_{\rm gr} = \sum_{N} \int d\Gamma_N \rho(N, p, x) \, Q_N(p, x) \tag{28}$$

$$Z_{\rm gr} = \sum_{N} e^{\mu N/\theta} \int d\Gamma_N e^{-(E/\theta)}$$
(29*a*)

$$Z_{\rm gr} = \sum_{N} \frac{1}{N!} \frac{1}{b^{3N}} e^{\mu N/\theta} \int d\Gamma_N e^{-(E/\theta)} \qquad (29b)$$

customarily used for a system consisting of N identical particles. From Eq. (27) it follows immediately

that Eq. (30) holds. From Ω , sometimes called the

$$Z_{\rm gr} = e^{-(\Omega/\theta)} \tag{30}$$

grand potential, all other thermodynamic functions may be computed. The parameter μ is the chemical potential. It is defined by $\mu = (\partial \xi / \partial N)_{R,T}$ where ξ is the Gibbs free energy or thermodynamic potential; $\xi = \psi + PV$. Formally, these results are written as Eqs. (31). From a knowledge of the grand partition

$$P = -\left(\frac{\partial\Omega}{\partial V}\right)_{\mu,T} \tag{31a}$$

$$\eta = -\left(\frac{\partial\Omega}{\partial T}\right)_{V,\mu} \tag{31b}$$

$$\overline{N}_{gr} = -\left(\frac{\partial\Omega}{\partial\mu}\right)_{V,T}$$
(31c)

function as a function of *V*, *T*, and μ , all thermodynamic functions follow. For an ideal gas, for example, Eq. (29b) in conjunction with Eq. (17) yields Eq. (32*a*). Here λ is the thermal de Broglie wave-

$$Z_{\rm gr} = \sum_{n=1}^{\infty} \frac{1}{N!} \left(\frac{2\pi m\theta}{b^2}\right)^{(3/2)N} V^N e^{\mu N/\theta}$$
$$= \exp\left(e^{\mu/\theta} \frac{V}{\lambda^3}\right) \tag{32a}$$
$$\hbar$$

$$\lambda = \frac{\hbar}{\sqrt{2\pi mkT}} \tag{32b}$$

length, defined by Eq. (32*b*). From Eqs. (32*a*) and (30) it follows that Eq. (33) holds. Application of

$$\Omega = -\theta e^{\mu/\theta} \frac{V}{\lambda^3} \tag{33}$$

Eqs. (31) yields the usual results for ideal gases, such as $PV = \overline{N}kT$; however, the grand average of \overline{N} enters the relations now, rather than just \overline{N} . From the definitions (28) and (29*b*) it can be shown that the fractional fluctuation in the number *N* is given by Eq. (34). Thus for gases, the fluctuations in num-

$$\left(\frac{\overline{N_{\rm gr}^2} - (\overline{N}_{\rm gr})^2}{\overline{N}_{\rm gr}^2}\right)^{1/2} = \frac{1}{\sqrt{N_{\rm gr}}} \tag{34}$$

ber are negligibly small, showing that the number of particles in the grand canonical ensemble is sharply peaked around the average value. The physical results deduced from the canonical and grand canonical ensemble can therefore be expected to be the same; the calculations, however, are frequently (especially in quantum problems) simpler in the grand canonical than in the canonical scheme. *See* CHEMI-CAL EQUILIBRIUM; CHEMICAL THERMODYNAMICS.

Applications. Two of the newest and most important applications of statistical physics involve the theory of nonideal gases and the theory of nonequilibrium states of dense gases and liquids. Theory of nonideal gases. The importance of relations such as Eq. (13), which connects the thermodynamic free energy to the partition function, lies in the general validity of these connections. Specifically, when dealing with a gas of interacting molecules, where the energy is given by Eq. (35) with the potential U

$$E(x_1, \dots, z_N; p_{x_1}, \dots, p_{z_N})$$

= $\sum_{i=1}^{N} \frac{p_{x_i}^2 + p_{y_i}^2 + p_{z_i}^2}{2m} + U(x_1, \dots, z_N)$ (35)

a known function, the relation is still valid. In that case, the partition function may be obtained directly in terms of U by performing the integrations over the momenta as in Eq. (36). Since from Z_N all thermody-

$$Z_N = \frac{1}{N!} \frac{1}{\lambda^{3N}} \int \dots \int d^3 x_1, \dots, d^3 x_N$$
$$\times \exp\left[-\frac{U(x_1, \dots, z_N)}{kT}\right] \quad (36)$$

namic relations follow, Eq. (36) is the appropriate starting point for all studies of nonideal gases. The canonical ensemble has been used in obtaining Z_N , which depends on *V* through the integration limits, on *T* through λ [see Eq. (32*b*)], and on *N* through the number of integrations. Since Z_N gives the free energy immediately, the equation of state (the relation between *P*, *V*, and *T*) can be derived by computing $P = -\partial \psi / \partial V$. The experimental results on the equations of state may be expressed in terms of the so-called virial expansion, Eq. (37). Here *B*(*T*) and *C*(*T*)

$$P = \frac{RT}{V} \left(1 + \frac{B(T)}{V} + \frac{C(T)}{V^2} + \dots \right)$$
(37)

are the second and third virial coefficients; they are determined experimentally. It is clear that Eq. (37) is a development starting from the ideal gas equation. A first task for statistical mechanics is to deduce the form (37), as well as an explicit expression for B(T), and so on. A first step in carrying this out is to set up relation (38*a*), and to assume that the potential

$$\exp\left[-\frac{U(x_1,\ldots,z_N)}{kT}\right] \equiv W(1,2,\ldots,N) \quad (38a)$$

$$U = U(\alpha) + U(\beta)$$
(38b)
$$W = W(\alpha)W(\beta)$$
(38c)

$$W = W(\alpha)W(\beta)$$
(586)
$$W(1) = S(1) = 1$$

W(1, 2) = S(1, 2) + S(1)S(2)

$$W(1, 2, 3) = S(1, 2, 3) + S(1)S(2, 3) + S(2)S(1, 3)$$

$$+ S(3)S(1, 2) + S(1)S(2)S(3)$$

W(1,...,N) = S(1,...,N) (38d)

$$+ S(1)S(2,...,N)$$

+ ... + S(1, 2)S(3,...,N)
+ ... + S(1)...S(N)

 $U(x_1, \ldots, z_N)$ is such that it is possible to make an unambiguous distinction between interacting and noninteracting particles. Any potential which vanishes for a separation larger than a critical amount has this property, and any short-range potential approximates it. A separated configuration may be defined as one in which none of the molecules of one group interacts with any of another. If two such groups are called α and β , Eq. (38b) is satisfied for such a configuration. Hence, by relation (38a), Eq. (38c) holds. A set of functions S may now be defined by Eqs. (38d). From Eqs. (38c) and (38d) the following important property can be proven: the S functions vanish for a separated configuration. For instance S(1, 2, ..., l) is zero, unless the molecules 1, 2, ..., *l* form a nonseparated configuration. This is an important property, for it allows the integration of the S functions in a simple fashion. This can be illustrated by the example of integral (39a). To evaluate this integral it is useful

$$\int \cdots \int d^3x_1 \cdots d^3x_l \mathcal{S}(1, \dots, l) \tag{39a}$$

$$\int \cdots \int dx_1 \cdots dz_1 S(1, \dots, l) = V l! b_l(T) \quad (39b)$$

to first imagine that molecule 1 is fixed somewhere in the middle of the vessel. Since the S functions vanish for a separated configuration, the *l* molecules must all be quite near to molecule 1. Roughly speaking, if *a* is the range of the molecular forces, integral (39*a*) gets contributions only from a range of about *la* around molecule 1. If *la* is smaller than $V^{1/3}$, the first l - 1 integrations are independent of V. The last integrations over the coordinates of 1 just contribute V, for molecule 1 can be placed anywhere in V. (Wall effects are ignored.) Thus Eq. (39b) holds. The *l*! is a normalization factor and b_l is independent of V. The b_l are called cluster integrals. Using Eqs. (36), (38a), (38d), and (39b), the pressure may be obtained in terms of the cluster integrals as in Eq. (40). Thus, the experimental form of the virial

$$P = \frac{RT}{B} \times \left(1 - \frac{Nb_2}{V} = 4\left(N^2b_2^2 - 2N^3b_3\right)\frac{1}{V^2} + \cdots\right)$$
(40)

development has been deduced, while the cluster integrals are related to the experimental virial coefficients by Eqs. (41a) and (41b). Using Eq. (39b) for

$$B(T) = -Nb_2 \tag{41a}$$

$$C(T) = N^2 (4b_2^2 - 2b_3) \tag{41b}$$

$$B(T) = 2\pi N \int_0^\infty r^2 dr (1 - e^{-[U(r)/kT]}) \qquad (41c)$$

l = 2 gives a direct relation between *B* and the intermolecular force potential U(r) as in Eq. (41*c*). This is a "perfect" statistical formula. In the case of helium, where both U(r) and *B* are known, Eq. (41*c*) may in fact be checked. [At low temperatures, quantum effects, not included in Eq. (41*c*), begin to play an important role.] The complete equation of state can be obtained using a similar procedure. The result comes out in an implicit form, as in Eqs. (42a) and (42b), derived

$$P = kT\Sigma b_l z^l \tag{42a}$$

$$\frac{N}{V} = \sum_{l=1}^{N} lb_l z^l \tag{42b}$$

by J. E. Mayer. In principle, the auxiliary variable zcan be eliminated between Eqs. (42a) and (42b) to obtain a relation between P, V, and T. Of course, to have a useful relation it is necessary to know the general character of the cluster integrals, and, in fact, this is not known, except for special molecular models. A question which has concerned physicists for some time is whether or not the system of equations (42) actually predicts a condensation phenomenon. It is well known that every real gas at a low enough temperature will condense when the volume is decreased. During the condensation process the pressure remains constant, the onset of condensation being marked by a discontinuity in the slope of the isotherm. The problem is now whether this information can be obtained from Eqs. (42). An interesting clue was obtained by Mayer. He showed that for low enough temperatures, $b_l(T) \cong \text{constant} \times b_0^l$. Substituting this relation in Eq. (42b) yields Eq. (43).

$$\frac{N}{V} = \sum_{l=1}^{N} l(\text{constant})(b_0 z)^N$$
(43)

Since the last power in the series is $N \cong 10^{24}$, the series given by Eq. (43) is radically different for $b_0 z < 1$ and $b_0 z > 1$. If $b_0 z > 1$, $(b_0 z)^N$ is a very large number and a change in N/V will cause a very slight change in z, hence a very slight change in P by Eq. (42a). Therefore, $b_0 z = 1$ separates two ranges: If $b_0 z < 1$, changes in N/V cause reasonable changes in P; if $b_0 z > 1$, P becomes quite insensitive to changes in N/V. This qualitative idea has been refined in many ways, by rigorously studying limit (44). These stud-

$$\lim_{N \to \infty \atop V \to \infty} (Z_N)^{1/N}$$
(44)

ies have clarified some aspects of the situation, but even so the problem is still not completely settled. *See* GAS.

Nonequilibrium theory of liquids. It was observed in the earlier discussion of Liouville's theorem that the Liouville equation usually does not help in setting up an appropriate ensemble. Yet the Liouville equation has become the starting point for an important development aimed at an understanding of the nonequilibrium states of dense gases and liquids. This development starts from the Liouville equation (45). The interpretation that $\rho(x_1, \ldots,$

$$\frac{\partial \rho}{\partial t} + \sum_{i=1}^{3N} \left(\frac{\partial \rho}{\partial x_i} \frac{dx_i}{dt} + \frac{\partial \rho}{\partial p_i} \frac{dp_i}{dt} \right) = 0$$
(45)

 $p_{z_N;t}$ $d\Gamma$ is the probability of finding the molecules of the system in their prescribed momentum and position ranges may now be used. A set of probability functions (or distribution functions) can be defined by Eqs. (46). The function f_1 is (apart from constants)

$$f_{1}(\mathbf{x}_{1}, \mathbf{p}_{1}, t)$$

$$= \int \cdots \int d(1)d(2) \cdots dN \ \rho(x, p, t) \qquad (46a)$$

$$f_{1}(\mathbf{x}_{1}, \dots, \mathbf{p}_{l}; t) = \int \cdots \int d(1) \cdots$$

 $d(l)d(l+1)\cdots d(N) \rho(x, p, t) \quad (46b)$

the Boltzmann distribution function. The canceled symbol d(1) means that the variables x_1 , y_1 , z_1 , p_{x1} , p_{y1} , p_{z1} are not integrated over. The higher distribution functions become important for dense systems. If it is assumed that the potential energy of the system consists of an external potential V_0 , and a potential U which is additive, then integration of Eq. (45) over all coordinates except those of molecule 1 yields an equation for f_1 , Eq. (47).

$$\frac{\partial f_1}{\partial t} + \frac{p_{\alpha}}{m_1} \frac{\partial f_1}{\partial x_{\alpha}} - \sum_{\alpha=1}^3 \frac{\partial V_0}{\partial x_{\alpha}} \frac{\partial f_1}{\partial p_0}$$

The equation for f_l involves f_{l+1} , and so on.

$$= \iint d^3 p_2 d^3 x_2 \sum_{\alpha} \frac{\partial U(x_1 x_2)}{\partial x_{\alpha}} \frac{\partial f_2(p_1 p_2, x_1 x_2)}{\partial p_{\alpha}} \quad (47)$$

The discussion of this hierarchy forms the basis of the study of the nonequilibrium phenomena in dense gases. The basic difficulty is that the equations for f_i always involve a function f_{i+1} . To obtain an equation for a single f_i function, it is necessary to make a guess or assumption about the way in which a higher f function can be expressed in terms of lower ones, if indeed this can be done at all. A frequently discussed possibility is the so-called superposition approximation relation (48). An appeal to probabilities of inde-

$$f_2(p_1x_1, p_2x_2) \cong f_1(p_1x_1)f_1(p_2x_2)$$
(48)

pendent events makes relation (48) appear reasonably plausible. However, it has not been proven that the approximation of relation (48) is indeed consistent with the system of Eq. (47). The use of relation (48) in Eq. (47) leads to (nonlinear) integral equations. In spite of the considerable amount of work done with these equations, the results still have not led to great advances in the theory of dense gases. However, for additive central potentials, defined by $U(1, ..., N) = \Sigma U(r_{ij})$, it is possible to express the equation of state in terms of the pair distribution function. This function is defined by Eq. (49*a*). Here

$$n_{2}(\mathbf{x}_{1}, \mathbf{x}_{2}) = N(N-1)$$

$$\times \frac{\int \cdots \int d(1) \ d(2) \ d(3) \cdots d(N) W(1, \dots, N)}{\int \cdots \int d(1) \cdots d(N) W(1, \dots, N)}$$

$$(49a)$$

$$n_{2}(\mathbf{x}_{1}, \mathbf{x}_{2}) = \iint d^{3}p_{1} \ d^{3}p_{2} \ f_{2}(\mathbf{x}_{1}\mathbf{p}_{1}, \mathbf{x}_{2}\mathbf{p}_{2}) \quad (49b)$$

W(1, ..., N) is defined by Eq. (38). The function n_2 gives the positional distribution of molecular pairs. As such it is less detailed than f_2 , which gives the

distribution of pairs both in positions and momenta. Apart from combinatorial factors, Eq. (49*b*) is valid. For central additive potentials, it may now be shown that n_2 is in fact a function of the distance r_{12} between the molecules: $n_2 \propto g(r)$, where g(r) is defined as giving the number of molecules between r and r + dr. *See* INTEGRAL EQUATION.

An analysis similar to the one given in the preceding discussion on nonideal gases now leads to an equation of state, Eq. (50), in terms of g(r) alone.

$$PV = NkT - \frac{2\pi N(N-1)}{3V} \int g(r) \frac{dU}{dr} r^3 dr \quad (50)$$

The integral equation referred to for the f_2 functions (once the superposition approximation is made) may now be expressed as an approximate integral equation for g(r). This equation, obtained by M. Born, H. S. Green, and J. Yvon, leads to approximations for the virial coefficient, which are fairly good but not excellent. The pair distribution function g(r) may be obtained directly from experimental x-ray scattering data. In principle experimental x-ray data could be used to obtain via Eq. (50) data about the equation of state. For this to be a feasible procedure would demand accuracies much beyond the present limits, as well as extensive measurements over a range of temperatures and densities. Unfortunately, therefore, this is not a possible procedure at present, although the formal relation (50) is generally valid. A current theory of liquid helium makes use of the relation between the pair distribution function and neutron-scattering data. See LIQUID HELIUM; NEUTRON DIFFRACTION; X-RAY DIFFRACTION.

Quantum statistical mechanics. The ensemble techniques can also be applied to systems which are described in terms of quantum mechanics. Consider an ensemble, where each system is described by a hamiltonian operator *H*. Let $\psi^{\alpha}(x,t)$ be a wave function (*x* stands for $\mathbf{x}_1, \ldots, \mathbf{x}_N$, and α characterizes an ensemble member). The Schrödinger equation is written as Eq. (51).

$$H\psi^{\alpha}(x,t) \doteq -\frac{\hbar}{i}\frac{\partial\psi^{\alpha}(x,t)}{\partial t}$$
(51)

See NONRELATIVISTIC QUANTUM THEORY.

If $\varphi_n(x)$ is a complete orthogonal set, Eq. (52), the

$$\int \varphi_n^* \varphi_m = \delta_{nm} \tag{52}$$

function ψ^{α} may be developed in terms of the set $\{\varphi\}$ as in Eq. (53). Here $|a_n^{\alpha}(t)|^2$ is the probability

$$\psi^{\alpha}(x,t) = \sum_{n} a_{n}^{\alpha}(t)\varphi_{n}(x)$$
(53)

that ensemble member α , at time *t*, is in state *n*. Of course, Eq. (54) is valid.

$$\sum_{n} \left| a_n^{\alpha}(t) \right|^2 = 1 \tag{54}$$

If *G* is a quantity which is defined for each ensemble system by G^{α} , then an ensemble average of

G is defined by Eq. (55a), where N_e is the number

$$\overline{\overline{G}} = \frac{1}{N_e} \Sigma(G^{\alpha}) \tag{55a}$$

$$\overline{G}^{\alpha} = \int (\psi^{\alpha})^* G_{\rm op} \psi^{\alpha} \tag{55b}$$

of ensemble members. It is important to distinguish the ensemble average from the ordinary quantummechanical average, which is defined for a single system as in Eq. (55*b*). The notation, double bar for ensemble average, single bar for quantum-mechanical average, stresses the difference. Of special importance in the calculation of averages is the density matrix. The matrix elements of ρ , the density matrix, relative to the set of functions φ are defined by Eq. (56*a*). The sum of the diagonal elements of ρ , called the trace of ρ , is given by Eq. (56*b*). The den-

$$\rho_{mn} = \frac{1}{N_e} \sum_{\alpha} (a_n^{\alpha})^* a_n^{\alpha} = \overline{\overline{a_n^* a_m}}$$
(56a)

$$\operatorname{Tr}(\rho) = \sum_{n} \rho_{mn} = \frac{1}{N_e} \sum_{\alpha} \sum_{n} |a_n^{\alpha}|^2 = 1$$
 (56b)

sity matrix ρ is the counterpart of the classical density function. From the Schrödinger equation (53) the analog of the Liouville theorem, Eq. (57), can

$$i\hbar\frac{d\rho_{mn}}{dt} = [H,\rho]_{mn} \tag{57}$$

be deduced immediately. Here $[a,b] \equiv ab - ba$ is the commutator. Finally the basic relation giving the ensemble average of any operator, Eq. (58), can be

$$\overline{G} = \operatorname{Tr}(\rho G) \tag{58}$$

obtained. The fact that observable entities, such as ensemble averages, come out in the form of a trace, implies that the calculated results are independent of the set of functions φ . This is true because a change in this basic set will induce a similarity transformation on both ρ and G, and the trace operation is invariant under such a transformation. *See* DENSITY MATRIX.

An ensemble in which all ensemble members are in the identical state (ψ^{α} is independent of α) constitutes a pure case. An ensemble in which the ψ^{α} do depend on α is a statistical mixture. A necessary and sufficient condition for a pure case is that $\rho^2 = \rho$.

At equilibrium, the system is the counterpart of the canonical ensemble. Then the density operator [see Eq. (12*a*)] is written as Eq. (59*a*). The free energy is

$$\rho = e^{(\psi - H)/kT} \tag{59a}$$

$$Z = e^{-(\psi/kT)} = \text{Tr}\left(e^{-(H/kT)}\right)$$
(59b)

again directly related to the partition function as in Eq. (59*b*). In Eq. (60),*Q* is an operator e^{Q} defined by

$$e^{Q} = \sum_{n} \frac{Q^{n}}{n!} \tag{60}$$

the series expansion. The thermodynamic relations follow as before.

The construction of an ensemble in quantum mechanics must be performed in harmony with the knowledge of the system. If the system is known to be in either one of two quantum states, the ensemble members must be evenly distributed over these states with, in addition, random phases. It is important to include the requirement of random phases; if this were not done, a uniform stationary ensemble would not even remain stationary. *See* QUANTUM STATISTICS. Max Dresden

Bibliography. D. S. Betts, *Introduction to Statistical Mechanics*, 1992; C. Garrod, *Statistical Mechanics and Thermodynamics*, 1995; W. Greiner, L. Neise, and H. Stocker, *Thermodynamics and Statistical Mechanics*, 1995; K. Huang, *Statistical Mechanics*, 2d ed., 1987; R. Kubo et al., *Statistical Physics*, 2 vols., 2d ed., 1991; L. D. Landau and E. Lifshitz, *Statistical Physics*, 3d ed., 1980; R. C. Tolman, *The Principles of Statistical Mechanics*, 1938, reprint 1980; D. H. Trevens, *Statistical Mechanics: An Introduction*, 1993.

Statistical process control

A collection of strategies and statistical methods for ensuring high-quality results in manufacturing and service processes. Traditionally, the statistical methods associated with statistical process control (SPC) have focused on process monitoring. As a result, some experts argue that statistical process control is a misnomer and that the term statistical process monitoring is more appropriate. Others argue that statistical quality control methods have become so entwined with strategic decisions and daily operations that the term statistical process control is entirely appropriate. Regardless, statistical process control methods and philosophies are essential tools for survival in today's highly competitive environment. See PROCESS CONTROL; QUALITY CONTROL; STATIS-TICS.

The origins of statistical process control coincide with a shift in focus by manufacturing companies from quantity to quality. The pioneers in this transformation include W. Shewhart, W. E. Deming, J. Juran, K. Ishakawa, G. Taguchi, and others. At the heart of this transformation were two simple ideas: understanding the customer's needs and meeting the customer's needs consistently. Accomplishing both of these tasks while minimizing wasted time, effort, and resources insured a competitive advantage. The success of organizations employing quality-oriented management styles led to a revolution in management tools and philosophies.

Total quality management. Total quality management (TQM) gained strength in the 1980s. The expression total quality management refers to both the philosophy and practices associated with continuously improving processes within a company. An all-encompassing view of processes included not

only manufacturing operations but also service areas such as sales and accounting functions. Companies adopted various strategies for continuous improvement with varied levels of commitment under an assortment of names. Total quality control, continuous quality improvement, and total quality excellence are synonyms for total quality management. These programs had several common themes. An important commonality was reliance upon actual process data to monitor systems, inform decision makers, and guide the continuous improvement activities. The tools associated with statistical process control are fundamental building blocks in these quality-driven programs. International standards were eventually established to ensure that total quality management programs, while varied, followed certain basic principles and practices. The standards known as ISO 9000:2000 are a product of the International Standards Organization (ISO), the American Society for Quality (ASQ), and the American National Standards Institute (ANSI).

Six Sigma. The ideas surrounding total quality management grew into the Six Sigma quality movement, which continues to evolve. Among the unique features of Six Sigma programs are the certification levels for managers trained in statistical problem solving and the use of the DMAIC process. The acronym DMAIC refers to five steps: define/deploy, measure, analyze, improve, and control.

The certification process begins with training in basic statistical methods and continuous quality improvement concepts. Managers mastering these initial topics are called "green belts." Through a series of training classes and real-time projects, managers ascend to the ranks of "black belt" or "master black belt." Managers receiving the higher-level designations must demonstrate an understanding of the philosophies and tools of continuous improvement, as well as the ability to put these ideas into practice with substantial success, as measured in long-term savings in real dollars.

The DMAIC problem-solving approach is central to Six Sigma's successful use of statistical process control methods throughout an organization. The define/deployment phase involves selecting a project leader (black belt), defining the project, and creating a project support structure. The measurement phase captures current project-related wisdom. Helpful tools used in this phase include many traditional quality control tools such as flow charts, check sheets, cause-and-effect diagrams, histograms, scatter plots, and control charts. These tools help identify key input and output variables used to control and measure process performance. The analysis phase provides a comprehensive understanding of process data and procedures. Process changes (improvements) are then recommended and tested using designed experiments or optimization techniques. Changes that result in improvements are permanently incorporated in the control phase, as are activities such as determining control plans, updating standard operation procedures, and mistakeproofing processes. See CONTROL CHART.

The preceding paragraphs highlight the seamless incorporation of statistical tools into the day-to-day working of management. Traditional quality control techniques such as control charts are now complemented by appropriate statistical methods such as designed experiments and modern technologies. Hence, statistical process control tools are used beyond quality engineering. The result is a blurring of the boundaries that traditionally defined statistical process control.

Quality control charts. All processes display variability. Processes include all meaningful activities that may be measured in some way. The process of getting to work in the morning or serving a customer may be measured in time. The process of surgery may be measured by survival rates, recovery times, or infection rates. The process of manufacturing an object may be measured by yield rates, dimensional properties of the object, or conformance to customer specifications. Determining whether a process is performing in a stable and predictable manner or displaying unusual behavior is the traditional domain of statistical process control.

All measurable processes exhibit variation that can be characterized as belonging to one of two classes: common-cause variation and special-cause variation. Common-cause variation is often unexplainable and unavoidable without fundamental changes to the underlying process. Special-cause variation is attributable to a specific cause that may be identified and corrected. The quality control chart is used to detect and address special-cause variation. Additionally, the control chart is used to quantify and avoid overreaction to common-cause variation. Overadjustment of equipment by operators can result in a deterioration of overall quality.

Underlying logic. To illustrate the usefulness of control charting, consider a process for manufacturing an automobile part. The logic of control charting is simple, as outlined in the following steps:

1. Select an appropriate measure for a given quality characteristic. One may be interested in the length of the manufactured part.

2. Collect samples at regular intervals. One might randomly select four parts every 2 hours for inspection.

3. Calculate an appropriate statistic from the sample data. The sample mean (average of the lengths for the four parts) would help determine if parts are of appropriate lengths. The sample range (difference in the lengths of the longest and shortest parts in the sample) would help determine if the parts are consistently the correct length.

4. Plot the calculated statistic(s) over time and compare the plotted values to predetermined values, called control limits. Unusual patterns or extreme observations beyond the control limit could indicate special-cause variation to be identified and corrected by operators. Such patterns or extreme values are called signals. Values falling between the control limits indicate common-cause variation, and



Control charts for length: (a) mean and (b) range. UCL = upper central limit. LCL = lower central limit. $\overline{\overline{X}}$ = average of all \overline{X} points. $\overline{\overline{R}}$ = average of all R points.

operators would allow the process to run without interruption.

The control chart is displayed graphically to allow easy use by operators. For example, the **illustration** shows the control charts for the mean length of parts and the range in the length of parts. For each sample, the average length for four randomly selected parts is calculated and plotted on the mean chart, and the sample range for the lengths is plotted on the range chart. Observations falling beyond the control limits of a control chart are called signals. The mean chart signal at time = 21 suggests that something unusual has occurred that requires the operator's attention. No values are beyond the control limits of the range chart, which indicates that the problem affects only the part length, not the process consistency (that is, the process is consistently making bad parts).

Guidelines for important questions. While the steps outlined in the previous section appear straightforward, careful consideration is required at each step.

Step 1: Get the right data. Selecting the best measurement method for a quality characteristic with regard to process information and practical considerations, such as safety, time, and ease of obtaining measurements, is obviously process-specific. There are, however, guidelines governing variable selection.

Control charts typically fall into two groups: attributes and variables. Attribute data involve counts or proportions. For example, a sample of 100 items might be inspected and the number of items not meeting customer requirements reported. Each item would be simply classified as good or bad depending on some quality attribute (length is acceptable or not acceptable). On the other hand, continuous variable measurements involve lengths, weights, times, or measurements that could obtain any value within a given range. For example, actual lengths of parts might be recorded. Continuous variable measurements are inherently more informative than attribute measurements. For example, suppose customer specifications require that parts have lengths of 12 ± 0.03 in. Simply knowing that a part is not the correct length is less informative than knowing that the part length is 12.04 in., which indicates both the direction and magnitude of the problem. The additional information contained in variable measurement allows more powerful monitoring schemes and helps operators in diagnosing problems. Continuous measurements should be used where possible. *See* INSPECTION AND TESTING.

Step 2: Get the right amount of data from the right place. The idea of rational subgroups is helpful in determining sampling locations. One should collect data at locations that allow operators to isolate potential problems. For example, if three production lines produce identical parts, sampling at the end of each line is preferred to sampling after parts from the three lines have been merged. After the parts have been mixed together, it may be difficult to determine which production line is experiencing problems.

Selecting how many items should be in each sample and how often samples should be collected requires technical knowledge of control chart performance. Generally, small samples taken frequently are preferred to large samples taken infrequently.

Step 3: Summarize the data in an appropriate manner. This involves the statistics that should be calculated. Data summaries (sample statistics) should answer two questions: Is the process on target (for example, parts are on average the correct length)? Is the process behaving in a consistent manner?

Continuous variables data provide better answers to these questions than do attribute data. Data summarized by averaging sample values (or weighted averages of past values) are used to determine if a process is on target. Practitioners refer to these statistics as measures of location. The sample range or sample standard deviation is useful in quantifying consistency of process behavior. Such statistics are referred to as measures of variation.

Step 4: Plot the summary statistics and control limits. The value(s) of the control limit(s) depends on the statistic being plotted, the sample size, and the tolerable frequency of false alarms. A signal on a control chart is called a false alarm if no actual problem exists. The average number of samples required for a control chart to signal a process problem is called the average run length (ARL). A large ARL is desired for problem-free (in-control) processes, while small ARLs are desired for processes with problems (out-of-control processes). General guide-lines for determining control limits are provided in most introductory texts on statistical process control methods.

Additional comments. The appropriate selection and design of control charts has received much attention over the years. Performances of the most common control charts are well understood. Information concerning the performance and selection of control charts may be found in many textbooks. In addition to common control charts, there are many specialized monitoring techniques available to practitioners. Multivariate control charts are used when several important process variables are related to one another. Forecast-based or regression-based charts are useful when data values change systematically over time. Process change due to tool wear is an example. Practitioners are advised to select an appropriate technique for each process.

A second area of caution associated with control charting is the relationship between control limits and specification limits. The relationship is extremely simple because there is no relationship. Specification limits reflect the customer's requirements. Control limits reflect the common-cause variation in a process and the statistic plotted on the control chart. The customer's needs are not determined by the process, nor are process variations determined by the customer.

The primary goals of control charting are to identify and eliminate special causes of variation, ensure that the process remains in control, and quantify common-cause variation. Once a process is behaving in a stable and predictable manner, one may evaluate the capability of the process and use designed experiments to gain further improvements.

Capability indices. One is always interested in knowing if inherent process variation allows one to meet customer needs. A capability index is a numerical measure relating the demands of the customer to the capability of the producer. There are many versions of capability indices, just as there are many types of control charts. The appropriate index de-

pends on the properties displayed by process data. Most capability indices are a ratio of customer specifications and common-cause process variation. For example, one capability index is defined by

$$C_p = \frac{(\text{USL} - \text{LSL})}{6\hat{\sigma}}$$

where $\hat{\sigma}$ represents the estimated standard deviation associated with common-cause variation, and USL and LSL represent the customer's upper specification limit and lower specification limit, respectively. A large C_p value indicates that the inherent process variation ($6\hat{\sigma}$) is small relative to customer specification limits (VSL–LSL). Thus, the process is very capable of meeting customer requirements.

A small capability ratio indicates that a process is not capable of meeting customer requirements without fundamental changes in the process and/or operation procedures. Process capability must be improved by identifying the most efficient and appropriate process changes that ensure that the output is on target, with minimal variation.

Designed experiments. Designed experiments are powerful techniques for improving process capability. Statistically designed experiments allow systematic investigation of factors that affect process performance. These efforts can improve processes in several ways. Greater understanding of process variables aids diagnostic efforts, leads to process optimization, and increases process robustness to nuisance variables that are difficult or impossible to control. As processes are improved, control charts must be updated to reflect the new and improved process behavior. Process capability ratios must be recalculated to reflect the increased process capability.

The cycle of monitoring and stabilizing processes, determining their capability, and improving process capability through statistically designed experiments is a powerful approach to quality improvement. Understanding customer needs in a quantifiable manner, communicating these requirements to operations, and focusing attention on consistently meeting customer needs seem only logical. Statistical process control techniques not only maximize the chance of gaining a competitive advantage but also allow gaining an advantage at a minimum of expense. These techniques have been useful in pure research and development for many years. As understanding of their usefulness grows statistical process control methods will remain a key tool in the management of manufacturing and service processes. See EXPER-IMENT; INDUSTRIAL ENGINEERING; MANUFACTURING ENGINEERING. Benjamin M. Adams

Bibliography. E. Grant and R. Leavenworth, *Statistical Quality Control*, 7th ed, 1996; G. Gryna, *Quality Planning & Analysis: From Product Development Through Use*, 4th ed., 2001; J. M. Juran and A. B. Godfrey, *Juran's Quality Handbook*, 5th ed., 1999; D. Montgomery, *Introduction to Statistical Quality Control*, 5th ed., 2005; T. Pyzdek, *The Six Sigma Handbook, Revised and Expanded: The Complete*

Guide for Greenbelts, Blackbelts, and Managers at All Levels, 2d ed., 2003; K. Yang and B. S. El-Haik, *Design for Six Sigma: A Roadmap for Product Development*, 2003.

Statistics

The field of knowledge concerned with collecting, analyzing, and presenting data. Not only workers in the physical, biological, and social sciences, but also engineers, business managers, government officials, market analysts, and many others regularly use statistical methods in their work. The methods range from simple counting to complex mathematical systems designed to extract the maximum amount of information from very extensive data.

In an important sense statistics may be regarded as a field of application of probability theory. The common problem faced by a physicist reading a meter, an engineer testing a material, an agronomist measuring the yield of a hybrid corn, a chemist determining the concentration of ascorbic acid, and an interviewer studying public opinion is the problem of random variation which prevents repetition of exactly the same result when a measurement is repeated. Statistical methods are employed to assess the magnitude of random variation, to minimize it, to balance it out, to remove it by calculation procedures, and to analyze it by suitably arranged patterns of observation. The theory of probability is concerned with the properties of random variables and hence furnishes the basis for developing techniques for controlling them. See PROBABILITY.

Viewing statistics from another direction, it is the science of deriving information about populations by observing only samples of those populations. A population is any well-specified collection of elements. Thus, one may refer to the population of adults in the continental United States viewing television screens at 8:14 P.M. on August 6, 1970; the population of automobiles less than 2 years old registered in Los Angeles County on a certain date; the population of vineyards in France; the hypothetical population of outcomes of tossing a given coin endlessly. Populations may be finite or infinite. An element of a univariate population is characterized by the value of a random variable which measures some single attribute of interest in the population. Thus, one may be interested in whether or not individuals of the television audience were or were not viewing program A; with each individual one may associate a random variable, let it be X, which takes on the value of 1 if the individual is watching A and 0 if he is not. If one were interested in a second characteristic of the elements of the television audience (such as age), one would be said to be dealing with a bivariate population; a third characteristic (such as economic status) would make it a trivariate or, less specifically, a multivariate population.

Random variables are either continuous, which means they can take on any numerical value (the length of a room), or discrete, which means they



Fig. 1. Distribution of incomes

can take on only a restricted set of values (number of windows in a room).

Distributions. In a univariate population, the population distribution is a curve (function of the random variable which characterizes the elements of the population) from which one can determine the proportion of the population which has elements in a certain range of the random variable. For example, the curve of **Fig. 1** provides the distribution of annual incomes of family units. The total area under the curve is 1. The area under the curve between any two vertical lines gives the proportion of the families having annual incomes between the two values marked on the horizontal scale by the two vertical lines. Thus, the fact that the area under the curve between \$2000 and \$6000 is 0.541 means that 54.1% of the family units have incomes in that range.

The distribution is also referred to as the distribution function, the density function, the frequency function, or the probability density.

The total area under the distribution curve to the left of each point can also be plotted to give a curve which starts at zero and reaches unity as the variable becomes large; the resulting curve is sometimes called the cumulative distribution function, the probability distribution, or simply the distribution. The cumulative form of the curve of Fig. 1 is shown in **Fig. 2**; the height of the curve at any point on the horizontal scale equals the area to the left of that point under the curve of Fig. 1 and is the proportion of the population having incomes less than the value at that point. The distribution (in either frequency or cumulative form) gives complete information about the way the characterizing variable is spread through the population.



Fig. 2. Cumulative distribution of incomes.

Population parameters. Populations (or population distributions) are often specified incompletely by certain population parameters. Some of these parameters are location parameters or measures of central tendency; a second class of important parameters consists of measures of dispersion or scale parameters.

The most widely used location parameters are the mean, the median, and the mode. The mean is the average over all the population of the values of the random variable. It is often represented by the Greek letter μ . In mathematical terms, if x is the random variable, f(x) the frequency function for a given population, and F(x) its cumulative form, then the mean is as shown in Eq. (1).

$$\mu = \int_{-\infty}^{\infty} x f(x) \, dx = \int_{-\infty}^{\infty} x \, dF(x) \tag{1}$$

The median, often designated by Med, M, or $X_{.50}$, is a number such that, at most, one-half the values of the variable associated with the elements of the population fall above or below it, as in Eq. (2). The mode

$$\int_{-\infty}^{M} dF(x) \ge \frac{1}{2} \le \int_{M}^{\infty} dF(x)$$
 (2)

is the most frequent value of the random variable; if the frequency function has a unique maximum value, the mode is the value of the random variable at which the frequency function reaches its maximum. Location parameters are numbers near the center of the range over which the random variable of the population varies; different ones arise from different definitions of center; generally the mean is used unless special circumstances make some other location parameter more appropriate.

The extent to which a population is scattered on either side of its center is roughly indicated by measures of dispersion such as the standard deviation, the mean deviation, the interquartile range, the range, and sometimes others. The standard deviation is the square root of the mean square of the deviations from the mean; it is usually denoted by the Greek letter σ ; σ^2 is called the variance and is expressed by Eq. (3). The mean deviation, shown in Eq. (4), is the average over the population of the

$$\sigma^{2} = \int_{-\infty}^{\infty} (x - \mu)^{2} f(x) dx$$
$$= \int_{-\infty}^{\infty} (x - \mu)^{2} dF(x)$$
(3)

Mean deviation
$$= \int_{-\infty}^{\infty} |x - \mu| f(x) dx$$
$$= \int_{-\infty}^{\infty} d|x - \mu| dF(x) \quad (4)$$

deviations from the mean, all taken to be positive. The interquartile range (often denoted by Q) is the difference $X_{.75} - X_{.25}$, where $X_{.75}$ is the value of the random variable such that one-quarter of the population has values larger than $X_{.75}$, and $X_{.25}$ is the number such that one-quarter of the population has values smaller than $X_{.25}$. The three numbers, $X_{.25}$, $X_{.50}$, $X_{.75}$, are called quartiles; these divide the population into quarters. The range is the difference between the largest and the smallest of the population elements.

Samples. If one examines every element of a population and records the value of the random variable for each, complete information is obtained about the distribution of the random variable in the population, and there is no statistical problem. It is usually impossible or uneconomical to make a complete enumeration (or census) of a population, and one must therefore be content to examine only a part or sample of the population. On the basis of the sample, one draws conclusions about the entire population; the conclusions thus drawn are not certain in the sense that they would likely have been somewhat different if a different sample of the population had been examined. The problem of drawing valid conclusions from samples and of specifying their range of uncertainty is known as the problem of statistical inference.

Statisticians distinguish two kinds of samples. A survey sample is one chosen from a population, all the elements of which actually exist. An experiment is a sample chosen from a hypothetical population. Thus if one was interested in the total number of pigs being fattened in a given season by the farmers of the state of Iowa, the researcher would survey-sample the existing population of pigs in Iowa at that time. If the effect of a certain hormone on the growth rate of pigs was the subject, an experiment would be performed by giving a group of pigs the hormone for a period of time; using the sample of observations obtained from the experiment, the researcher would draw inferences about the hypothetical population of observations that would have resulted had all pigs been given the hormone.

Random sampling. In planning a sample survey, the manner in which the data to be gathered will fulfill the purpose should be clearly stated. The population to be sampled must be explicitly defined. The method of sampling should be efficient and lead to a straightforward analysis. The question of what elements should be included in the population depends on the purpose of the survey. Thus the population of vineyards in France might include as a vineyard a dozen vines in the backyard of a person living in the heart of Paris if the purpose were to estimate total grape production of France; but the population might be defined to exclude such a small vineyard if the purpose were to estimate the size of the harvest to be available to commercial wineries

The type of sampling of interest in statistics is probability sampling, because it eliminates subjective aspects from the selection of the sample. In probability sampling, all possible distinct samples are known, the selection of the sample is done randomly according to a preassigned probability, and the method of analysis is predetermined and unambiguous. Only from such samples can inferences about populations be made with measurable precision. Simple random sampling is a method of selecting a sample of n elements out of a population of N elements so that all such samples have an equal probability of being drawn. This may be done by selecting a first element at random from the population, then a second element at random from the remaining population, and so on until the n elements are selected. Because an element cannot appear more than once in the sample, this is a form of sampling without replacement. The sampling ratio or sampling fraction is n/N.

Often the purpose of drawing a sample is to estimate the mean or average of a characteristic of the population. If y_i is the value of the characteristic of the *i*th unit, then the population mean is as shown in Eq. (5). The sample mean is the average of the *n* units

$$\mu = \frac{1}{N} \sum_{i=1}^{N} y_i$$

= $\frac{1}{N} (y_1 + y_2 + y_3 + \dots + y_N)$ (5)

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} (x_1 + x_2 + \dots + x_n)$$
 (6)

in the sample and is expressed in Eq. (6). Here the *n x*'s are the *n y*'s selected from the population as the sample. The population total is $N\mu$ and is estimated by Nx^- . The population variance is defined as shown in Eq (7), and the sample variance is usually defined as in Eq. (8), although some authors use *n* instead

$$\sigma^{2} = \frac{1}{N} \sum_{i=1}^{N} (y_{i} - \mu)^{2}$$
(7)

$$s^{2} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{i} - \bar{x})^{2}$$
(8)

of n - 1. The quantities \bar{x} and s^2 , the sample mean, and the sample variance are called sample statistics; they are the first and second sample moments and are also estimators of the corresponding population parameters, μ and σ^2 .

The estimators are themselves random variables. If one repeatedly drew samples of size *n* from the population and computed \bar{x} from each sample, he would obtain a population of \bar{x} 's with its own distribution which would differ from the distribution of *x*. It can be shown that the \bar{x} population has exactly the same mean μ as the *x* population. Further, the variance of the \bar{x} population is as shown in notation (9), where

$$\frac{(N-n)}{(N-1)}\frac{\sigma^2}{n} \tag{9}$$

 σ^2 is the variance of the *x* population. The first fraction is ordinarily nearly unity so that the variance of \bar{x} is approximately the fraction 1/n of the original population variance. As *n*, the sample size, becomes large, the \bar{x} population becomes more concentrated about μ and the reliability of a particular value of \bar{x} as an estimate of the population mean increases.

The observations of a sample, besides providing estimates of population parameters, can also be used



Fig. 3. Histogram.

to obtain an estimate of the population's frequency function. This estimate is determined by dividing the range of the sample observations into several intervals of equal length L and counting the number of observations occurring in each interval; these numbers are then divided by nL to determine fractions giving the relative density of the sample occurring in each interval; then on a sheet of graph paper one lays out the intervals on a horizontal axis and plots horizontal lines above each interval at a height equal to the fraction corresponding to the interval; finally the successive plotted horizontal lines are connected by vertical lines to form a broken line curve known as a histogram (Fig. 3). The area under the curve is unity, and the area between any two points gives the fraction of the sample observations lying between those two points. If one takes larger and larger samples, the chosen intervals can be made smaller and the broken line curve will come closer and closer to the underlying population frequency function. Often one does not trouble to divide the interval frequencies by nL to normalize the area of the histogram, but merely plots the frequencies themselves; the resulting broken line curve is still referred to as a histogram.

Sampling techniques. When a population can be regarded as being made up of several nonoverlapping subpopulations, one may draw a sample from it by drawing a simple random sample from each subpopulation (or stratum); this procedure is called stratified random sampling. The method is employed when it is desired to have specific information about each stratum individually, when it is administratively convenient to subdivide the population, when there are natural strata, or when a gain in precision would be realized because each stratum is more homogeneous than the whole population.

Let *n* be the number of units sampled in the entire population of *N* units, and let n_b be the number of units sampled in stratum *b* containing N_h units. Therefore, the sum of all the n_b is *n*, and of all the N_b is *N*. Let σ_b be the true standard deviation within stratum *b*. When $n_b/n = N_b/N_n$, the n_h are said to be proportionately allocated. When $n_b/n = N_b\sigma_b/\Sigma N_b\sigma_b$, the n_b are said to be optimally allocated because the variance of the estimated mean is then smallest for fixed total size of sample. Ordinarily, if the σ_b are well estimated, optimum allocation gives greater precision than proportional allocation, and proportional allocation gives greater precision than simple random sampling.

Systematic sampling may be regarded as a sampling of units at regular intervals in the population, for example, every tenth unit. Usually this is done with a random start; that is, the first unit in the sample is selected at random from a small group at the beginning. A systematic sample is usually relatively easy and fast to execute, and if the analogy to stratified sampling is valid, a systematic sample will be more accurate. However, because of the essentially nonrandom nature of a systematic sample, it is difficult to estimate the sample variance unless certain assumptions are made about randomness in the order of the population. If the population has any periodicity, then the estimates made from a systematic sample can be quite poor.

When the elements of a population are in mutually exclusive groups called the primary units of the population, then a sample of these primary units might be made, and then, from those selected, a sample of the individual elements would be made. This procedure is called two-stage sampling or subsampling. Multistage sampling can involve more than two stages of sampling. Although more complicated and difficult to apply and to analyze, multistage sampling offers considerable flexibility for balancing between statistical precision and the cost of sampling.

When too little is known about a population to plan a sample, then two samples are made. The manner in which the second sample is taken is determined from the results of the first sample. In Stein's method of two-stage sampling, the size of the additional sample is decided by using the estimate of the variance from the first sample. In double sampling or two-phase sampling, the first sample is used to gather information on a variate associated with the variate of interest. This information is used to establish strata for drawing a stratified random sample involving the variate of interest.

Ratio and regression estimates are often used when observations are made on one or more variates related to the variate of primary interest.

Sequential sampling and sequential analysis refer to a method of sampling in which the size of the sample is not specified in advance. Observations are drawn one by one until a specified degree of confidence in the information to be obtained has been achieved.

Sampling distributions. Many important sampling distributions are derived for random samples drawn from a normal or gaussian distribution, which is a



Fig. 4. Normal distribution.

bell-shaped symmetrical distribution centered at its mean μ . The distribution is illustrated in **Fig. 4** for three different values of σ , the standard deviation. Equation (10) is the mathematical equation for these

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-u)^2/\sigma^2}$$
(10)

curves. The area under the curves for the indicated limits is as follows:

Area	Limits
0.50	$\mu - 0.675\sigma$ to $\mu + 0.675\sigma$
0.683	$\mu - \sigma$ to $\mu + \sigma$
0.90	$\mu - 1.645\sigma$ to $\mu + 1.645\sigma$
0.95	$\mu - 1.960\sigma$ to $\mu + 1.960\sigma$
0.99	$\mu - 2.326\sigma$ to $\mu + 2.326\sigma$

Referring to the above limits, 0.675σ is called the probable error inasmuch as the odds are even that a randomly drawn observation will lie within 0.675σ of the mean.

If samples of size *n* are drawn from a normal population and the sample mean \bar{x} is computed for each, the \bar{x} 's will have a normal distribution with the same mean and with variance σ^2/n . Thus, if samples of size 4 were drawn from a population distributed by the curve marked $\sigma = 1$ in Fig. 4, then the means of those samples would be distributed by the curve marked $\sigma = 0.5$ (that is, $1/\sqrt{4}$). Further, the probability would be 0.95 that a particular \bar{x} so drawn would lie with (1.960)(0.5) = 0.98 of the population mean μ .

The central limit theorem of the theory of probability states that under very general conditions the sample mean \bar{x} is approximately normally distributed whatever may be the distribution function for the underlying population. This powerful theorem enables one to make probability statements such as the one at the end of the preceding paragraph in ignorance of the actual population distribution.

Besides the distribution of the mean, two other sampling distributions derived from the normal distribution have wide application. One is the chi-square distribution, χ^2 , which provides the distribution of the sample variance, as shown in Eq. (11).

$$s^{2} = \sum_{i=1}^{n} \frac{(x_{1} - \bar{x})^{2}}{n - 1}$$
(11)

Here x_1, x_2, \ldots, x_n are the observations of a random sample of size *n* drawn from a normal population.

The quantity given in Eq. (12) has a distribution

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \tag{12}$$

illustrated in **Fig. 5** for three values of n. The mathematical form for the distribution curve is shown in notation (13) and is often referred to as the chi-square

$$\frac{1}{[(n-3)/2]!} \left(\frac{n-1}{2\sigma^2}\right)^{(n-1)/2} (\chi^2)^{(n-3)/2} e^{-(n-1)\chi^2/2\sigma^2}$$
(13)

distribution with n - 1 degrees of freedom.

χ

The most useful random variable for interval



Fig. 5. Chi-square distributions.

estimation of a population mean is shown in Eq. (14),

$$t = \frac{\sqrt{n}(\bar{x} - \mu)}{s} \tag{14}$$

which has a symmetrical distribution very similar in appearance to the curves plotted in Fig. 4. The mathematical form for the distribution is notation (15). This is referred to as the *t* distribution or the Student

$$\frac{[(n-2)/2]!}{\sqrt{(n-1)\pi}[(n-3)/2]!} \left/ \left(1 + \frac{t^2}{n-1}\right)^{n/2}$$
(15)

distribution with n - 1 degrees of freedom. The following gives limits which include 95% of the area under the curve for a few values of n:

n	Limits			
2	-12.71 to $+$ 12.71			
3	-4.30 to + 4.30			
4	-3.18 to $+3.18$			
10	-2.20 to $+$ 2.20			
25	-2.06 to $+$ 2.06			
120	-1.98 to $+1.98$			

For very large sample sizes, the limits become -1.96 to +1.96, as for the normal distribution. An illustration of the use of these limits is given below.

Estimation. In making an estimate of the value of a parameter of a population from a sample, a function (called the estimator) of the observations is used. For example, for estimating the mean of a normal population the mean of the sample observations is usually taken as the estimator. Another estimator is the average of the two most extreme observations. In fact, there is an infinitude of estimators. The problem of estimation is to find a "good" estimator.

A good estimator may be regarded as one which results in a distribution of estimates concentrated near the true value of the parameter and which can be applied without excessive effort. There is no single way of deciding how good an estimator is, but there are several criteria by which an estimator may be judged.

An unbiased estimator is one which results in a distribution of estimates which has a mean exactly equal to the value of the parameter being estimated. Otherwise the estimator is called biased. The bias is the mean of the distribution of the estimator minus the value of the parameter it estimates.

An estimator is said to be consistent if the probability that an estimate will differ from the value of the parameter by more than any fixed amount can be made arbitrarily small by increasing the number of observations.

The variance of an estimator is the mean squared deviation of the estimates from the value of the parameter. The estimator with the smallest variance is called most efficient. The relative efficiency of two estimators is the ratio of the variances. When the numerator of this ratio is the variance of a most efficient estimator, this ratio is simply called the efficiency of the other estimator.

An estimator is said to be sufficient if it contains all the information in the sample regarding the parameter. This is so when the conditional distribution of the sample for a given estimate is independent of the parameter.

There are several methods of constructing estimators of parameters. The method of moments is applied by assuming that the first few sample moments are equivalent to the moments of some distribution, and then solving for the parameters of that distribution. The methods known as least squares, minimum variance, and minimum chi-square all have as their basis the estimation of the values of the parameters which minimize some linear function of the squares of deviations of the observations from the values of the parameters. In applying Bayes' method, the distribution of possible values of the parameter before the sample is taken, called the a priori distribution, is used in conjunction with the observations in the sample to yield an estimator. The method of maximum likelihood uses as the estimate that value of the parameter for which the probability of the sample is highest.

A confidence interval is an interval constructed in such a way that the true parameter value is within this interval with a predetermined probability in repeated sampling; this probability is called the confidence level of the interval. For example, a sample mean μ known to be normally distributed with variance σ^2 will differ less than 1.96 σ from the true but unknown mean μ with probability 0.95. Thus expression (16) can be formulated. On solving the two inequalities for μ , expression (16) may be written as expression (17).

Probability that

$$(-1.9\sigma < \bar{x} - \mu < 1.96\sigma) = 0.95 \tag{16}$$

Probability that

$$(\bar{x} - 1.96\sigma < \mu < \bar{x} + 1.96\sigma) = 0.95$$
 (17)

When a particular value of x^- is computed from the observations, then $x^- \pm 1.96\sigma$ is a pair of numbers which define a particular interval called the confidence interval. Because μ is a fixed number, it is within this particular interval with probability zero or one. However, μ is unknown, so that the confidence to be associated with the interval is stated in terms of the proportion of all intervals constructed in this manner which would include μ , rather than

this particular interval. A probability of this sort, valid for a population of outcomes, when used in reference to a particular outcome is called a fiducial probability.

Ordinarily σ is unknown and the estimate *s* derived from the sample must be used to form a confidence interval; in this case, the *t* distribution rather than the normal distribution must be used. As an example, suppose a chemist has made four determinations of the atomic weight μ of hydrogen as follows: 1.0066, 1.0090, 1.0084, and 1.0086. The average of these values is given in Eq. (18), and the sample estimate of

$$\bar{x} = \frac{1}{4}(1.0066 + 1.0090 + 1.0084 + 1.0086)$$

= 1.0082 (18)

the variance of his technique is as shown in Eq. (19), the estimate of σ being therefore Eq. (20). It follows

$$s^{2} = \frac{1}{3} [(1.0066 - 1.0082)^{2} + (1.0090 - 1.0082)^{2} + (1.0084 - 1.0082)^{2} + (1.0086 - 1.0082)^{2}] = .00000123$$
(19)

 $s = \sqrt{0.00000123} = 0.0011 \tag{20}$

from the definition of t in the preceding section that in this case Eq. (21) holds. Using the short t table

$$t = \frac{2(1.0082 - \mu)}{0.0011} \tag{21}$$

given in the preceding section, one finds that expression (22) can be formulated. Solving these inequalities for μ gives expression (23), and the chemist can

Probability that $\left(-3.18 < \frac{2(1.0082 - \mu)}{0.0011} < 3.18\right) = 0.95 \quad (22)$

Probability that

$$(1.00645 < \mu < 1.00995) = 0.95 \tag{23}$$

assert with 95% confidence that the atomic weight of hydrogen lies between 1.00645 and 1.00995. For greater precision, more observations are required.

In a similar fashion one may obtain a confidence interval for σ by using s^2 in connection with the chi-square distribution.

A confidence region is the generalization of a confidence interval and refers to the simultaneous estimation of several population parameters. The confidence level is the proportion of the time that the region actually includes the true values of the parameters.

In general, the most desirable confidence interval or region is the smallest one which can be constructed for the selected confidence level.

Tests of hypotheses. Besides estimation of parameters, another major area of statistical inference is the testing of hypotheses. A hypothesis is merely an assertion that a population has a specific property. The test consists of drawing a sample from the population and determining whether or not it is consistent with the assertion. Very often the hypothesis is a statement about the mean of a population;



Fig. 6. Simple test of a hypothesis.

that it has a given value, that it is the same as that of another population, that it exceeds that of another population by at least 10 units, and the like. Thus, one may be comparing a new blend of gasoline with a current blend, a new drug with a standard one, a new manufacturing process with an existing one.

For a very simple illustration of the basic ideas involved, suppose a population is known to have either of two distributions which differ mainly in location. Let the random variable be x and let the two functions of x which determine the two distributions be represented by the symbols f(x) and g(x) as indicated in **Fig. 6**; that is, f(x) represents the height of the left curve at x, and g(x) represents the height of the right curve at x. Assume that the population has the distribution f(x); there is in this case only a single alternative, g(x), so that rejection of the hypothesis implies acceptance of the alternative. Suppose that for testing the hypothesis one can afford only a single observation, that is, a sample of size one.

It is obvious in this instance that one should choose some number *b* in advance, then accept the hypothesis if the observation falls to the left of *b* and reject it if the observation falls to the right. Values of *x* to the left of *b* are said to constitute the acceptance region of the test; values to the right of *b* constitute the critical region. The area under f(x) to the left of *b*, which is denoted mathematically by notation (24), is

$$\int_{-\infty}^{\infty} f(x) \, dx \qquad (24)$$

the probability that the hypothesis will be accepted when it is true. The area to the right of *b* under f(x)is the probability that the hypothesis will be rejected when it is true; this probability is called the type I error of the test. The area to the left of *b* under g(x) is the probability that the hypothesis will be accepted when it is false and is called the type II error of the test. The area to the right of *b* under g(x) is called the power of test and is, of course, one minus the type II error.

By moving the point b to the right one can make the type I error as small as one likes, but in doing so the type II error is increased. This dilemma is characteristic of the construction of tests of hypotheses. Ordinarily one arbitrarily chooses the type I error to be some small number such as 0.05 or 0.01 and then chooses the critical region so as to minimize the type II error.

The fraction f(x)/g(x) is called the likelihood ratio or simply the likelihood function of *x*. From Fig. 6 it is evident that the likelihood is relatively large when the hypothesis is true and small when it is false. For a sample of size n with observations x_1, x_2, \ldots, x_n the sample likelihood is defined to be the product of the individual likelihoods given in expression (25). Let all possible samples of size n be divided

$$\frac{f(x_1)f(x_2)}{g(x_1)g(x_2)} \cdots \frac{f(x_n)}{g(x_n)}$$
(25)

into two sets, with one set (the acceptance region) containing those for which the likelihood is larger than some number *b* and with the other set (the critical region) containing those samples for which the likelihood is less than *b*. It can be proved mathematically that this is the best critical region for testing the hypothesis in the sense that it minimizes the type II error for given type I error. The likelihood criterion therefore furnishes a procedure for constructing specific tests of hypotheses. Other approaches to the testing problem, such as Bayes' method or the minimax principle of game theory, lead to the same criterion.

If, for example, one applies the likelihood criterion to the problem of testing whether a normal population has a given mean, perhaps 10, and specifies that the type I error shall be 0.01, then the following procedure results: Draw a random sample of size n; construct an 0.99 confidence interval for the population mean μ ; if the number 10 lies within the interval, accept the hypothesis; otherwise, reject it.

An important class of hypotheses has to do with tests of independence in multivariate populations. As an example, one may consider the population of registered voters in the United States as a bivariate population with one variable being their opinions (yes, no, undecided) on some political proposition of the moment, and the other variable being geographical location. Elements of the population may be classified into a so-called contingency table as shown in Fig. 7. The hypothesis of independence asserts that the division of political opinion is unaffected by location. To test that hypothesis, a random sample of the population may be interviewed and classified into the 12 categories or cells of the contingency table. Let n_{ii} be the number of individuals falling in the cell in the *i*th row and the *j*th column of the table. The sum of all the n_{ii} is, of course, *n*, the sample size. The row sums may be denoted by r_1 , r_2 , r_3 , r_4 , and the column sums by c_1 , c_2 , c_3 ; the sum of each of the sets is n. On applying the likelihood criterion to the test of the hypothesis, one finds that it rests on

location	yes	no	undecided	
east	n ₁₁	n ₁₂	n ₁₃	
south	n ₂₁	n ₂₂	n ₂₃	
midwest	n ₃₁	n ₃₂	n ₃₃	
west	n ₄₁	n ₄₂	n ₄₃	

Fig. 7. Contingency table.

expression (26), which is approximately distributed

$$2\left(n \log n + \sum_{ij} n_{ij} \log n_{ij} - \sum_{i} r_i \log r_i - \sum_{j} c_j \log c_j\right) \quad (26)$$

according to the chi-square distribution with six degrees of freedom. To test at the 0.05 level for the type I error, one would compute the above expression and compare it with 12.6, which is the value that marks 95% of the area under the chi-square distribution curve. If the expression turned out to be less than 12.6, one would accept the hypothesis of independence; otherwise, one would reject it.

Had there been *R* rows and *C* columns instead of four and three, one would have used the chi-square distribution with (R - 1)(C - 1) degrees of freedom. For a trivariate population there would have been a three-way contingency table and four hypotheses of independence that could have been tested; three of them assert that a given criterion of classification is independent of the other two, and the fourth asserts that all three are mutually independent.

Design of experiments. An experiment is performed to obtain information about the relations between several variables. For example, one may study the effect of storage temperature and duration of storage on the flavor of a frozen food. Three variables (flavor, temperature, and duration) are involved; one (flavor) is called the subject of the experiment; the other two are called factors which influence the subject. Sometimes the factors have intrinsic value in themselves; sometimes they are merely nuisance variables which must be taken into account because it is impossible to perform the experiment without them.

There exist in the statistical literature great numbers of specific experimental designs. These are patterns for making experimental observations; the actual construction of the designs requires quite advanced mathematics based on group theory, finite geometries, and combinatorial analysis. The mathematical problem is to find a pattern from which it is possible to extract the desired information and yet minimize the number of observations.

Experimental designs are most important to the experimenter when observations are expensive to make and when more than one factor is involved in the experiment. In the past it was believed that the best experimental procedure was to vary factors one at a time. Thus, in the frozen food experiment one might have held storage temperature constant and studied the effect of duration of storage only on flavor. Having determined that relationship, one would then hold duration constant and study the temperature effect. This procedure is not only wasteful of time and resources but may very well lead to erroneous conclusions because in all likelihood there is interaction between the two factors; that is, duration effect probably changes in a not obvious way when temperature is changed. Even if one believes that one can extrapolate accurately

the duration effect to other temperatures, there is much to be said for checking the extrapolation procedure in the experiment because it can probably be done at no additional cost if the experiment is well designed.

As an illustration of the use of a design, consider a very simple and useful one called the randomized block design. Suppose a manufacturer contemplates purchasing a machine which can be obtained from one of three sources, X, Y, or Z. He obtains one of each on trial in order to compare their performance using several operators from his own plant. Perhaps five operators, A, B, C, D, and E, are to be used in the proposed experiment, which is a two-factor experiment with performance being the subject and the two factors being humans and machines. The operators are not really a factor of interest in evaluating the machines but are, of course, necessary to the experiment.

It would be a mistake to use perhaps 15 operators, putting 5 on each machine, because operators differ in ability and a machine that turned out well might have been fortunate in having a good set of operators assigned to it. It is necessary that the machines be compared in as nearly equivalent circumstances as possible; in this case, that is done by having each operator operate every machine. The machines may then be compared in blocks (operators) which are individually homogeneous although they may be quite different among themselves.

The data are obtained by measuring production of each operator on each machine for a specified period of time and filling in the table of **Fig. 8**. It is essential that the order in which an operator works on the three machines be randomized (by tossing dice, for example) so that minor factors not taken into account in the design will not bias the results. If, for example, learning is an important factor in operating the machines, then the experiment must be three-factor and a more elaborate design is required.

In the resulting experimental observations the effects of humans and of machines are entangled, but there exists a computational technique for well-designed experiments known as the analysis of variance which will disentangle them. This enables the total variation to be broken down into parts, a variance attributable to machines, one to humans, and one to interactions; if the experiment were duplicated, one could also break out a variance corresponding to experimental error. One can estimate

machine	X	Y	Z
operator			
A			
В			
С			
D			
E			

Fig. 8. Data form for randomized block experiment.

the individual main effects of machines, the main effects of humans, and individual interaction effects. In experiments involving more factors, one may be able (depending on the experimental design selected) to estimate higher-order interactions such as the second-order interactions between the main effects of one factor and the first-order interactions of two other factors.

A common hypothesis in an analysis of variance is the null hypothesis that the variance associated with some factor or interaction is not larger than the experimental error variance; the test criterion is essentially the ratio of the two variances and has the so-called F distribution when the null hypothesis is true. Rejection of the null hypothesis implies that the factor or interaction in question had a significant effect on the subject of the experiment.

Regression and correlation. The regression problem is that of estimating certain unknown constants or parameters occurring in a function which relates several variables; the variables may be random or not. By far the most easily handled cases are those in which the function is linear in the unknown parameters, and it is worth considerable effort to transform the function to that form if at all possible.

The eventual adult height H of a 5-year-old boy may be quite well predicted by a linear function of three variables: his own present height B, his father's height F, and his mother's height M. The linear function is given in Eq. (27), where a, b, c, and d are

$$H = a + bB + cF + dM \tag{27}$$

the unknown parameters. The three variables B, F, and M are called independent variables, the variable H is called the dependent variable, and the parameters b, c, and d are often referred to as regression coefficients.

To estimate the parameters, one might draw a random sample of 5-year-old boys, measure their heights and those of their parents, then some years later measure their adult heights. One would then have sufficient data from which the parameters could be estimated by procedures entirely analogous to the methods for estimating location and scale parameters described above. In practice, one would not take so long to get the data but would use men whose childhood records were available. This method involves subtle sampling problems, however; for example, persons whose records are available may have been better cared for on the average and hence taller on the average.

The data which supply the estimates of regression coefficients can also be used to estimate the standard deviation, or standard error of regression, σ . Using that estimate and a table of the *t* distribution, one can compute a prediction interval (analogous to a confidence interval) which will have the desired probability of including the correct adult height of a given boy.

Tall fathers sometimes have short sons and short fathers sometimes have tall sons, but generally a father's height is a good indicator of his son's height; that is, fathers' heights and their sons' heights are positively correlated. For a different example: The price of a commodity in a free economy is negatively correlated with the supply of that commodity; there is not a fixed relation between price and supply, but as a general proposition one goes up when the other goes down.

Statisticians have developed measures of the degree of such imprecise relationships called coefficients of correlation. The most widely used one is the Pearson, or product moment, correlation, which is generally denoted by ρ . It measures the degree of linear association or correlation between two random variables of a bivariate or multivariate population and is defined as the mean over all the population of the product of the deviations of the two variables from their means divided by the product of their standard deviations. Mathematically, it is as shown in Eq. (28),

$$\rho = \int (x - \mu)(y - \nu)f(x, y) \, dx \frac{dy}{\sigma_x \sigma_y} \tag{28}$$

where *x* and *y* are the random variables, μ and ν their means, σ_x and σ_y their standard deviations, and f(x,y) their frequency distribution.

If the two variables are completely unrelated, then $\rho = 0$; if they have a fixed linear relation so that one can be calculated directly from the other, then $\rho = +1$ or $\rho = -1$ depending upon whether their relation is direct or reverse. Otherwise ρ will be some fraction between -1 and +1 with the fraction being near zero if there is poor correlation between them.

When the independent variables of a regression function are random variables, there is an equivalence between correlation coefficients and regression coefficients; after one set has been defined, the definition of the other follows automatically; mathematical formulas connecting them may be found in any statistics textbook.

Nonparametric inference. Most techniques of statistical inference rely on the central limit theorem or on sampling distributions derived from normal populations. They are practically always valid for large samples and are often valid for samples of intermediate size. However, there are occasions when one cannot rely on these techniques, particularly when samples are small or when some evident peculiarity of the population (such as marked asymmetry) makes ordinarily used sampling distributions suspect. In such instances one uses nonparametric methods, which are valid whatever form the population distribution might take.

Nonparametric techniques use the so-called order statistics, which are merely the sample observations arranged in ascending order of magnitude. One may let x_1 be the smallest sample observation, x_2 the next smallest, and so on with x_n being the largest. A basic theorem states that on the average the sample divides the population into n + 1 equal parts, that is, that 1/(n + 1) of the population lies between any two successive order statistics. Many nonparametric methods rest on this fact and its consequences.

As a simple illustration of a nonparametric estimate and confidence interval, one may consider the estimation of the population median, M, given an ordered sample of size 5 consisting of x_1, x_2, x_3, x_4 , and x_5 . The estimate of *M* among the observations is simply the central sample observation x_3 . The extreme observations x_1 and x_5 provide limits for a confidence interval for M; the probability level for the interval is calculated as follows. One-half of the population lies to the right of *M*, hence the probability is 1/2 that a randomly drawn observation will lie to its right. The probability is $(1/2)^5$ that all five observations will lie to its right. Similarly, the probability is $(1/2)^5$ that all five will lie to its left. In all other cases the sample will have at least one observation on each side of M; therefore, expression (29) can be formulated.

Probability that

$$(x_1 < M < x_5) = 1 - (1/2)^5 - (1/2)^5$$

= 30/32 \approx 0.94 (29)

See ANALYSIS OF VARIANCE; BIOMETRICS; DISTRIBU-TION (PROBABILITY); EXPERIMENT; QUALITY CON-TROL. Alexander M. Mood

Bibliography. K. A. Brownlee, *Statistical Theory* and Methodology in Science and Engineering, 2d ed., 1965, reprint 1984; W. J. Dixon and F. J. Massey, *Introduction to Statistical Analysis*, 4th ed., 1983; M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, vol. 2, 5th ed., 1991, vol. 3, 4th ed., 1983, reprint 1987, vol. 4, 1983; A. M. Mood, *Introduction to the Theory of Statistics*, 3d ed., 1974; D. H. Sanders, *Statistics: A First Course*, 5th ed., 1995; M. F. Triola, *Elementary Statistics*, 8th ed., 2000.

Statoblasts

Chitin-encapsulated bodies, resistant to freezing and a limited amount of desiccation. They serve as a special means of asexual reproduction in the Phylactolaemata, a class of fresh-water Bryozoa. They are 0.01-0.06 in. (0.26-1.5 mm) long, and their shape and structure are important to the taxonomy of the group.

Statoblasts are classified as follows: sessoblasts, those which attach to zooecial tubes (structures of the outer layer or zooecial of an individual in the colony) or the substratum; floatoblasts and spinoblasts, both of which have a float of "air" cells and therefore are free-floating; and piptoblasts, which are free but have no float (see **illus.**).

These bodies are produced in enormous quantities from spring to autumn. They develop by organization of masses of peritoneal cells and epidermal cells that bulge into the coelom. Each mass then secretes protective upper and lower chitinous valves, the rims of which often project peripherally. Statoblasts are therefore somewhat disk-shaped.

These bodies remain dormant for variable periods of time and serve to tide species over adverse ecological conditions, such as freezing or drying, which



Some types of statoblasts. (a) Spinoblast of Pectinatella magnifica. (b) Floatoblast of Plumatella repens. (c) Piptoblast or sessoblast of Fredericella sultana. (d) Sessoblast of Stolella indica.

kill the colony. During this time they may be dispersed over considerable distances, being carried by animals, floating vegetation, or the action of water currents. When environmental conditions become favorable, which is usually in the spring, statoblasts germinate and a zooid develops from the mass of cells lying between the two valves. Statoblasts of *Lophopodella* have germinated after 50 months of drying. *See* BRYOZOA. Sybil P. Parker

Staurolite

A nesosilicate mineral occurring in metamorphic rocks. The chemical formula of staurolite may be written as $A_4B_4C_{18}D_4T_8O_{40}X_8$, where $A = Fe^{2+}$, Mg, \square (\square = vacancy, \square > 2); B = Fe²⁺, Zn, Co, Mg, Li, Al, Fe³⁺, Mn²⁺, \Box ; C = Al, Fe³⁺, Cr, V, Ti; D = Al, Mg, \Box (\Box > 2); T = Si, Al; X = OH, F, O. Staurolite is an order-disorder series between an ordered monoclinic structure (with $\beta = 90.64^{\circ}$) and a disordered orthorhombic structure (with $\beta = 90^{\circ}$). The structure of staurolite consists of slabs of the kyanite structure [Al₂SiO₅] and layers of Fe-Al oxyhydroxide $[\sim^{[6]}Al_{0.67}$ ^[4]Fe₂ (OH)₂] alternating along [010], the monoclinic symmetry axis. The kyanite slab consists of edge-sharing chains of AlO₆ octahedra (Fig. 1a, regular dots) extending along the c direction, flanked by AlO₆ octahedra (random dots) and cross-linked by SiO₄ tetrahedra (crosses); only small amounts of Mg and Fe substitute for Al in this slab. The oxide-hydroxide layer consists of edgesharing chains of (Al, D)O₆ octahedra (Fig. 1b, regular dots) extending along the c axis, alternating with chains of $(Al, \Box)O_6$ octahedra (random dots) and (FeO₄) tetrahedra (crosses). In the latter chains, the $(Al,\Box)O_6$ octahedra share edges; the (FeO₄) tetrahedra share faces with adjacent (Al, D)O₆ octahedra, and polyhedra adjacent across a shared face cannot both be locally occupied. The extensive coupling of substitutions involving the cations Al, Fe²⁺, H and \Box (vacancy) gives rise to extensive short-range order, and local bond-valence arrangements are optimal for the formula $[Fe^{2+}_{4}Al_{18}$ - Si₈O₄₄(OH)₄]²⁺. However, at the long-range level, a crystal cannot have a net charge, and hence the principal heterovalent substitutions in staurolite $[Al \leftrightarrow Si, Mg \leftrightarrow Al, \Box \leftrightarrow H, Li \leftrightarrow Fe^{2+}, \Box \leftrightarrow Fe^{2+}]$ all act such as to reduce this charge to zero. Thus the chemical complexity of staurolite results from the interaction of long-range and short-range charge-balance requirements.

Staurolite occurs as well-formed, often-twinned, prismatic crystals (**Fig. 2**). It is brown-black, reddish brown, or light brown in color and has a vitreous to dull luster. Light color and dull luster can result from abundant quartz inclusions. There is no cleavage, specific gravity is 3.65–3.75, and hardness is 7–7.5 (Mohs scale).

Staurolite occurs in pelitic (former shale) schists of medium grade (degree) of metamorphism. In typical rocks, it forms as temperature increases from the reaction of muscovite, garnet, and chlorite to form staurolite, biotite, and water, or from the breakdown



Fig. 1. Staurolite crystal structure: (a) kyanite slab and (b) oxide-hydroxide layer, where a is the shortest axis and c is the longest axis.



Fig. 2. Staurolite crystal habits. (After C. Klein and C. S. Hurlbut, Jr., Manual of Mineralogy, 21st ed., John Wiley, 1993)

of chloritoid to form staurolite, garnet, and water at temperatures of about 930°F (500°C) as shown by laboratory experiments. With increasing temperature in such rocks, staurolite reacts with muscovite to form garnet, biotite, and aluminosilicate minerals (kyanite, sillimanite, or andalusite, depending on pressure) at temperatures near 1110° F (600°C). Thus, most staurolite forms between 930 and 1110°F (500 and 600°C). Staurolite may form at pressure as low as 2.5 kilobars (250 megapascals), corresponding to depths in the Earth's crust of 5.5. mi (9 km); at lower pressure, cordierite and andalusite take its place in pelitic schists or contact metamorphic rocks. Staurolite may occur at pressures corresponding to the deepest levels in the Earth's crust (>10 kbar). However, it is less common and tends to be Mg-rich in deep crustal rocks. Typical minerals occurring with staurolite are quartz, micas (muscovite and biotite), garnet (almandine), tourmaline, and kyanite, sillimanite, or andalusite. Staurolite is common where pelitic schists reach medium-grade metamorphism. Examples are the Swiss and Italian Alps (notable at Saint Gotthard, Switzerland), and all the New England states, Virginia, the Carolinas, Georgia, New Mexico, Nevada, and Idaho. See META-MORPHISM; SILICATE MINERALS. Frank C. Hawthorne

Bibliography. W. A. Deer, R. A. Howie, and J. Zussman, *Rock Forming Minerals: Orthosilicates*, vol. 1A, 1997; F. C. Hawthorne et al., The crystal chemistry of staurolite, *Can. Mineral.*, 31:551-616, 1993; M. J. Holdaway, B. L. Dutrow, and P. Shore, A model for the crystal chemistry of staurolite, *Amer. Mineral.*, 71:1142-1159, 1986; C. Klein and C. S. Hurlbut, Jr., *Manual of Mineralogy*, 21st ed., 1993.

Stauromedusae

An order of the cnidarian class Scyphozoa, usually found in circumpolar regions. *Haliclystus auricula* is typical. The egg develops into a planula which can only creep since it lacks cilia. The planula changes into a polyp that metamorphoses directly into a combined polyp and medusa form. The medusa is composed of a cuplike bell called a calyx (medusan part) and a stem (polyp part) which terminates in a pedal disk (see **illus.**). The calyx is eight-sided and has eight groups of short, capped tentacles and eight sensory bodies, called anchors (rhopalioids), on its margin. The mouth, situated at the center of the calyx, has four thin lips and leads to the stomach in which gastral filaments are arranged in a row on either side of each interradius. Though sessile, the medusa can



Stauromedusae. (a) Lucernaria. (b) Haliclystus. (After L. H. Hyman, The Invertebrates, vol. 1, McGraw-Hill, 1940)

move in a leechlike fashion by alternate attachment and release of the pedal disk, using the substratum as an anchor. *See* SCYPHOZOA. Tohru Uchida

Steam

Water vapor, or water in its gaseous state. Steam is the most widely used working fluid in external combustion engine cycles, where it will utilize practically any source of heat, that is, coal, oil, gas, nuclear fuel (uranium and thorium), waste fuel, and waste heat. It is also extensively used as a thermal transport fluid in the process industries and in the comfort heating and cooling of space. The universality of its



Fig. 1. Vapor-tension curve for water. 1 lb/in.² abs. = 6.895 kPa, absolute pressure. $^{\circ}C = (^{\circ}F - 32)/1.8$.

Thermal properties of steam (approximate)*						
		Enthalpy, Btu/lb [‡]				
Pressure, Ib/in. ² abs. (kPa)	Temperature, °F (°C)	Saturated liquid [†]	Latent heat	Saturated vapor [†]		
1 (6.895)	102 (39)	69.7	1036	1106		
10 (68.95)	193 (89)	161	982	1143		
14.7 (101.3)	212 (100)	180	970	1150		
100 (689.5)	328 (164)	298	889	1187		
1000 (6895)	545 (285)	542	649	1192		
3208 (22,120)	705 (374)	906	0	906		

*For more exact values see J. H. Keenan and F. G. Keyes, Thermodynamic Properties of Steam, Wiley, 1936, or American Society of Mechanical Engineers, Thermodynamic and Transport Properties of Steam (tables and charts), 1967.

[†]Enthalpy is measured from saturated liquid at 32°F (0°C). ^{*}1 Btu/lb = 2326 J/kg.

availability and its highly acceptable, well-defined physical and chemical properties also contribute to the usefulness of steam. The engineering profession has developed so much evaporative and condensing equipment for changing the state of water that its properties can be utilized from pressures ranging from 0.18 in. Hg or 0.61 kilopascal (ice point) to supercritical (>3208.2 lb/in.² absolute or 22.120 megapascals absolute pressure and 705.47°F or 374.15°C). These temperature ranges are readily adaptable for use with ferrous and nonferrous metal structures at costs which dominate the design, selection, and use of many applications.

The temperature at which steam forms depends on the pressure in the boiler. The steam formed in the boiler (and conversely steam condensed in a condenser) is in temperature equilibrium with the water. The relation is rigorously specified by the vapor tension curve (Fig. 1). (The table gives a few selected values in the first two columns.) Under these conditions, with steam and water in contact and at the same temperature, the steam is termed saturated. Steam can be entirely vapor when it is 100% dry, or it can carry entrained moisture and be wet. It may also be contaminated, as in all practical and engineering applications, with other gases, especially air, in which case it is a mixture and Dalton's law prevails. After the steam is removed from contact with the liquid phase, the steam can be further heated without changing its pressure. If initially wet, the additional heat will first dry it and then raise it above its saturation temperature. This is a sensible heat addition, and the steam is said to be superheated.

The heat energy in steam can be divided into three parts: (1) enthalpy of the liquid required to raise the water from its initial temperature (usually considered to be 32° F or 0° C) to the boiling temperature, (2) enthalpy of vaporization required to convert the water to steam at the boiling temperature, and (3) enthalpy of superheat that raises the steam to its final temperature. As the steam performs its thermodynamic function of giving up its heat energy, it loses its superheat, becomes wet, and finally condenses to hot water. While wet, the steam has a quality that decreases as the percent of dry, saturated steam present

in the wet steam decreases. Dry steam has a quality of 100% (see the table and **Fig. 2**). *See* ENTHALPY.

Properties as a gas. Superheated steam at temperatures well above the boiling temperature for the existing steam pressure follows closely the laws of a perfect gas. Thus, $pv \approx 85.8T$ and $k \approx 1.3$, where p is pressure in lb/ft², v is volume in ft³/lb, T is absolute temperature in °R, and k is the ratio of specific heat at constant pressure to specific heat at constant volume. (In SI units, $p'v' \approx 461.5 T$; where p' is pressure in Pa, v' is volume in m³/kg, and T' is absolute temperature in K.) However, the behavior of dry, saturated steam departs from that of a perfect gas, and wet steam is a mixture. Therefore, the properties of steam near its vaporization temperature are determined experimentally.



Fig. 2. Temperature-enthalpy chart for steam. 1 Btu/lb = 2326 J/kg. $^{\circ}C = (^{\circ}F - 32)/1.8$.


Fig. 3. Temperature-entropy chart for steam. $^{\circ}C = (^{\circ}F - 32)/1.8$; 1 Btu/lb $^{\circ}F = 4186.8$ J/kg $^{\circ}C$; 1 Btu/lb = 2326 J/kg; 1 ft³/lb = 62.4 \times 10⁻³ m³/kg; 1 lb/in.² = 6.895 kPa. (After J. H. Keenan and F. G. Keyes, Thermodynamic Properties of Steam, John Wiley and Sons, 1936)

Data can be presented in tabular form, giving pressure, volume, entropy, enthalpy, and temperature for saturated liquid (water at the boiling point), saturated steam, and steam vapor at various temperatures of superheat. Alternatively, thermodynamic properties of steam can be presented diagrammatically. For analysis of thermodynamic cycles, the temperatureentropy chart is widely used (**Fig. 3**). Area on this chart is proportional to heat energy. At the critical point, steam condenses directly into water without releasing energy. In the uncharted area the pressure is higher than is usually encountered in commercial practice.

For engineering application the Mollier diagram presents steam data in a convenient form (**Fig. 4**). In the wet region, lines of constant temperature and of constant pressure are straight and coincide with each other. Total enthalpies above 32° F (0°C) are plotted as ordinates and total entropies as abscissas.

The specific volume of steam varies widely with temperature and pressure. At 1 lb/in.² absolute (6.895 kPa absolute pressure) the specific volume of saturated steam is 334 ft³ (20.85 m³/kg), at atmospheric pressure it is 26.8 (1.29), at 1000 lb/in.² absolute (6.895 MPa) it is 0.45 (28×10^{-3}), at 2000 lb/in.²

absolute (13.79 MPa) it is 0.19 (12×10^{-3}), and at the critical pressure (3208 lb/in.² absolute or 22.12 MPa) it is 0.05 (3×10^{-3}). Superheating substantially increases values of specific volume of a steam at saturation conditions.

Viscosity of steam increases with temperature and pressure. Saturated steam at atmospheric pressure has a viscosity of about 2.6×10^7 (lb)(s)/ft² (1.2 × 10⁹ Pa·s). At 1000°F (538°C) and 2000 lb/in.² (13.79 MPa) pressure it is 13.4×10^7 (lb)(s)/ft² (6.4 × 10⁹ Pa·s). Similarly, the heat conductivity of steam increases with temperature and pressure. The thermal conductivities for the conditions for which viscosities are given are about 14, 37, and 109 Btu/(h)(ft)(°F) or 24, 64, and 189 W/(m)(°C).

Application. Chiefly because of its availability, but also because of its nontoxicity, steam is widely used as the working medium in thermodynamic processes. It has a uniquely high latent heat of vaporization: 1049 Btu/lb (2.440 MJ/kg) at 1 in. Hg abs. (3.39 kPa) and 79°F (26°C), 970 (2.256 MJ/kg) at 14.7 lb/in.2 abs. (101 kPa) and 212°F (100°C), 889 (2.068 MJ/kg) at 1000 lb/in.2 abs. (6.895 MPa) and 328°F (164°C), 649 (1.510 MJ/kg) at 1000 lb/in.² abs. (6.895 MPa) and 545°F (285°C), and 463 (1.077 MJ/kg) at 2000 lb/in.² abs. (13.79 MPa) and 636°F (336°C). Steam has a specific heat in the vicinity of half that of water. For comparison, its specific heat is about twice that of air and comparable to that of ammonia. Except for a few gases such as hydrogen, with a specific heat seven times that of steam, the specific heat of steam is relatively high so that it can



Fig. 4. Mollier diagram for steam. 1 Btu/lb = 2326 J/kg; 1 Btu/lb $^{\circ}F = 4186.8$ J/kg $^{\circ}C$; $^{\circ}C = (^{\circ}F - 32)/1.8$; 1 lb/in.² = 6.895 kPa. (After Mollier's Steam Tables and Diagrams, Pitman, 1927)

carry more thermal energy at practical temperatures than can other usable gases. *See* BOILER; BOYLE'S LAW; CHARLES' LAW; DALTON'S LAW; ENTROPY; STEAM ENGINE; STEAM-GENERATING UNIT; STEAM HEATING; STEAM TURBINE; THERMODYNAMIC CYCLE; THERMO-DYNAMIC PRINCIPLES; WATER. Theodore Baumeister

Bibliography. American Society of Heating, Refrigeration, and Air-Conditioning Engineers, *Handbook: Applications*, 1993, *Equipment*, 1992, *Fundamentals*, 1994, *Systems*, 1992; E. A. Avallone and T. Baumeister III (eds.), *Marks' Standard Handbook for Mechanical Engineers*, 10th ed., 1996; J. H. Keenan et al., *Steam Tables: Thermodynamic Properties of Water Including Vapor, Liquid, and Solid Phases*, 1978; K. S. Pitzer, *Thermodynamics*, 3d ed., 1994; E. Schmidt, *Properties of Water and Steam*, 1981.

Steam condenser

A heat-transfer device used for condensing steam to water by removal of the latent heat of steam and its subsequent absorption in a heat-receiving fluid, usually water, but on occasion air or a process fluid. Steam condensers may be classified as contact or surface condensers.

In the contact condenser, the condensing takes place in a chamber in which the steam and cooling water mix. The direct contact surface is provided by sprays, baffles, or void-effecting fill such as Rashig rings. In the surface condenser, the condensing takes place separated from the cooling water or other heatreceiving fluid (or heat sink). A metal wall, or walls, provides the means for separation and forms the condensing surface.

Both contact and surface condensers are used for process systems and for power generation serving engines and turbines. Modern practice has confined the use of contact condensers almost entirely to such process systems as those involving vacuum pans, evaporators, or dryers, and to condensing and dehumidification processes inherent in vacuumproducing equipment such as steam jet ejectors and vacuum pumps. The steam surface condenser is used chiefly in power generation but is also used in process systems, especially in those in which condensate recovery is important. Air-cooled surface condensers are used increasingly in process systems and in power generation when the availability of cooling water is limited.

Water-cooled surface condensers used for power generation are of the high-vacuum type. Their main purpose is to effect a back pressure at the turbine exhaust. To achieve an economical station heat rate and fixed cost, these condensers must be designed for high heat-transfer rates, minimum steam-sidepressure loss, and effective air removal. The usual power plant steam condenser incorporates a single steam-condensing chamber into which the steam turbine exhausts. Multipressure condensers, having two or more steam-condensing chambers operating at different pressures, have been found to be more economical than single-pressure condensers for large installations, which generally use cooling towers as the final heat sink. Most multipressure condensers are compartmented internally, on the steamcondensing side, with the cooling water arranged to flow through the tubes in one direction from inlet to outlet. The mean cooling water temperature progressively increases as the water flows from compartment to compartment, with back pressure on the turbine increasing proportionately. Economical performance is effected with multipleexhaust pressure turbines when each exhaust pressure section of the turbine is designed for the back pressure produced in its respective pressure section in the condenser. A rise in water temperature in the order of 20-30°F (11-17°C) is necessary to effect required pressure differences in the turbine.

Air-cooled surface condensers used for power generation are designed to operate at higher back pressures than water-cooled condensers. This is consistent with their lower overall heat-transfer rate, which is of the order of 10-12 Btu/($^{\circ}F$)(ft²)(h) $[57-68 \text{ W/(°C)(m^2)}]$ in contrast with the normal 600-700 Btu/($^{\circ}$ F)(ft²)(h) [3.4-4.0 kW/($^{\circ}$ C)(m²)] characteristic of water-cooled steam condensers. To compensate for these large differences in heat transfer, greater temperature differences are required between the condensing steam and the cooling air than those needed with conventional water cooling. Steam is condensed on the inside of the heattransfer tubes, and airflow is usually produced by fans across their outside surface. The tubes are of the extended-surface type with fins on the outside. The normal ratio of outside to inside surface is on the order of 10 or 12 to 1. The use of air as the cooling medium for steam condensers used in smaller industrial steam power plants is increasing rapidly. Air-cooled steam condensers for the production of electric power have been confined to smaller plants; however, with the decreasing availability of water for cooling, the application of air-cooled steam condensers to the larger steam electric plants may be anticipated.

Condenser sizes have increased with the increase in size of turbine generators. The largest, in a single shell, is approximately 485,000 ft² (45,000 m²); the steam-condensing space occupies a volume of 80,000 ft³ (2265 m³). Large units have been built with the surface divided equally among two or three shells. Power plant condensers require 35-100 lb of cooling water to condense 1 lb (35-100 kg for 2 kg) of steam, or, if air-cooled, 2500-5000 ft³ of air for 1 lb (150-300 m³ for 1 kg) of steam. Normally, about 0.5 ft² (0.05 m²) of surface is required for watercooled condensers, and about 10 ft² (0.9 m²) of inside tube surface for air-cooled condensers is required for each kilowatt of generating capacity. *See* STEAM; STEAM TURBINE; VAPOR CONDENSER. Joseph F. Sebald

Bibliography. H. P. Bloch, *A Practical Guide to Steam Turbine Technology*, 1995; J. P. Holman, *Heat Transfer*, 9th ed., 2001.

Steam electric generator

An alternating-current (ac) synchronous generator driven by a steam turbine for 50- or 60-Hz electrical generating systems operation. *See* STEAM TURBINE.

Basic operation and construction. The synchronous generator is made of two parts: a stator (stationary) and a rotor (rotating) [see illus.]. The stator is a cylindrical steel frame, inside of which a cylindrical iron core made of thin insulated laminations is mounted on a support system. The iron core has equally spaced axial slots on its inside circumference, and wound within the core slots is a stator winding. The stator winding copper is electrically insulated from the core. The rotor consists of a forged solid steel shaft. Wound into axial slots on the outside diameter of the shaft is a copper rotor winding that is held in the slots with wedges. Retaining rings support the winding at the rotor body ends. The rotor winding, referred to as the field, is electrically insulated from the shaft and is arranged in pole pairs (always an even number) to form the magnetic field which produces the flux. The rotor shaft (supported by bearings) is coupled to the steam turbine, and rotates inside the stator core. See ELECTRIC ROTATING MACHINERY; MAGNETIC FLUX; WINDINGS IN ELECTRIC MACHINERY

The stator winding (armature) is connected to the ac electrical transmission system through the bushings and output terminals. The rotor winding (field) is connected to the generator's excitation system. The excitation system provides the direct-current (dc) field power to the rotor winding via carbon brushes riding on a rotating collector ring mounted on the generator rotor. The synchronous generator's output voltage amplitude and frequency must remain constant for proper operation of electrical load devices. During operation, the excitation system's voltage regulator monitors the generator's output voltage and current. The voltage regulator controls the rotor winding dc voltage to maintain a constant generator stator output ac voltage, while allowing the stator current to vary with changes in load. Field windings typically operate at voltages between 125 and 575 V dc. The synchronous generator's output frequency is directly proportional to the speed of the rotor,

Frequency = (rpm \times number of poles) \div (120)

and the speed of the generator rotor is held constant by a speed governor system associated with the steam turbine. *See* ALTERNATING-CURRENT GEN-ERATOR; GENERATOR.

The first commercial use of the ac synchronous generator was in the mid-1880s. Synchronous generators range in size from a few kilovolt-amperes to 1,650,000 kVA. 60-Hz steam-driven synchronous generators operate at speeds of either 3600 or 1800 rpm; for 50-Hz synchronous generators these speeds would be 3000 or 1500 rpm. These two- and four-pole generators are called cylindrical rotor units. For comparison, water (hydro)-driven and air-driven synchronous generators operate at lower speeds,



Cross section of a typical hydrogen- and water-cooled, two-pole, 3600-rpm, steam electric generator. The dimensions shown are for a 900,000-kVA synchronous generator. The hydrogen coolers are not shown.

some as low as 62 rpm (116 poles). The stator output voltage of large (generally greater than 100,000 kVA) units ranges 13,800-27,000 V. The operating characteristics of the steam-turbine-driven synchronous generators are defined in IEEE Std. C50.13-2005, *IEEE Standard for Cylindrical-Rotor 50 Hz and 60 Hz Syncbronous Generators Rated 10 MVA and Above. See* ELECTRIC POWER GENERATION; HYDRO-ELECTRIC GENERATOR.

The physical size of a steam generator is determined by the transportation limitations of size and weight. For the generator shown in the illustration, the stator would weigh about 800,000 lb (363,000 kg), and the rotor about 140,000 lb (64,000 kg).

Cooling. There are five sources of heat loss in a synchronous generator: stator winding resistance, rotor winding resistance, core, windage and friction, and stray losses. Removing the heat associated with these losses is the major challenge to the machine designer.

As the demand for power increased, generators grew in output rating and physical size. The cooling requirements for the stator windings, rotor windings, and core increase proportionally to the cube of the machine size. The early synchronous generators were air-cooled. Later, air-to-water coolers were required to remove the heat.

In 1937, hydrogen gas was introduced as a cooling medium in place of air. The generator frame had to be sealed against hydrogen leakage. The first units were filled with pure hydrogen at pressures of 0.5-5 psi (3.5-35 kPa), with a hydrogen purity of greater than 97%. In today's large generators, hydrogen pressures of 75 psi (520 kPa) are common. An elaborate seal-oil system (gland seal) is used to form a hydrogen seal near the bearings where the rotating shaft exits the frame. The seal-oil system operates at a pressure that is slightly higher than the hydrogen pressure. Hydrogen is the most effective gas available for generator cooling, and is used because it removes approximately ten times more heat than air; its density is one-tenth that of air; windage losses are lower; hydrogen is a better electrical insulator; it is a nonoxidizing agent; and pure hydrogen will not support combustion in the case of a winding failure. The hydrogen is circulated throughout the stator frame by blowers or fans mounted on the rotor, and is cooled by hydrogen-to-water heat exchangers mounted within the stator frame. In the early machines, the hydrogen gas cooled the surface of the stator and rotor windings (as did air), relying on conduction for heat transfer through the winding insulation. This cooling method is called indirect cooling. For larger machines, the stator windings were designed with hydrogen cooling passages manufactured within the copper winding, which allowed the hydrogen to flow through the winding in direct contact with the copper. These designs were called directly cooled windings. The rotor windings were also directly cooled with hydrogen gas. Helium is not used because it has higher density, has higher specific heat, and costs five times more than hydrogen. See HYDROGEN.

Synchronous generators continued to increase in output, and in 1957 oil was introduced as a cooling medium for stator windings. Oil was quickly replaced by water in 1961. The stator winding losses are responsible for about 25-40% of the total losses. The copper windings were manufactured with hollow copper conductors (strands). Water flows through the conductors in direct contact with the copper windings, thus removing the heat. Water enters and leaves the stator winding through Teflon hoses, which electrically insulate the steel water piping from the stator winding copper. A stator cooling water system contains a storage tank and pumps to move the water through the stator winding, filters, strainers, heat exchangers, and demineralizers to maintain low conductivity. The demineralizer contains a resin bed which removes minerals and other conductive impurities from the James R. Michalec water.

Bibliography. T. C. Elliott, K. Chen, and R. Swanekamp (eds.), Standard Handbook of Powerplant Engineering, McGraw-Hill, New York, 2d ed., 1998; A. E. Fitzgerald, C. Kingsley, Jr., and S. D. Umans, Electric Machinery, McGraw-Hill, New York, 6th ed., 2003; A. J. Gonzalez et al., Monitoring and Diagnosis of Turbine-Driven Generators, Prentice Hall, Englewood Cliffs, NJ, 1995; Institute of Electrical and Electronics Engineers, IEEE Guide for Operation and Maintenance of Turbine Generators, IEEE Standard 67-1990, 1990; Institute of Electrical and Electronics Engineers, IEEE Standard Dictionary of Electrical and Electronics Terms, 7th ed., IEEE Std. 100-2001, 2001; Institute of Electrical and Electronics Engineers, IEEE Standard for Cylindrical-Rotor 50 Hz and 60 Hz Synchronous Generators Rated 10 MVA and Above, IEEE Std. C50.13-2005, 2005.

Steam engine

A machine for converting the heat energy in steam to mechanical energy of a moving mechanism, for example, a shaft. The steam engine dominated the industrial revolution and made available a practical source of power for application to stationary or transportation services. The steam power plant could be placed almost anywhere, whereas other means of power generation were more restricted, experiencing such site limitations as an elevated water supply, wind, animal labor, and so on. The steam engine can utilize any source of heat in the form of steam from a boiler. It was developed in sizes which ranged from that of children's toys to 25,000 hp (18.6 MW), and it was adaptable to pressures up to 200 lb/in.² (1.4 megapascals). It reached its zenith in the nineteenth century in stationary services such as drives for pumping plants; drives for air compressor and refrigeration units; power supply for factory operations with shafting for machine shops, rolling mills, and sawmills; and drives for electric generators as electrical supply systems were perfected. Its adaptability to portable and transportation services rested largely on its development of full-rated torque at any speed from rest to full throttle; its speed variability at the will of the operator; and its reversibility, flexibility, and dependability under the realities of stringent service requirements. These same features favored its use for many stationary services such as rolling mills and mine hoists, but the steam engine's great contribution was in the propulsion of small and large ships, both naval and merchant. Also, in the form of the steam locomotive, the engine made the railroad the practical way of land transport. Most machine elements known today had their origin in the steam engine: cylinders, pistons, piston rings, valves and valve gear crossheads, wrist pins, connecting rods, crankshafts, governors, and reversing gears. *See* LO-COMOTIVE.

The twentieth century saw the practical end of the steam engine. The steam turbine with (1) its high speed (for example, 3600 rpm); (2) its utilization of maximum steam pressures (2000-5000 lb/in.2 or 14-34 MPa), maximum steam temperatures ($\sim 1100^{\circ}$ F or 600°C), and highest vacuum (~29 in. Hg or 98 kilopascals); and (3) its large size (~1,000,000 kW) led to such favorable weight, bulk, efficiency, and cost features that it replaced the steam engine as the major prime mover for electric generating stations. The internal combusion engine, especially the high-speed automotive types which burn volatile (gasoline) or nonvolatile (diesel) liquid fuel, offers a self-contained, flexible, low-weight, low-bulk power plant with high thermal efficiency that has completely displaced the steam locomotive with the diesel locomotive and marine steam engines with the motorship and motorboat. It is the heart of the automotive industry, which produces in a year 10,000,000 vehicles that are powered by engines smaller than 1000 hp (750 kW). Because of the steam engine's weight and speed limitations, it was excluded from the aviation field, which has become the exclusive preserve of the internal combustion piston engine or the gas turbine. See DIESEL ENGINE; GAS TURBINE; INTERNAL COMBUSTION ENGINE; STEAM TURBINE; TURBINE.

Cylinder action. A typical steam reciprocating engine consists of a cylinder fitted with a piston (**Fig. 1**). A connecting rod and crankshaft convert the piston's to-and-fro motion into rotary motion. A flywheel tends to maintain a constant-output angular velocity in the presence of the cyclically changing steam pressure on the piston face. A D slide valve



Fig. 1. Principal parts of horizontal steam engine.



Fig. 2. Single-ported slide valve on counterflow doubleacting cylinder.



Fig. 3. Events during one cycle of a piston operation. (a) In ideal engine. (b) As depicted on the indicator card of a noncondensing steam engine.

admits high-pressure steam to the cylinder and allows the spent steam to escape (**Fig. 2**). The power developed by the engine depends upon the pressure and quantity of steam admitted per unit time to the cylinder.

Indicator card. The combined action of valves and piston is most conveniently studied by means of a pressure-volume diagram or indicator card (**Fig. 3**). The pressure-volume diagram is a thermodynamic analytical method which traces the sequence of phases in the cycle. It may be an idealized operation (Fig. 3*a*), or it may be an actual picture of the phenomena within the cylinder (Fig. 3*b*) as obtained with an instrument commonly known as a steam engine indicator. This instrument, in effect, gives a graphic picture of the pressure and volume for all phases of steam admission, cutoff, expansion, release, exhaust, and compression. It is obtained as

the engine is running and shows the conditions which prevail at any instant within the cyclinder. The indicator is a useful instrument not only for studying thermodynamic performance, but for the equally important operating knowledge of inlet and exhaust valve leakage and losses, piston ring tightness, and timing correctness. *See* THERMODYNAMIC PROCESSES.

The net area of the indicator card shows, thermodynamically, the work done in the engine cylinder. By introducing the proper dimensional quantities, power output can be measured. Thus if the net area within the card is divided by the length, the consequent equivalent mean height is the average pressure difference on the piston during the cycle and is generally called the mean effective pressure p, usually expressed in lb/in.² With a cylinder dimension of piston area a (in in.²), length of piston stroke l (in ft), and n equal to the number of cycles completed per minute, the equation given below holds. It has often

Indicated horsepower =
$$\frac{plan}{33,000}$$

been referred to as the most important equation in mechanical engineering.

Engine types. Engines are classified as single- or double-acting, and as horizontal (Fig. 1) or vertical depending on the direction of piston motion. If the steam does not fully expand in one cylinder, it can be exhausted into a second, larger cylinder to expand further and give up a greater part of its initial energy. Thus, an engine can be compounded for double or triple expansion. In counterflow engines, steam enters and leaves at the same end of the cylinder; in uniflow engines, steam enters at the end of the cylinder and exhausts at the middle.

Steam engines can also be classed by functions, and are built to optimize the characteristics most desired in each application. Stationary engines drive electric generators, in which constant speed is important, or pumps and compressors, in which constant torque is important. Governors acting through the valves hold the desired characteristic constant. Marine engines require a high order of safety and dependability.

Valves. The extent to which an actual steam piston engine approaches the performance of an ideal engine depends largely on the effectiveness of its valves. The valves alternately admit steam to the cylinder, seal the cylinder while the steam expands against the piston, and exhaust steam from the cylinder. The many forms of valves can be grouped as sliding valves and lifting valves (**Fig. 4**). *See* CARNOT CYCLE; THERMODYNAMIC CYCLE; VALVE.

D valves (Fig. 2) are typical sliding valves where admission and exhaust are combined. A common sliding valve is the rocking Corliss valve; it is driven from an eccentric on the main shaft like other valves but has separate rods for each valve on the engine. After a Corliss valve is opened, a latch automatically disengages the rod and a separate dashpot abruptly closes the valve. Exhaust valves are closed



double-beat poppet valve seated in cage

Fig. 4. Steam engine valves in closed positions. The arrows indicate the path steam will travel when the valves (colored parts) are open.

by rods, as are other sliding valves.

Lifting values are more suitable for use with hightemperature steam. They, too, are of numerous forms. The poppet value is representative.

Valves are driven through a crank or eccentric on the main crankshaft. The crank angle is set to open the steam port near dead center, when the piston is at its extreme position in the cylinder. The angle between valve crank and connecting-rod crank is slightly greater than 90° , the excess being the angle of advance. So that the valves will open and close quickly, they are driven at high velocity with consequently greater travel than is necessary to open and close the ports. The additional travel of a sliding valve is the steam lap and the exhaust lap. The greater the lap, the greater the angle of advance to obtain the proper timing of the valve action.

Engine power is usually controlled by varying the period during which steam is admitted. A shifting eccentric accomplishes this function or, in releasing Corliss and in poppet valves, the eccentric is fixed and cutoff is controlled through a governor to the kickoff cams or to the latch that allows the valves to be closed by their dashpots.

For high engine efficiency, the ratio of cylinder volume after expansion to volume before expansion should be high. The volume before expansion into which the steam is admitted is the volumetric clearance. It may be determined by valve design and other structural features. For this reason, valves and ports are located so as not to necessitate excessive volumetric clearance. *See* BOILER; STEAM.

Theodore Baumeister Bibliography. E. A. Avallone and T. Baumeister III (eds.), *Marks' Standard Handbook for Mechanical*

Steam-generating furnace

Engineers, 10th ed., 1996.

An enclosed space provided for the combustion of fuel to generate steam. The closure confines the products of combustion and is capable of withstand-



Water-cooled furnace-wall constructions. (a) Tangent tube. (b) Welded membrane and tube. (c) Flat stud and tube. (d) Full stud and refractory-covered tube. (e) Tube and tile. (f) Tubes backed by refractory. (After T. Baumeister, ed., Marks' Standard Handbook for Mechanical Engineers, 8th ed., McGraw-Hill, 1978)

ing the high temperatures developed and the pressures used. Its dimensions and geometry are adapted to the rate of heat release, to the type of fuel, and to the method of firing so as to promote complete burning of the combustible and suitable disposal of ash. In water-cooled furnaces the heat absorbed materially affects the temperature of gases at the furnace outlet and contributes directly to the generation of steam. *See* STEAM.

Furnace walls. Prior to 1925, most furnaces were constructed of firebrick. As the capacity and the physical size of boiler units increased and the suspension burning of pulverized coal was developed, limitations were imposed by the height of the refractory walls that could be made self-supporting at high temperatures and by the inability of refractories to resist the fluxing action of molten fuel ash. Thus refractory is used primarily for special-purpose furnaces, for the burner walls of small boilers, or for the arches of stoker-fired boilers to promote the ignition of the fuel. *See* REFRACTORY.

The limitations of refractory constructions can be extended somewhat by cooling the brickwork with air flowing through channels in the structure, or by sectionalizing the wall into panels and transferring the structural load to external air-cooled steel or castiron supporting members. The heat absorbed can be recovered by using the cooling air for combustion, and the low rate of heat transfer through the refractory helps to maintain high furnace temperatures, thus accelerating ignition and the burning of the fuel.

The low tensile strength of refractories restricts the geometrical shapes that can be built, making it difficult to provide overhanging contours or roof closures. Thus sprung arches or shaped tiles suspended from steel must be used. Many refractory mixtures having air-setting or hydraulic-setting properties are available, and they can be used to form monolithic structures by ramming, guniting, or pouring in forms.

Water-cooling of all or most of the furnace walls is almost universally used in boiler units of all capacities and for all types of fuel and methods of firing. The water-cooling of the furnace walls reduces the transfer of heat to the structural members and, consequently, their temperatures are within the limits that provide satisfactory strength and resistance to oxidation. Water-cooled tube constructions facilitate large furnace dimensions and optimum arrangements of the roof, hopper, and arch, as well as the mountings for the burners and the provision for screens, platens, or division walls to increase the amount of heat-absorbing surface exposed in the combustion zone. External heat losses are small and are further reduced by the use of insulation.

Furnace wall tubes are spaced on close centers to obtain maximum heat absorption and to facilitate ease of ash removal (see **illus.**). The so-called tangent tube construction, wherein adjacent tubes are almost touching with only a small clearance provided for erection purposes, has been frequently used. However, most new boiler units use membrane-tube walls, in which a steel bar or membrane is welded between adjacent tubes. This construction facilitates the fabrication of water-cooled walls in large shopassembled tube panels. Less effective cooling is obtained when the tubes are wider spaced and extended metal surfaces, in the form of flat studs, are welded to the tubes; if even less cooling is desired, the tube spacing can be increased and refractory installed between or behind the tubes to form the wall enclosure.

Furnace walls must be adequately supported, with provision for thermal expansion and with reinforcing buckstays to withstand the lateral forces resulting from the difference between the furnace pressure and the surrounding atmosphere. The enclosure walls must prevent air infiltration when the furnace is operated under suction, and must prevent gas leakage when the furnace is run at pressures above atmospheric.

Additional furnace cooling, in the form of tubular platens, division walls, or wide-spaced screens, often is used; in high heat input zones the tubes may be protected by refractory coverings anchored to the tubes by studs. *See* FURNACE CONSTRUCTION.

Heat transfer. Heat-absorbing surfaces in the furnace receive heat from the products of combustion and, consequently, lower the furnace gas temperature. The principal mechanisms of heat transfer take place simultaneously and include intersolid radiation from the fuel bed or fuel particles, nonluminous radiation from the products of combustion, convection from the furnace gases, and conduction through any deposits and the tube metal. The absorption effectiveness of the surface is greatly influenced by the deposits of ash or slag.

Analytical solutions to the transfer of heat in the

furnaces of steam-generating units are extremely complex, and it is very difficult, if not impossible, to calculate furnace-outlet gas temperatures by theoretical methods. Nevertheless, the furnace-outlet gas temperature must be predicted accurately because this temperature determines the design of the remainder of the unit, particularly that of the superheater. Calculations therefore are based upon test results supplemented by data accumulated from operating experience and judgment predicated upon knowledge of the principles of heat transfer and the characteristics of fuels and slags. *See* HEAT TRANSFER; STEAM-GENERATING UNIT. George W. Kessler

Bibliography. E. A. Avallone and T. Baumeister III, Marks' Standard Handbook for Mechanical Engineers, 10th ed., 1996; R. D. Reed, Furnace Operations, 3d ed., 1981; M. W. Thring, The Science of Flames and Furnaces, 1962.

Steam-generating unit

The wide diversity of parts, appurtenances, and functions needed to release and utilize a source of heat for the practical production of steam at pressures to 5000 lb/in.² (34 megapascals) and temperatures to 1100° F (600°C), often referred to as a steam boiler for brevity. *See* STEAM.

The essential steps of the steam-generating process include (1) a furnace for the combustion of fuel, or a nuclear reactor for the release of heat by fission, or a waste heat system; (2) a pressure vessel in which feedwater is raised to the boiling temperature, evaporated into steam, and generally superheated beyond the saturation temperature; and (3) in many modern



Fig. 1. Large, central-station steam-generating unit, with pump-assisted circulation, corner-fired with tangential, tilting burners for pulverized coal, oil, or gas. (a) Plan; (b) burners tilted down; (c) burners tilted up. Output is 2,000,000 lb steam/h (250 kg/s) at 2400 lb/in.² (16.5 MPa), 1000°F (540°C), and 1000°F (540°C) reheat. (*Combustion Engineering, Inc.*)

central station units, a reheat section or sections for resuperheating steam after it has been partially expanded in a turbine. This aggregation of functions requires a wide assortment of components, which may be variously employed in the interests, primarily, of capacity and efficiency in the steam-production process. The selection, design, operation, and maintenance of these components constitute a complex process which, in the limit, calls for the highest technical skill.

Service requirements. The wide diversity of commercial steam-generating equipment stems from attempts to accommodate design to the dictates of the imposed service conditions.

Safety. The pressure parts, a potential bursting hazard, must be designed, maintained, and operated under the most stringent codes, such as the ASME Boiler and Pressure Vessel Code. The codes and reg-



Fig. 2. Boiling-water reactor (BWR) for a large central-station electric generating plant showing essential components. Typical performance is 6,000,000 lb/h (750 kg/s) at 1000 lb/in.² (7 MPa), saturated steam. (*General Electric Co.*)

ulations are backed by the police power of the state.

Shape. Physical shape must fit the limits imposed. The differences between marine and stationary applications are an example.

Bulk. Space occupied has various degrees of value, and the equipment must best comply with these dictates.

Weight. Frequently, weight is the lowest common denominator in determining portability and economic suitability.

Setting. Confinement of heat source operations and the ensuing transport function will vary the details of the setting, for example, the use of metal or of refractory enclosure.

Character of labor. Highly skilled labor, as in atomic power plants, can utilize design details that would be prohibitive in areas where the labor is primitive.

Cleanability. Impurities on the heat-transfer and heat-release surfaces must not impair safe or efficient operation; surfaces must be cleanable to give acceptable reliability and performance.

Life. Short or long life and short or long operating periods between overhauls are vital to the selection of a preferred design.

Efficiency. Inherent high efficiency can be built into a design, but economics will dictate the efficiency chosen; the full realization of built-in high efficiency requires the services of skilled operators and firefighters.

Cost. The overall cost of steam in cents per 1000 lb or per 1000 kg is the ultimate criterion, reflecting investment, operating, and maintenance expenses in commercial operations.

Adaptation to use. These service requirements differ in each installation so that a wide variety of designs is available in the commercial markets (Figs. 1 and 2). There is no one best design for universal application, but propriety rests in the proper selection (see table). The designs are characterized by the various degrees of applicability of component parts such as steam drums and tubes; fuel-burning equipment and furnaces; draft systems for air supply and flue gas removal; blowers, fans, and stacks; heat traps such as economizers and air preheaters; structural steel for the support of parts with ample provision for expansion; a casing or setting with possible utilization of water walls, refractory, and insulation; pumps for boiler feed and for water circulation; fuel- and refuse-handling systems; feedwater purification systems; blowdown systems and soot-blowing equipment; and a wide assortment of accessories and instruments such as pressure gages, safety valves, water-level indicators, and sophisticated automatic control with its elaborate interlocks for foolproof, efficient operation and complete computerization.

The maintenance of a safe working temperature for metal parts requires ample circulation of water over the steam-generating parts and ample circulation of steam over the superheating and reheating parts. Water circulation in the generating sections may be by natural convection processes where ample physical height gives the circulatory pressure difference between a column of water in the

Performance characteristics of some steam-generating units					
Boiler type	Fuel (potential)	Steam output, Ib/h (kg/s)	Steam pressure, Ib/in. ² (MPa)	Steam temperature, °F (°C)	Boiler efficiency, %
Horizontal return tubular (HRT), fire-tube	Coal, oil, gas	10,000 (1.3)	150 (1.0)	Dry and saturated	70
Large industrial, natural-circulation, water-tube	Coal, oil, gas	300,000 (38)	1400 (9.7)	900 (480)	88
Large central station, natural-circulation	Coal, oil, gas	2,000,000 (250)	2700 (18.6)	1050/1050 (570/570)	90
Large central station, supercritical, once-through	Coal, oil, gas	5,000,000 (630)	3600 (24.8)	1000/1000/1000 (540/540/540)	90
Boiling-water reactor (BWR)	Uranium	6,000,000 (750)	1000 (6.9)	Dry and saturated (550 ± °F or 290 ± °C)	98±
Pressurized-water reactor (PWR), hot water, 4 loops	Uranium	6,000,000 (750)	700 (4.8)	Dry and saturated (500 [±] °F or 260 [±] °C)	98
Electric boiler, small	200± volts	2,000 (0.25)	150 (1.0)	Dry and saturated	98±

downcomer and a column of steam mixed with water in the riser parts. With high operating steam pressures, such as 1400 lb/in.2 (9.7 MPa), pumpassisted circulation is often selected. With supercritical operation (above 3200 lb/in.² or 22.1 MPa) there is no difference in density between liquid and vapor so that forced-flow, once-through principles supersede water recirculation practice. Such supercritical steam generators have no steam or separating drums and, like nuclear steam generators, require the highest purity of feed and boiler water (typically 10⁻⁹ part of impurities) to avoid deposits in heated circuits or the transport of solids to the turbine. See BOILER; BOILER ECONOMIZER; FIRE-TUBE BOILER; MA-RINE BOILER; RAW WATER; REHEATING; STEAM SEPARA-TOR; STEAM TEMPERATURE CONTROL; SUPERHEATER; WATER-TUBE BOILER. Theodore Baumeister

Bibliography. American Society of Mechanical Engineers, *Boiler And Pressure Vessel Code 2004*, 2003; E. A. Avallone, T. Baumeister III, and A. Sadegh (eds.), *Mark's Standard Handbook for Mechanical Engineers*, 11th ed., 2006; J. J. Jackson, *Steam Boiler Operation; Principles and Practices*, 2d ed., 1987; E. B. Woodruff, H. B. Lammers, and T. F. Lammers, *Steam Plant Operation*, 8th ed., 2004.

Steam heating

A heating system that uses steam generated from a boiler. The steam heating system conveys steam through pipes to heat exchangers, such as radiators, convectors, baseboard units, radiant panels, or fandriven heaters, and returns the resulting condensed water to the boiler. Such systems normally operate at pressure not exceeding 15 lb/in.² gage or 103 kilopascals gage, and in many designs the condensed steam returns to the boiler by gravity because of the static head of water in the return piping. With utilization of available operating and safety control devices, these systems can be designed to operate automatically with minimal maintenance and attention.

One-pipe system. In a one-pipe steam heating system, a single main serves the dual purpose of supplying steam to the heat exchanger and conveying

condensate from it. Ordinarily, there is but one connection to the radiator or heat exchanger, and this connection serves as both the supply and return; separate supply and return connections are sometimes used. Because steam cannot flow through the piping or into the heat exchanger until all the air is expelled, automatic air-venting valves must be provided on all exchangers and at the ends of all mains. These valves may be of a type which closes whenever steam or water comes in contact with the operating element but which also permits air to flow back into the system as the pressure drops. A vacuum valve closes against subatmospheric pressure to prevent return of air.

Two-pipe system. A two-pipe system is provided with two connections from each heat exchanger, and in this system steam and condensate flow in separate mains and branches (**Fig. 1**). A vapor two-pipe system operates at a few ounces above atmospheric pressure, and in this system a thermostatic trap is located at the discharge connection from the heat exchanger which prevents steam passage, but permits air and condensation to flow into the return piping.

When the steam condensate cannot be returned by gravity to the boiler in a two-pipe system, an alternating return lifting trap, condensate return pump, or vacuum return pump must be used to force the condensate back into the boiler. In a condensate return-pump arrangement, the return piping is arranged for the water to flow by gravity into a collecting receiver or tank, which may be located below the steam-boiler waterline. A motor-driven pump controlled from the boiler water level then forces the condensate back to the boiler.

In large buildings extending over a considerable area, it is difficult to locate all heat exchangers above the boiler water level or return piping. For these systems a vacuum pump is used that maintains a suction below atmosphere up to ± 25 in. or ± 64 cm (max) of mercury in the return piping, thus creating a positive return flow of air and condensate back to the pumping unit. Subatmospheric systems are similar to vacuum systems, but in contrast provide a means of partial vacuum control on both the supply and



Fig. 1. Two-pipe up-feed system with automatic return trap.

return piping so that the steam temperature can be regulated to vary the heat emission from the heat exchanger in direct proportion to the heat loss from the structure. tem which uses a condensation pump as a mechanical lift for systems where a part of the heating system is below the boiler room. Note that the low section of the system is maintained under the same vacuum conditions as the rest of the system. This is



Figure 2 depicts a two-pipe vacuum heating sys-

Fig. 2. Layout of a vacuum heating system with condensation and vacuum pumps.

accomplished by connecting the vent from the receiver and pump discharge to the return pipe located above the vacuum heating pump.

With the wide acceptance of all-year air conditioning, low-pressure steam boilers have been used to produce cooling from absorption refrigeration equipment. With this system the boiler may be used for primary or supplementary steam heating.

Exhaust from gas- or oil-driven turbines or engines may be used in waste heat boilers or separators, along with a standby boiler to produce steam for a heating system.

Another source for steam for heating is from a high-temperature water source (350–450°F or 180–230°C) using a high-pressure water to lowpressure steam heat exchanger. *See* BOILER; COM-FORT HEATING; OIL BURNER; STEAM-GENERATING FUR-NACE. John W. James

Bibliography. M. Greenwald, *Residential Hot Water and Steam Heating with Gas, Oil and Solid Fuels,* 1987; E. B. Woodruff, H. B. Lammers, and T. F. Lammers, *Steam-Plant Operation,* 8th ed., 2004.

Steam jet ejector

A steam-actuated device for pumping compressible fluids, usually from subatmospheric suction pressure to atmospheric discharge pressure. A steam jet ejector is most frequently used for maintaining vacuum in process equipment in which evaporation or condensation takes place. Because of its simplicity, compactness, reliability, and generally low first cost, it is often preferred to a mechanical vacuum pump for removing air from condensers serving steam turbines, especially for marine service. *See* VACUUM PUMP.

Principle. Compression of the pumped fluid is accomplished in one or more stages, depending upon the total compression required. Each stage consists of a converging-diverging steam nozzle, a suction chamber, and a venturi-shaped diffuser. Steam is supplied to the nozzle at pressures in the range 100-250 lb/in.² gage (0.7-1.7 megapascals). A portion of the enthalpy of the steam is converted to kinetic energy by expanding it through the nozzle at substantially constant entropy to ejector suction pressure, where it reaches velocities of 3000-4500 ft/s (900-1400 m/s). The air or gas, with its vapor of saturation, which is to be pumped and compressed is entrained, primarily by friction in the high-velocity steam jet. The impulse of the steam produces a change in the momentum of the air or gas vapor mixture as it mixes with the motive steam and travels into the converging section of the diffuser. In the throat or most restricted area of the diffuser, the energy transfer is completed and the final mixture of gases and vapor enters the diverging section of the diffuser, where its velocity is progressively reduced. Here a portion of the kinetic energy of the mixture is reconverted to pressure with a corresponding increase in enthalpy. Thus the air or gas is compressed to a higher pressure than its entrance pressure to the ejector. The compression ratios selected for each stage of a steam jet



Fig. 1. Typical multistage steam jet ejector with contact barometric condensers; first and second stages condensing and third stage noncondensing.

ejector usually vary from about 4 to 7. See DIFFUSER.

Application. Two or more stages may be arranged in series, depending upon the total compression ratio required (**Fig. 1**). Two or more sets of series stages may be arranged in parallel to accommodate variations in capacity.

Vapor condensers are usually interposed between the compression stages of multistage steam jet ejectors to condense and remove a significant portion of the motive steam and other condensable vapors (**Fig. 2**). This action reduces the amount of fluid to be compressed by the next higher stage and results in a reduction in the motive steam required. Both surface and contact types of vapor condensers are used for this purpose. *See* VAPOR CONDENSER.

Ejectors used as air pumps for steam condensers that serve turbines are usually two-stage and are equipped with inter- and aftercondensers of the surface type. The steam condensed is drained through traps to the main condenser and returned to the boiler feed system. Ejectors used as vacuum pumps in process systems may be equipped with either surface condensers or contact condensers of the barometric type between or after stages or both. They



Fig. 2. Multistage steam jet ejector with surface inter- and aftercondenser; first and second stages condensing.

may be single or multistage machines. High-vacuum process ejectors with as many as seven stages in series have been built. Industrial or process ejectors are frequently used instead of mechanical vacuum pumps to pump corrosive vapors because they can be manufactured economically from almost any corrosion-resistant material.

The number of stages of compression usually used for various suction pressures with atmospheric discharge pressure is as follows:

No. of	Range of suction pressure
ejector stages	
1	3-30 in. Hg (10-102 kPa) abs.
2	0.4-4 in. Hg (1.4-14 kPa) abs.
3	1-25 mmHg (0.13-3.3 kPa) abs
4	0.15-3 mmHg (20-400 Pa) abs.
5	20-300 µm (2.7-40 Pa)
6	5-20 µm (0.7-2.7 Pa)
7	1-5 μm (0.13-0.7 Pa)

Ejectors are also made which use air or other gases instead of steam as the energy source.

Joseph F. Sebald

Steam separator

e

A device for separating a mixture of the liquid and vapor phases of water. Steam separators are used in most boilers and may also be used in saturated steam lines to separate and remove the moisture formed because of heat loss.

In boilers, steam separators must perform efficiently because both steam-free water and waterfree steam are required. The force promoting water circulation in a natural-circulation boiler is derived from the density difference between the water in the downcomer tubes and the steam-water mixture in the riser tubes. Solid (steam-free) water in the downcomers maximizes this force. Steam-free water is equally important in a forced-circulation boiler to prevent cavitation in the pump.

Water-free steam is quite necessary because dissolved and suspended solids entering the boiler with the feedwater concentrate in the boiler water, and such solids contaminate the steam if all the water is not separated. Impurity concentrations as low as a few parts per billion can cause solid deposits to form in turbines or can cause product contamination if the steam is used in direct contact with a product. Solid deposits in turbines decrease turbine capacity and efficiency and may even cause unbalance of the rotor in extreme cases. Consequently, the sampling of steam and the measurement of impurities in steam have received much attention.

Many low-pressure boilers operate satisfactorily without steam separators because their capacity ratings are quite conservative. Steam separation in such boilers may be compared with a pan of water boiling on a stove at a moderate rate. However, it is known that if the fire under the pan is turned too high or



Steam dryers and cyclone steam separators separate steam and water in high-capacity boiler unit. Arrows show direction of flow.

if the water contains certain solids (like milk, for example), separation becomes inadequate and the pan boils over. Similar phenomena take place in natural separation boilers if the load is too high or the concentration of solids in the water is too great.

Operating pressure also influences natural separation. The force effecting separation results from the difference in density between water and steam. This difference is quite great at low pressures, but as pressure is increased, the density of water decreases while that of steam increases, resulting in a steadily declining density difference.

Because of these effects, steam separators are used on the majority of boilers. Steam separators have many forms and may be as fundamental as a simple baffle that utilizes inertia caused by a change of direction. Most modern, high-capacity boilers use a combination of cyclone separators and steam dryers, as shown in the **illustration**.

Cyclone separators have two major advantages. First, they slice the steam-water mixture into thin streams so that the steam bubbles need travel only short distances through the mixture to become disengaged. Second, they whirl the mixture in a circular path, creating a centrifugal force many times greater than the force of gravity.

Cyclone separators remove practically all of the steam from the water, producing the steam-free water needed for the boiler downcomer tubes. They also remove the major portion of the water from the steam. After leaving the cyclone separators, the steam is passed through the dryers, where the last traces of moisture are removed. Steam dryers remove small droplets of water from the steam by providing a series of changes in direction and a large surface area to intercept the droplets.

Because of the efficient operation of steam separators, steam is one of the purest mass-produced commodities made. *See* BOILER; BOILER FEEDWATER; STEAM; STEAM TURBINE. Earl E. Coulter

Bibliography. American Society for Testing and Materials, *Methods of Sampling Steam*, ASTM

Designation D1066; ASTM, Methods of Test for Deposit-Forming Impurities in Steam, ASTM Designation D2186; ASTM, ASTM Standards, pt. 23: Water:Atmospheric Analysis, annually; P. M. Brister, F. G. Raynor, and E. A. Pirsh, The Influence of Boiler Design and Operating Conditions on Steam Contamination, ASME Pap. 51-A-95; M. J. Moore and C. H. Sieverding, Two-Phase Steam Flow in Turbines and Separators: Theory, Instrumentation, Engineering, 1976.

Steam temperature control

Means for regulating the operation of a steamgenerating unit to produce steam at the required temperature. The temperature of steam is affected by the change in the relative heat absorption as load varies, by changes in ash or slag deposits on the heatabsorbing surface, by changes in fuel, by changes in the proportioning of fuel and combustion air, or by changes in feedwater temperature. Low steam temperature lowers the efficiency of the thermal cycle. However, high steam temperature, which increases thermal efficiency, is restricted by the strength and durability of materials used in superheaters. Control of steam temperature is, therefore, a matter of primary concern in the design of modern steamgenerating units.

Steam temperature can be controlled by one or more of several methods (see **illus.**). These include



Methods of controlling steam temperature. (*a*) Bypass dampers. (*b*) Tilting burners. (*c*) Spray attemperation. (*d*) Gas recirculation.

(1) the damper control of gases to the superheater, to the reheater, or to both, thus changing the heat input; (2) the recirculation of low-temperature flue gas to the furnace, thus changing the relative amounts of heat absorbed in the furnace and in the superheater, reheater, or both; (3) the selective use of burners at different elevations in the furnace or the use of tilting burners, thus changing the location of the combustion zone with respect to the furnace heat-absorbing surface; (4) the attemperation, or controlled cooling, of the steam by the injection of spray water or by the passage of a portion of the steam through a heat exchanger submerged in the boiler water; (5) the control of the firing rate in divided furnaces; and (6) the control of the firing rate relative to the pumping rate of the feedwater to forced-flow oncethrough boilers. Generally, these various controls are adjusted automatically. See STEAM-GENERATING UNIT. George W. Kessler

Steam turbine

A machine for generating mechanical power in rotary motion from the energy of steam at temperature and pressure above that of an available sink. By far the most widely used and most powerful turbines are those driven by steam. In the United States well over 85% of the electrical energy consumed is produced by steam-turbine-driven generators. Individual turbine ratings historically have tended to follow the increasing capacity trend but have reached limits imposed by material and machine design considerations. The largest unit shipped during the 1950s was rated 500 MW. Units rated about 1100 MW were in service by the close of the 1960s, and ratings up to 1300 MW saw frequent application in the 1970s and early 1980s. Units of all sizes, from a few horsepower to the largest, have their applications.

Until the 1960s essentially all steam used in turbine cycles was produced in boilers burning fossil fuels (coal, oil, and gas) or, in minor quantities, certain waste products. The 1960s marked the beginning of the introduction of commercial nuclear power. *See* ELECTRIC POWER GENERATION; NUCLEAR POWER.

Turbine parts. Figure 1 shows a small, simple mechanical-drive turbine of a few horsepower. It illustrates the essential parts for all steam turbines regardless of rating or complexity: (1) a casing, or shell, usually divided at the horizontal center line, with the halves bolted together for ease of assembly and disassembly; it contains the stationary blade system; (2) a rotor carrying the moving buckets (blades or vanes) either on wheels or drums, with bearing journals on the ends of the rotor; (3) a set of bearings attached to the casing to support the shaft; (4) a governor and valve system for regulating the speed and power of the turbine by controlling the steam flow, and an oil system for lubrication of the bearings and, on all but the smallest machines, for operating the control valves by a relay system connected with the governor; (5) a coupling to connect with the driven machine; and (6) pipe connections to the steam supply



Fig. 1. Cutaway of small single-stage steam turbine. (General Electric Co.)

at the inlet and to an exhaust system at the outlet of the casing or shell.

Applications. Steam turbines are ideal prime movers for applications requiring rotational mechanical input power. They can deliver constant or variable speed and are capable of close speed control. Drive applications include centrifugal pumps, compressors, ship propellers, and, most importantly, electric generators.

The turbine shown in Fig. 1 is a small mechanicaldrive unit. Units of this general type provide 10– 1000 hp (7.5–750 kW) with steam at 100–600 lb/in.² gage (0.7–4.1 megapascals) inlet pressure and temperatures to 800°F (427° C). These and larger multistage machines drive small electric generators, pumps, blowers, air and gas compressors, and paper machines. A useful feature is that the turbine can be equipped with an adjustable-speed governor and thus be made capable of producing power over a wide range of rotational speeds. In such applications efficiency varies with speed (**Fig. 2**), being 0 when the rotor stalls at maximum torque and also 0 at the runaway speed at which the output torque is 0. Maximum efficiency and power occur where the product



Fig. 2. Illustrative stage performance versus speed.

of speed and torque is the maximum.

Many industries need steam at one or more pressures (and consequently temperatures) for heating and process purposes. Frequently it is more economical to produce steam at high pressure, expand it partially through a turbine, and then extract it for further processing, rather than using a separate boiler at the process steam pressure. Figure 3 is a cross section through an industrial automatic extraction turbine. The left set of valves admits steam from the boiler at the flow rate to provide the desired electrical load. The steam flows through five stages to the controlled extraction point. The second set of valves acts to maintain the desired extraction pressure by varying the flow through the remaining 12 stages. Opening these internal valves increases the flow to the condenser and lowers the controlled extraction pressure.

Industrial turbines are custom-built in a wide variety of ratings for steam pressures to 2000 lb/in.² gage (14 MPa), for temperatures to 1000°F (538°C), and in various combinations of nonextracting, single and double automatic extraction, noncondensing and condensing. Turbines exhausting at or above atmospheric pressure are classed as noncondensing regardless of what is done with the steam after it leaves the turbine. If the pressure at the exhaust flange is less than atmospheric, the turbine is classed as condensing.

Turbines in sizes to about 75,000 hp (56 MW) are used for ship propulsion. The drive is always through reduction gearing (either mechanical or electrical) because the turbine speed is in the range of 4000– 10,000 rpm, while 50–200 rpm is desirable for ship propellers. Modern propulsion plants are designed for steam conditions to 1450 lb/in.² gage (10.0 MPa) and 950°F (510°C) with resuperheating to 950°F (510°C). Fuel consumption rates as low as 0.4 lb of oil per shaft-horsepower-hour (0.07 kg/megajoule) are achieved. *See* SHIP POWERING, MANEUVERING, AND SEAKEEPING.

Central station generation of electric power provides the largest and most important single application of steam turbines. Ratings smaller than 50 MW are seldom employed today; newer units are rated as large as 1300 MW. Large turbines for electric power production are designed for the efficient use of steam in a heat cycle that involves extraction of steam for feedwater heating, resuperheating of the main steam flow (in fossil-fuel cycles), and exhausting at the lowest possible pressure economically consistent with the temperature of the available condenser cooling water.

In fossil-fuel-fired cycles, steam pressures are usually in the range of 1800-3500 lb/in.² gage (12.4-24.1 MPa) and tend to increase with rating. Temperatures of 950-1050°F (510-570°C) are used, with 1000°F (540°C) the most common. Single resuperheat of the steam to 950-1050°F (510-570°C) is almost universal. A second resuperheating is occasionally employed. **Figure 4** shows a typical unit designed for fossil-fuel steam conditions. Tandemcompound double-flow machines of this general



Fig. 3. Cross-section view of single-automatic-extraction condensing steam turbine. (General Electric Co.)

arrangement are applied over the rating range of 100–400 MW. Initial steam flows through the steam admission valves and passes to the left through the high-pressure portion of the opposed flow rotor. After resuperheating in the boiler it is readmitted through the intercept valves and flows to the right through the reheat stages, then crosses over to the double-flow low-pressure rotor, and exhausts downward to the condenser. *See* SUPERHEATER.

The water-cooled nuclear reactor systems common in the United States provide steam at pressures of about 1000 lb/in.² gage. (6.9 MPa), with little or no initial superheat. Temperatures higher than about 600°F (315° C) are not available. Further, reactor containment and heat-exchanger considerations preclude the practical use of resuperheating at the reactor. The boiling-water reactor, for example, provides steam to the turbine cycle at 950-lb/in.² gage (6.55-MPa) pressure and the saturation temperature of 540°F (282°C). Such low steam conditions mean that each unit of steam flow through the turbine produces less power than in a fossil-fuel cycle.



Fig. 4. Partial cutaway view of 3600-rpm fossil-fuel turbine generator. (General Electric Co.)

Fewer stages in series are needed but more total flow must be accommodated for a given output. Nuclear steam conditions often produce a turbine expansion with water of condensation present throughout the entire stream path. Provisions must be made to control the adverse effects of water: erosion, corrosion, and efficiency loss. In consequence of the differences in stream conditions, the design for a nuclear turbine differs considerably from that of a turbine for fossil fuel application. The former tend to be larger, heavier, and more costly. For example, at 800 MW, a typical fossil-fuel turbine can be built to run at 3600 rpm, is about 90 ft (27 m) long, and weighs about 1000 tons (900 metric tons). A comparable unit for a water-cooled nuclear reactor requires 1800 rpm, is about 125 ft (38 m) long, and weighs about 2500 tons (2300 metric tons). See NUCLEAR REACTOR.

Figure 5 represents a large nuclear turbine generator suitable for ratings of 1000 to 1300 MW. Steam from the reactor is admitted to a double-flow highpressure section, at the left, through four parallel pairs of stop and control valves, not shown. The stop valves are normally fully open and are tripped shut to prevent dangerous overspeed of the unit in the event of loss of electrical load combined with control malfunction. The control valves regulate output by varying the steam flow rate. The steam exhausting from the high-pressure section is at 150 to 200 lb/in.² abs. (1.0 to 1.4 MPa absolute pressure) and contains about 13% liquid water by weight. The horizontal cylinder alongside the foundation is one of a pair of symmetrically disposed vessels performing three functions. A moisture separator removes most of the water in the entering steam. Two steam-to-steam reheaters follow. Each is a U-tube bundle which condenses heating steam within the tubes and superheats the main steam flow on the shell side. The first stage uses heating steam extracted from the high-pressure turbine. The final stage employs reactor steam which permits reheating to near initial temperature. Alternate cycles employ reheat with initial steam only or moisture separation alone with no reheat. Reheat enhances cycle efficiency at the expense of increased investment and complexity. The final choice is economic, with two-stage steam reheat, as shown, selected most frequently.

Reheated steam is admitted to the three doubleflow low-pressure turbine sections through six combined stop-and-intercept intermediate valves. The intermediate valves are normally wide open but provide two lines of overspeed defense in the event of load loss. Exhaust steam from the low-pressure sections passes downward to the condenser, not shown in Fig. 5.

Turbine cycles and performance. Figures 6 and 7 are representative fossil-fuel and nuclear turbine thermodynamic cycle diagrams, frequently called heat balances. Heat balance calculations establish turbine performance guarantees, provide data for sizing the steam supply and other cycle components, and are the basis for designing the turbine generator.

The fossil-fuel cycle (Fig. 6) assumes a unit rated 500 MW, employing the standard steam condi-



Fig. 5. An 1800-rpm nuclear turbine generator with combined moisture separator and two-stage steam reheater. (General Electric Co.)

tions of 2400 lb/in.2 gage (2415 lb/in.2 abs. or 16.65 MPa absolute) and 1000°F (538°C), with resuperheat to 1000°F (538°C). As can be seen in the upper left corner, the inlet conditions correspond to a total heat content, or enthalpy, of 1461 Btu/lb (3.398 MJ/kg) of steam flow. A flow rate of 3,390,000 lb/h (427.1 kg/s) is needed for the desired output of 500 MW (500,000 kW). For efficiency considerations the regenerative feedwater heating cycle is used. Eight heaters in series are employed so that water is returned to the boiler at 475°F (246°C) and 459 Btu/lb (1.068 MJ/kg) enthalpy, rather than at the condenser temperature of 121°F (49°C). Because of the higher feedwater temperature, the boiler adds heat to the cycle at a higher average temperature, more closely approaching the ideal Carnot cycle, in which all heat is added at the highest cycle temperature. The high-pressure turbine section exhausts to the resuperheater at 530 psia (3.65 MPa absolute) pressure and 1306 Btu/lb (3.038 MJ/kg) enthalpy. The reheat flow of 3,031,000 lb/h (381.9 kg/s) returns to the reheat or intermediate turbine section at 490 lb/in.² abs. (3.38 MPa absolute) pressure and 1520 Btu/lb (3.536 MJ/kg) enthalpy. These data are sufficient to calculate the turbine heat rate, or unit heat charged against the turbine cycle. The units are Btu of heat added in the boiler per hour per kilowatt of generator output. Considering both the initial and reheat steam, the heat rate is given by Eq. (1).

Turbine heat rate =

 $\frac{3,390,000\ (1461-459)+3,031,000\ (1520-1306)}{500,000}$

= 8090 Btu/kWh

= 2.37 kW of heat/kW of output(1)

The typical power plant net heat rate is poorer



Fig. 6. Typical fossil-fuel steam turbine cycle.

than the turbine heat rate because of auxiliary power required throughout the plant and because of boiler losses. Assuming 3% auxiliary power (beyond the boiler-feed pump power given in the turbine cycle in Fig. 6) and 90% boiler efficiency, the net plant heat rate is given by Eq. (2).

Net plant heat rate

$$= 3,390,000 (1461 - 459) +3,031,000 (1520 - 1306) $\overline{500,000 [(100 - 3)/100][90/100]}$
= 9270 Btu/kWh
= 2.717 kW of heat/kW of output (2)$$

The heat rates of modern fossil-fuel plants fall in the range of 8600-10,000 Btu/kWh (2.52-2.93 kW of heat/kW of input). Considering that the heat-energy equivalent of 1 kWh is 3412 Btu, the thermal efficiency of the example is given by Eq. (3).

$$\eta_t = (3412/9270) \ 100 = 37\% \tag{3}$$

Figure 6 shows 2,191,000 lb/h (276.1 kg/s) of steam exhausting from the main unit to the condenser at an exhaust pressure of 3.5 in. of mercury (11.9 kilopascals) absolute. The theoretical exhaust enthalpy (ELEP), without considerations of velocity energy loss and friction loss between the last turbine stage and the condenser, is 1040 Btu/lb (2.419 MJ/kg). The actual used energy end point (UEEP) is 1050 Btu/lb (2.442 MJ/kg). The exhaust heat at the condenser pressure is thermodynamically unavailable and is rejected as waste heat to the plant's surroundings. The exhaust steam is condensed at a constant $121^{\circ}F$ (49°C) and leaves the condenser as water at 89 Btu/lb (0.207 MJ/kg) enthalpy.

On a heat-rate basis, this cycle rejects heat to the



Fig. 7. Typical nuclear steam turbine cycle.

condenser at the approximate rate which is given by Eq. (4).

Net station condenser heat rejection rate

$$= 2,191,000 (1050 - 89) + 144,000 (1097 - 89) + 417,000 (100 - 89) 500,000 (0.97)$$

= 4650 Btu/kWh

$$= 1.363$$
 kW of heat/kW of output

(4)

If evaporating cooling towers are used, each pound of water provides about 1040 Btu (each kilogram provides about 2.419 MJ) cooling capacity, which is equivalent to a required minimum cooling-water flow rate of 4.5 lb/kWh (0.57 kg/MJ). The cooling-water needs of a large thermal plant are a most important consideration in plant site selection.

The nuclear cycle (Fig. 7) assumes a unit rated 1210 MW and the steam conditions of the boilingwater reactor. Many similarities can be seen to Fig. 6. The major differences include moisture separation and steam reheating and the lack of need for an intermediate pressure element. The low steam conditions are apparent. The consequent turbine heat rate is given in Eq. (5).

Turbine heat rate

$$= \frac{15,400,000 (1191 - 398)}{1,210,000}$$

$$= 10,090 \text{ Btu/kWh}$$

$$= 2.957 \text{ kW of heat/kW of output} (5)$$

A typical nuclear plant also requires about 3% auxiliary power beyond the reactor feed-pump power already included in Fig. 7. The equivalent boiler efficiency approaches 100%, however, and leads to the equivalent net plant heat rate given by Eq. (6).

Net plant heat rate

$$= \frac{15,400,000 (1191 - 398)}{1,210,000 [(100 - 3)/100]}$$

$$= 10,400 \text{ Btu/kWh}$$

$$= 3.048 \text{ kW of heat/kW of output} \qquad (6)$$

The corresponding thermal efficiency is given by Eq. (7).

$$\eta_t = (3412/10, 400) \ 100 = 33\% \tag{7}$$

Heat is rejected at the condenser at a rate given approximately by Eq. (8).

Net station condenser heat rejection rate

$$= [8,350,000 (1020 - 89) + 200,000 (1028 - 89) + 2,350,000 (101 - 89)] + 2,350,000 (101 - 89)] - 1,210,000 (0.97)$$

= 6810 Btu/kWh

= 1.996kW of heat/kW of output (8)

Comparison of the heat rates shows that the nuclear cycle requires about 12% more input heat than does the fossil-fuel cycle and rejects about 46% more heat to the condenser, thus requiring a correspondingly larger supply of cooling water. It consumes heat priced at about half that from coal or one quarter that from oil, and is essentially free of rejection to the atmosphere of heat and combustion products from the steam supply.

Turbine classification. Steam turbines are classified (1) by mechanical arrangement, as single-casing, cross-compound (more than one shaft side by side), or tandem-compound (more than one casing with a single shaft); (2) by steam flow direction (axial for most, but radial for a few); (3) by steam cycle, whether condensing, noncondensing, automatic extraction, reheat, fossil fuel, or nuclear; and (4) by number of exhaust flows of a condensing unit, as single, double, triple flow, and so on. Units with as many as eight exhaust flows are in use.

Often a machine will be described by a combination of several of these terms.

The least demanding applications are satisfied by the simple single-stage turbine of Fig. 1. For large power output and for the high inlet pressures and temperatures and low exhaust pressures which are required for high thermal efficiency, a single stage is not adequate. Steam under such conditions has high available energy, and for its efficient utilization the turbine must have many stages in series, where each takes its share of the total energy and contributes its share of the total output. Also, under these conditions the exhaust volume flow becomes large, and it is necessary to have more than one exhaust stage to avoid a high velocity upon leaving and consequent high kinetic energy loss. Figure 5 is an example of a large nuclear turbine generator which has six exhaust stages in parallel.

Machine considerations. Steam turbines are highspeed machines whose rotating parts must be designed for high centrifugal stress. Difficult stress problems are found in long last-stage blading, hot inlet blading, wheels, and rotor bodies.

Casing or shell stresses. The casings or shells at the high-pressure inlet end must be high-strength pressure vessels to contain the internal steam pressure. The design is made more difficult by the need for a casing split at the horizontal center line for assembly. The horizontal flange and bolt design must be leakproof. Shell design problems lead to the use of

small-diameter, high-speed turbines at high pressure, and the use of double shell construction (Fig. 4).

Rotor buckets or blades. Turbine buckets must be strong enough to withstand high centrifugal, steam bending, and vibration forces. Buckets must be designed so that their resonant natural frequencies avoid the vibration stimulus frequencies of the steam forces, or are strong enough to withstand the vibrations.

Sealing against leakage. It is necessary to minimize to the greatest possible extent the wasteful leakage of steam along the shaft both at the ends and between stages. The high peripheral velocities between the shaft and stationary members preclude the use of direct-contact seals. Seals in the form of labyrinths with thin, sharp teeth on at least one of the members are utilized. In normal operation these seals do not touch one another, but run at close clearance. In the case of accidental contact, the sharp teeth can rub away without distorting the shaft.

Vibration and alignment. Shaft and bearings should be free of critical speeds in the turbine operating range. The shaft must be stable and remain in balance.

Governing. Turbines usually have two governors, one to control speed and a second, emergency governor to limit possible destructive overspeed. The speed signal is usually mechanical or electrical. A power relay control, usually hydraulic, converts speed signals to steam valve position. Great reliability is required. *See* GOVERNOR.

Lubrication. The turbine shaft runs at high surface speed; consequently its bearings must be continuously supplied with oil. At least two oil pumps, a main pump and a standby driven by a separate power source, are usually provided on all but the smallest machines. A common supply of oil is often shared between the governing hydraulic system and the lubrication system.

Aerodynamic design. The vane design for highest efficiency, especially for the larger sizes of turbines, draws upon modern aerodynamic theory. Classic forms of impulse and reaction buckets merge in the three-dimensional design that is required by modern concepts of loss-free fluid flow. To meet the theoretical steam flow requirements, vane sections change in shape along the bucket. To minimize centrifugal forces on the vanes and their attachments, long turbine buckets are tapered toward their tips. *See* CARNOT CYCLE; HEAT BALANCE; STEAM-GENERATING UNIT; TURBINE. Frederick G. Baily

Bibliography. E. A. Avallone and T. Baumeister III (eds.), *Marks' Standard Handbook for Mechanical Engineers*, 10th ed., 1996; F. G. Baily, K. C. Cotton, and R. C. Spencer, Predicting the performance of large steam turbine-generators operating with saturated and low superheat steam conditions, *Combustion*, 3(3):8–13, 1967; H. P. Bloch, *A Practical Guide to Steam Turbine Technology*, 1996; R. H. Kehlhofer et al., *Combined-Cycle Gas and Steam Turbine Power Plants*, 2d ed., 1999; J. K. Salisbury, *Steam Turbines and Their Cycles*, 1950, reprint 1974; B. G. A. Strotzki, Steam turbines, *Power*, 106(6):S1–S40, 1962.

Steel

Any of a great number of alloys that contain the element iron as the major component and small amounts of carbon as the major alloying element. These alloys are more properly referred to as carbon steels. Small amounts, generally on the order of a few percent, of other elements such as manganese, silicon, chromium, molybdenum, and nickel may also be present in carbon steels. However, when large amounts of alloying elements are added to iron to achieve special properties, other designations are used to describe the alloys. For example, large amounts of chromium, over 12%, are added to produce the important groups of alloys known as stainless steels.

Phases. There are three unique thermodynamically stable arrangements of iron and carbon atoms in carbon steels. The arrangements are crystalline and referred to as phases. Ferrite is the phase with a body-centered cubic unit cell, austenite is the phase with a face-centered cubic unit cell, and cementite is the phase with an orthorhombic unit cell. The phase or combination of phases that actually exists in a given steel depends on temperature and composition. A phase diagram is a type of map that shows where the different phases exist under equilibrium conditions. Equilibrium assumes that a steel has been held long enough at a temperature to ensure that the most stable arrangement of atoms and phases has developed (**Fig. 1**).

The phase diagram for the Fe-C system shows that austenite is stable only at high temperatures and that it is replaced at low temperatures by other phases. This characteristic of iron-carbon alloys is the fundamental basis for the processing and great versatility of steels. When a steel is heated in the austenite phase field, it becomes quite ductile, and heavy sections can be hot-rolled or forged to smaller sizes and more complex shapes. On cooling after hot work, the austenite transforms to structures of higher



Fig. 1. Critical temperatures in plain carbon steels and constituents which are present in iron-carbon alloys upon slow cooling.

strength. Another important use of the austenite phase field is the heat treatment of steels. By heating a finished steel part into the austenite phase, that is, by heating above the critical temperatures identified by A_3 and A_{cm} on the phase diagram (Fig. 1), the microstructure existing at room temperature is replaced by austenite. The resulting austenite may then be cooled at various rates to form new arrangements of the phases with a variety of mechanical properties. Cooling at very high rates or quenching produces the hardest structure in steels, and is discussed in the sections on medium- and high-carbon steels. *See* ALLOY STRUCTURES.

Microstructure and properties. When steels are cooled slowly from the austenite phase field, the austenite transforms to microstructures of ferrite and cementite. In steels with carbon content below 0.8%, the austenite first transforms to ferrite, as indicated by the ferrite-plus-austenite field of the phase diagram for the Fe-C system (Fig. 1). Steels with carbon content above 0.8% first form cementite. At or below the A1 temperature, any austenite not transformed to ferrite or cementite transforms to a mixture of ferrite and cementite referred to as pearlite. The ferrite and cementite form as parallel plates, or lamellae, in pearlite. Thus, when a steel reaches room temperature, its microstructure consists of either ferrite and pearlite or cementite and pearlite, depending on its carbon content (Fig. 1).

Typical microstructures exist for steels cooled very slowly in a furnace, a process referred to as full annealing, or cooled in air, a process referred to as normalizing. The 0.06% carbon steel consists almost entirely of ferrite in the form of fine crystals or grains. The higher-carbon steels contain increasingly greater amounts of pearlite which, at the low magnifications (**Fig. 2**), appears dark because the lamellae are too close together to be resolvable.

The mechanical properties of annealed, normalized, or hot-rolled steels are directly dependent on the amounts of pearlite present in the microstructure. Cementite is a hard carbide phase and, when present in the pearlite, significantly strengthens a steel. The yield and tensile strengths increase with increasing carbon content or increasing pearlite content of steels containing up to 0.8% carbon. However, as strength increases, ductility falls and the steels become less formable, as shown by the decrease in reduction of area and elongation (**Fig. 3**).

The change in mechanical properties with carbon content separates steels into low-, medium-, and high-carbon classifications that are related to the need for formability in manufacturing and strength and toughness for good performance in service.

Low-carbon steels. These steels, sometimes referred to as mild steels, usually contain less than 0.25% carbon, and therefore have microstructures that consist mostly of ferrite and small amounts of pearlite. These steels are easily hot-worked in the austenite condition and are produced in large tonnages for beams and other structural applications. The relatively low strength and high ductility (Fig. 3)



Fig. 2. Micrographs illustrating appearance of the constituents shown in Fig. 1. (a) 0.06% carbon steel: white is ferrite, grain boundaries appear as a dark network. (b) 0.18% carbon steel: white is ferrite, dark is pearlite. (c) 0.47% carbon steel: same as b, but larger amount of pearlite in this higher-carbon steel. (d) 1.2% carbon steel: dark is pearlite, white grain boundaries are cementite.



Fig. 3. Effect of carbon content on the tensile properties of hot-worked carbon steels. $10^3\,psi=6.9$ MPa; 2 in. =5 cm.

of the low-carbon steels make it possible also to cold-work these steels in the ferritic condition. Coldrolled low-carbon steels are extensively used for sheet applications in the appliance and automotive industries. Cold-rolled steels have excellent surface finishes, and both hot- and cold-worked mild steels are readily welded.

The low-carbon ferritic steels have undergone interesting developments in order to meet the challenges of the energy crisis. Although mild steels have excellent formability, their low strength requires relatively thick sections for a given load-bearing application. New compositions of ferritic steels combined with new processing have significantly raised the strength of low-carbon steels, and therefore made possible significant reductions in vehicle weight by permitting the use of thinner sections.

One approach to improved low-carbon ferritic steels has led to the development of a class of steels referred to as high-strength low-alloy (HSLA). HSLA steels are alloyed with very small amounts of strong carbide-forming elements, such as vanadium and niobium. Very fine carbides are formed when the microalloyed steels are finish-hot-worked low in the austenite phase field. The alloy carbides limit the growth of the austenite grains, and therefore the size of the ferrite grains that form from the austenite. The resulting very fine ferrite grain size, together with the fine-alloy carbides, significantly increases strength. Special attention is also paid to inclusion content and shape in HSLA steels, and as a result of inclusion control and the very fine ferritic grain sizes, HSLA steels have excellent toughness for critical applications.

Another approach to producing high strength and good formability in low-carbon steel is to heat a steel with a ferrite-pearlite microstructure into the ferriteplus-austenite phase field (Fig. 1). Such intercritically annealed steels are referred to as dual-phase steels. The intercritical annealing treatment converts the pearlite areas into pools of austenite and, by controlling cooling rate or alloying the austenite, or both, transforms to structures other than pearlite on cooling. Good formability and high strength are achieved by the elimination of the pearlite and by the introduction of dislocations into the ferrite matrix surrounding the tranformed austenite. The dislocations are crystal defects that make possible the shaping of metals.

Medium-carbon steels. The medium-carbon steels contain between 0.25 and 0.70% carbon, and are most frequently used in the heat-treated condition for machine components that require high strength and good fatigue resistance (resistance to cyclic stressing). The heat treating consists of austenitizing, cooling at a high enough rate to produce a new phase, martensite, and then tempering at a temperature below the A_1 . The martensite is essentially a supersaturated ferrite with carbon atoms trapped between the iron atoms. This structure is very hard and strong. The hardness and strength of martensite increase rapidly with the carbon content of the steel, and the greatest benefits of forming martensite occur above 0.3% carbon content (**Fig. 4**).



Fig. 4. Relation of hardness and tensile strength to the carbon content in steels which are rapidly quenched from above critical temperature. $10^3 \, \text{psi} = 6.9 \, \text{MPa}.$

However, the same structural factors that make martensite very hard also make it very brittle. Therefore, martensitic steels are tempered to increase their toughness. The tempering essentially makes it possible for the carbon atoms to diffuse from the martensitic structure to form independent carbide particles. Thus the microstructure changes to a mixture of ferrite and carbides on tempering. The extent of this process is controlled primarily by temperature, and a large range of hardness and toughness combinations can be produced.

Martensite can be formed only when austenite transformation to ferrite or pearlite is suppressed. The ferrite and pearlite phases require the diffusion of carbon, a time-dependent process, for their formation. In plain-carbon steels the diffusion can be suppressed only by quenching or cooling very quickly in brine or water. Even when martensite is formed on the surface of a bar of plain-carbon steel, however, the center of the bar may cool too slowly to form martensite.

Small amounts of alloying elements such as chromium, nickel, and molybdenum reduce the rate at which ferrite or pearlite form, and therefore make it possible to form martensite in heavier sections or at slower cooling rates. Slower cooling reduces the tendency to distortion or cracking that sometimes accompanies the high residual stresses introduced by severe quenching. Hardenability is the term used to define the ease of martensite formation relative to cooling rates, section sizes, and steel composition.

High-carbon steels. Steels containing more than 0.7% carbon are in a special category because of their high hardness and low toughness. This combination of properties makes the high-carbon steels ideal for bearing applications where wear resistance is important and the compressive loading minimizes brittle fracture that might develop on tensile loading.

Two types of microstructure are produced in high-carbon steels. One consists entirely of a pearlite with a very fine lamellar spacing produced in steels with about 0.80% carbon. Figure 1 shows that no ferrite or cementite forms above A_1 at this carbon content. An important application of pearlitic 0.8% carbon steels is in railroad rail. Another application in which pearlitic microstructure is important is highstrength wire for cables and wire rope. In a process referred to as patenting, rods of 0.8% carbon steel are transformed to very fine pearlite. The rods are then drawn to wire to produce an aligned pearlite with strengths up to 350 ksi (2415 MPa).

The other type of microstructure produced in high-carbon steels is tempered martensite. The most common bearing steel contains 1% carbon and about 1.5% chromium. This steel is oil-quenched to martensite and tempered to retain as high a hardness as possible. Austenitizing is performed in the austenitecementite phase field. As a result, the microstructure contains very fine cementite particles which not only contribute to wear resistance but also help to maintain a fine austenite grain size. *See* CAST IRON; FERROALLOY; HEAT TREATMENT (METALLURGY); IRON; IRON ALLOYS; METAL, MECHANICAL PROPER-TIES OF; STAINLESS STEEL; STEEL MANUFACTURE.

George Krauss

Mechanical properties. The mechanical properties of steel are determined from a uniaxial tension test in which a cylindrical bar of measured diameter is subjected to tension along its length while the magnitude of the force applied and the resulting elongation are measured. The force, divided by the original cross-sectional area of the bar, is referred to as the stress, and the elongation of a specific portion of the bar, divided by the original length of that portion, is the strain. The strain is clearly dimensionless, but the stress has the units of force per unit area. Since the forces involved are large (in the American system, forces are expressed in kips, where 1 kip is 1000 lb), stresses are expressed in kips/in.2 (ksi). The SI unit of force is the newton, and 1 N/m^2 is the pascal (Pa). For the large stresses in steel, the usual unit is the megapascal, (MPa); 1 ksi = 6.895 MPa. A diagram plotted with stress as the ordinate and strain as the abscissa is referred to as a stress-strain diagram. For all steels with carbon generally making up less than 2% (structural steels), the initial portion of such a diagram is linear. The slope of this straight-line portion, the ratio of stress to strain, is the modulus of elasticity; it is about 29,000 ksi (200,000 MPa). At this time, the test specimen is also elastic, recovering its original dimensions if the load is removed, with unloading following down the same straight line plotted during loading. See ELASTICITY; STRESS AND STRAIN; STRUC-TURAL STEEL.

In the low-carbon and medium-carbon steels, linearity ceases abruptly in the yield zone, with a large increase in strain with little or no increase in load. The magnitude of the strain occurring during yield is 10–20 times as large as the strain at the end of the elastic region.

At the conclusion of the yield, strain-hardening occurs, and additional load is once again required to produce additional deformation, but at a much lower ratio of stress to strain. Eventually, the load reaches a maximum value, the ultimate tensile strength, where the specimen begins to neck down, developing a severely reduced cross section at some point, and quickly elongates a great deal and fractures. The total elongation is an order of magnitude larger than that during yield, two orders of magnitude above the elastic strain.

In the higher-carbon steels, there is generally no abrupt yield. The linear zone makes a transition to a curve going up to the ultimate strength and down to fracture. It is still important to mark the end of elastic behavior, and therefore a yield is defined for these steels, either in terms of a measured amount of strain (typically 0.005) or as an amount of inelastic, nonlinear strain, found as the point where the stressstrain curve intersects a straight line having the same slope as the initial portion but with an offset (typically 0.002).

The behavior of steel in compression is similar, provided that buckling of the specimen is prevented.

Of course, there is no neck-down or fracture, the specimen simply being squeezed down and shortened, and flowing out.

Commercially available steels have yield points ranging 33–100 ksi (230–700 MPa) and ultimate tensile strengths up to 130 ksi (900 MPa), although building codes and specifications generally restrict operating stress levels to values below the yield.

G. Donald Brandt

Bibliography. American Society for Metals, *Metals Handbook, Properties and Selection: Irons, Steels, and Higb Performance Alloys*, 10th ed., vol. 1., 1990; Association of Iron and Steel Engineers, *The Making, Shaping, and Treating of Steel*, 11th ed., 1998; H. K. D. H. Bhadeshia and R. W. K. Honeycombe, *Steels: Microstructure and Proper ties*, 3d ed., 2006; D. Llewellyn and R. Hudd, *Steels: Metallurgy and Applications*, 3d ed., 1998.

Steel manufacture

A sequence of operations in which pig iron and scrap steel are processed to remove impurities and are separated into the refined metal and slag.

Ironmaking

Ironmaking, one of the earliest of all metallurgical operations, is the initial step in most steel-producing operations. The blast furnace process for ironmaking came into being with the stone-built furnaces of the Middle Ages. Since then, ironmaking furnaces have evolved into towering structures that dominate the integrated steel mill.

Blast furnace. The iron blast furnace process consists of charging iron ore, coke, and limestone flux into the top of a shaft furnace and blowing heated air or blast into the bottom. This process yields a high-carbon iron containing silicon and manganese which is converted to steel in steelmaking furnaces.

The reactions occurring in iron- and steelmaking primarily involve adjustment of the oxygen potential—first, to decrease the oxygen potential and reduce the iron ore to an impure pig iron, and then to increase the oxygen potential and oxidize impurities from that iron in the steelmaking operation. In the iron blast furnace, the oxygen potential is extremely low in the hearth zone and continually increases as the reducing gases move up the furnace, reacting with iron oxide and coke.

Direct reduced iron. Another source of iron units for production of steel, in addition to pig iron and scrap, is direct-reduced iron. Metallic iron is produced from iron ores by removing most of the oxygen at temperatures below the melting points of materials in the process. Direct-reduced iron can be refined to steel or used to increase productivity of the blast furnace or other melting processes. Several processes have been developed to contact reducing agents with iron oxides, including fluidized beds, moving-bed shaft furnaces, fixed-bed retorts, and rotary-kiln processes.

Direct-reduction processes represent a small

fraction of iron production, but are being adopted where suitable ores and reducing agents, particularly natural gas, are available. Direct-reduced iron is very suitable as a charge material in steelmaking processes, and has special advantages in electric furnace steelmaking. *See* FLUIDIZATION; IRON METALLURGY; PRESSURIZED BLAST FURNACE.

Steelmaking Processes

Reduction of iron ores by carbonaceous fuel directly to a steel composition was practiced in ancient times, but liquid processing was unknown until development of the crucible process, in which iron ore, coal, and flux materials were melted in a crucible to produce small quantities of liquid steel. Modern steelmaking processes began with the invention of the airblown converter by H. Bessemer in 1856. The Thomas process was developed in 1878; it modified the Bessemer process to permit treatment of highphosphorus pig iron. The Siemens-Martin process, also known as the open-hearth process, was developed at about the same time. The open-hearth process utilizes regenerative heat transfer to preheat air used with a burner; it can generate sufficient heat to refine solid steel scrap and pig iron in a reverberatory furnace. After World War II, various oxygen steelmaking processes were developed.

Steelmaking can be divided into acid and basic processes depending upon whether the slag is high in silica (acid) or high in lime (basic). The furnace lining in contact with the slag should be a compatible material. A silica or siliceous material is used in acid processes, and a basic material such as burned dolomite or magnesite is used in basic processes. Carbon, manganese, and silicon, the principal impurities in pig iron, are easily oxidized and separated; the manganese and silicon oxides go into the slag, and the carbon is removed as carbon monoxide and carbon dioxide in the off-gases. Phosphorus is also oxidized but does not separate from the metal unless the slag is basic. Removal of sulfur occurs to some extent by absorption in a basic slag. Thus, the basic steelmaking processes are more versatile in terms of the raw materials they can handle, and have become the predominant steelmaking processes.

A typical pig iron charged to the steelmaking process might contain roughly 4% carbon, 1% manganese, and 1% silicon. The phosphorus and sulfur levels in the pig iron vary. The composition of the steel tapped from the steelmaking furnace generally ranges from 0.04 to 0.80% carbon, 0.06 to 0.30% manganese, 0.01 to 0.05% phosphorus, and 0.01 to 0.05% sulfur, with negligible amounts of silicon. The chemistry of steelmaking can be represented as reactions (1)-(5), which are treated in terms of oxidation by

$$C + (FeO) \rightarrow CO + Fe(I)$$

$$Si + 2(FeO) \rightarrow (SiO_2) + 2Fe(I)$$
 (2)

(1)

$$Mn + (FeO) \rightarrow (MnO) + Fe(I)$$
 (3)

$$2P + 4(CaO) + 5(FeO) \to 4CaO \cdot P_2O_5 + 5Fe(I)$$
 (4)

$$S + CaO \rightarrow CaS + O$$
 (5)

liquid iron oxide in the slag. Several alternative mechanisms are considered possible, such as direct oxidation by gaseous oxygen, reaction (6), or alternatively,

$$C + \frac{1}{2}O_2(g) = CO(g)$$
 (6)

the oxygen may dissolve in the liquid iron and react according to reaction (7). The removal of phospho-

С-

$$+0 = C0 \tag{7}$$

rus from the bath requires a basic oxidizing slag as illustrated by reaction (4). The actual form of the phosphorus in the slag is thought to be as tricalcium or tetracalcium phosphate, although evidence is not entirely clear in this regard. Sulfur is absorbed by a basic slag, although it may or may not form calcium sulfide as indicated by reaction (5).

Bessemer converter processes. The original Bessemer converter employed an acid lining. This acid process is no longer used extensively. The process requires a very low phosphorus- and sulfur-bearing pig iron since these elements are not removed in the process. A cross section of a Bessemer converter is shown in Fig. 1. Oxidation of the charge is accomplished by blowing air through a wind box on the bottom of the vessel. Elements are eliminated essentially in the same order as the stability of the oxide, that is, silicon is eliminated first, along with manganese, and finally carbon. The Bessemer process is autogenous-that is, no additional heat is supplied to the process-and the oxidation of the metalloids; particularly silicon, is sufficient to keep the charge molten and raise the temperature as purification proceeds so that a low-carbon steel at a temperature near 2900°F (1600°C) is produced at the end of the process. A small amount of cold scrap can be charged.

The Thomas or basic Bessemer process, used more extensively in Europe, is carried out in a basic lined vessel, and utilizes lime or limestone as a charge material to form a basic slag. Pig iron charged to this process contains up to 1 wt % phosphorus, which is oxidized from the bath after most of the carbon has been removed in a so-called afterblow. The



Fig. 1. Operating positions of a Bessemer converter. (a) Charging and pouring. (b) Blowing. (*After R. D. Pehlke*, *Unit Processes of Extractive Metallurgy*, p. 89, *American Elsevier*, 1973)



Fig. 2. Diagram of the principal parts of an open-hearth furnace. The arrows indicate the direction of the flow of preheated air, flame, and waste gases when a liquid fuel is fired through a burner in the trench at the right end of the furnace. The five charging doors shown are in the front wall of the furnace, and the checker chambers extend under the charging floor, which is not shown. When reversal of firing takes place, the function of the uptake is reversed, producing the downtake. (*After H. E. McGannon, ed., The Making, Shaping and Treating of Steel, United States Steel Corporation, 8th ed., p.* 471, 1964)

oxidation of the phosphorus is exothermic and adds considerable thermal energy to the process.

One of the disadvantages of Bessemer steelmaking is the resulting higher phosphorus and nitrogen contents in the steel. The nitrogen content of a Bessemer steel is approximately 0.015 wt %, although lower nitrogen contents can be achieved by careful control of temperature and by carrying out the final oxidation of the bath with iron ore additions.

An important aspect of converter steelmaking is control, because of the rapidity with which the end point of the process is approached, and the necessity for providing a careful thermal balance for the autogenous process. End-point control has been greatly enhanced by the use of photocells to observe the character of the flame and emission from the vessel toward the end of the blow.

Open-hearth process. From the turn of the century through the late 1950s, the basic open-hearth process was the workhorse of the American industry, accounting for approximately 85% of the steel produced in the United States in 1955. Because of its

melting capabilities, the basic open-hearth process can handle a wide range of pig iron and scrap ratios, the optimum operation being a metallic charge of 40 to 60% molten pig iron with the balance as steel scrap.

A typical open-hearth furnace is shown in **Fig. 2**. In an effort to obtain a high flame temperature and good thermal efficiency, the open hearth employs a regenerative heat transfer system. By reversing the direction of firing in the furnace, hot combustion gases pass out of the furnace through a set of brick checkers which recover heat from the expanded combustion gases. The incoming air is passed through the hot checker work on the opposite end of the furnace before being reacted with fuel at the burner. This technique for preheating the combustion air provides fuel economy.

In a basic operation, a wide range of raw materials can be treated, and the phosphorus and sulfur concentrations are kept under suitable control. An advantage of the open hearth over the Bessemer process is that the steels produced contain about 0.005 wt % nitrogen, which is low enough to provide good properties in deep-drawing sheet material.

The solid constituents-steel scrap, limestone, and iron ore-are charged to the furnace, and the burner turned on. Burners are built into both ends of the furnace, and the direction of the flame is changed from time to time to take advantage of the regenerative heat transfer system for the air which is located on either side of the furnace at a lower level. As soon as the charge starts to melt down, pig iron is added and the iron oxide reacts with the molten pig iron to produce carbon monoxide in a so-called ore boil. As the temperature increases, the decomposition of limestone at the bottom of the furnace occurs and a lime boil starts, with carbon dioxide being evolved at a relatively rapid rate. As the calcined lime floats up through the charge, phosphorus is removed as well as some sulfur. The productivity of the openhearth furnace had been limited by both charging time and the extensive melting and refining periods, requiring approximately 6 to 8 h for melting and refining alone. The availability of low-cost oxygen has permitted oxygen lances to be inserted through the roof of open-hearth furnaces, which decreases the time required to process a heat (single batch of steel) from about 12-14 h to approximately 6 h. Oxygen blowing, however, creates fumes and requires gascleaning equipment. This limitation on productivity and the added expense of gas-cleaning equipment in an older installation have led to rapid displacement of the open-hearth furnace by the oxygen top-blowing process.

Oxygen steelmaking. The high dissolved nitrogen content in Bessemer steel limits its application, and a nitrogen-free gas for blowing in the converter is needed. The use of pure oxygen in a bottom-blown vessel will result in extremely severe erosion of the bottom lining and tuyere area. An alternative is to blow a Bessemer converter with oxygen and steam to produce low-nitrogen steel. This process has received considerable attention in Europe.

In the early 1950s, commercial development of the oxygen steelmaking process was achieved at Linz, Austria. Since Austrian ores are low in phosphorus, they cannot support the basic Bessemer process thermally; furthermore, a shortage of scrap in Austria following World War II eliminated economic application of the open hearth. These factors along with availability of oxygen at much lower cost in tonnage quantities led to development of the top-blown oxygen process, which is based on experience with a small vessel at Gerlafingen, Switzerland. The first large-scale oxygen steelmaking converters were 30ton vessels. Further development of the process, particularly in the United States, has increased the size of the vessel until most oxygen steelmaking shops employ 250-ton vessels, with some vessels as large as 400 tons in operation.

A schematic cross section of an oxygen steelmaking vessel is shown in **Fig. 3**. **Figure 4** shows the oxygen jetting of a 300-ton heat of steel. The process is autogenous and uses about 25 to 35% scrap, the remainder being basic pig iron which is relatively low



Fig. 3. Schematic section of oxygen steelmaking vessel.

in sulfur. Burned lime is used as a flux, with enough added to produce a lime-silica ratio of around 3 to 1 in the final slag. Oxygen is blown down vertically into the bath through a Laval nozzle, which produces supersonic flow. With larger vessels, the use of a threehole lance has improved operation.

Copious fumes in the form of iron oxide particles around a micrometer in size are evolved during the oxygen blow. This fume loss amounts to 2 to 3% of the metallic charge, and gas-cleaning equipment represents a large portion of the capital expenditure for a basic oxygen furnace plant. Significant operating variables include the lance height and oxygen flow



Fig. 4. Large oxygen steelmaking furnace in operation. (Bethlehem Steel Corporation)

rate. The required blowing time for a heat, including those approaching 250 tons, is approximately 20 min. By suitable adjustment of the lance height, some iron can be oxidized early in the blow to dissolve the lime and produce an early basic oxidizing slag, which assists in phosphorus removal. Later in the heat, lowering the lance or increasing the oxygen flow rate provides the deep penetration necessary to agitate metal and slag and provide for relatively low iron losses in the slag.

The processing of high phosphorus irons in the top-blown oxygen steelmaking process can be accomplished by injecting powdered lime into the oxygen stream. This technique, known as the LD-AC process, is used in Europe to process pig iron containing more than 0.5 wt % phosphorus.

Oxygen converter gas recovery. Many oxygen steelmaking vessels have open hoods into which air is ingested to burn the hot evolved converter gases, which contain about 90% carbon monoxide during much of the oxygen-blowing period. Considerable effort has been made, particularly in Japan, to recover the converter gas, which averages over 80% carbon monoxide, for use as a fuel in the steel plant. The use of a closed hood and special precautions to prevent explosion of the carbon monoxide has resulted in an economical process for recovering the thermal energy in the carbon monoxide. In most hoods, a waste heat boiler is employed to recover the sensible heat of the stack gas; but a marked improvement in thermal efficiency is achieved by collecting the carbon monoxide gas for use as a supplementary fuel in the steel plant. Furthermore, the gas-cleaning problems are decreased because the volume of gas is only approximately one-sixth as great as when air is ingested in an open-hood system.

Submerged injection processes. Submerged injection processes were developed first for several European plants, in which bottom-blown Thomas (basic Bessemer) vessels were converted to use oxygen and hydrogen-bearing gases such as hydrocarbons, steam, or natural gas. In France, G. Savard and R. Lee in 1966 patented a dual tuyere with which oxygen could be injected through an inner tuyere and a gaseous hydrocarbon injected through the surrounding annular space formed by the central pipe and the surrounding outer pipe. This arrangement, utilizing the effect of cracking the hydrocarbon to cool the surrounding refractory, permitted bottom blowing of oxygen in steelmaking converters.

Several conversions of Thomas converter plants to the bottom-blown oxygen process were made by using natural gas in the outer surrounding tuyere. These steelmaking vessels varied in size up to 50 tons capacity; in 1973 the scale-up to 200-ton vessels was accomplished by the United States Steel Corporation.

The bottom-blown oxygen steelmaking process utilizes a converter with a replaceable bottom containing dual tuyeres with oxygen introduced through the central outlet. The number of tuyeres increases with the size of the furnace. Tuyere placement is critical in order to take advantage of the stirring capability of bottom injection, particularly under conditions where oxygen flow rates are decreased. Replacement of the bottom refractory plug for the vessel, which contains the tuyeres, is required only as often as the replacement of the remaining vessel refractory lining, typically after more than 1000 heats.

The absence of a lance in the bottom-blowing process has offered opportunities to decrease total plant height and has allowed relatively easy conversion of older open-hearth buildings to oxygen steelmaking shops.

Simultaneous top and bottom blowing has been explored to combine the advantages of each type of oxygen steelmaking. The joint effects of simultaneous top and bottom blowing of oxygen and mixed gases have resulted in higher metallic yield, improved manganese recovery, better control of oxygen content, and lower oxygen concentrations, as well as improved dephosphorization.

Electric furnace process. Electric arc furnace technology began late in the nineteenth century with the original design of P. L. T. Heroult. The three-graphite electrode furnace with a swinging roof for top charging and a rocker base for tilting to tap the finished molten steel has been continuously improved and developed further. *See* ELECTRIC FURNACE.

In its early form, the electric arc furnace had a small capacity and was used only for high-alloy and stainless steels, and for 5- or 10-ton heats of various steels cast into sand molds to produce shaped castings in foundries. Modern direct-arc furnaces ranging in size up to 400 tons are used to produce all types of steel. The electric furnace may be either acid- or basic-lined (Fig. 5). An advantage over the open-hearth and oxygen converter processes is that very high temperatures can be produced. The high temperature and the flexibility of using a 100% solid charge, or up to 60% molten pig iron, provides considerable versatility for this melting unit. The very high temperatures are particularly useful in producing alloy steels. If large amounts of alloying elements are added to a ladle from an oxygen converter or an open-hearth furnace, the steel is often cooled to the extent that good-quality steel cannot be poured, or if the alloving elements are added in the furnace, they may be oxidized into the slag. This is particularly true for elements such as niobium (columbium), vanadium, and chromium. Furthermore, the requirement in certain types of stainless steel to provide for very low carbon contents requires that the carbon be removed without excessive chromium losses. This is possible at high temperatures such as can be achieved in the electric arc furnace.

The rapid development of steelmaking technology using the electric arc furnace, not only for alloy and stainless steels but especially for carbon steel production, has increased its share of production capacity to about 20% of the steel industry. This increased capacity reflects the ascension of the "mini-mill" as a major steel supplier. The size of the mini-mill also has dramatically increased to the point where this label is not descriptive of this type of steel plant,



Fig. 5. Schematic cross section of a Heroult electric arc furnace with a flat-bottomed shell and stadium-type subhearth construction, indicating typical refractories employed in (a) an acid lining and (b) a basic lining. Only two of the three electrodes are shown in this section. (*After H. E. McGannon, ed., The Making, Shaping and Treating of Steel, 8th ed., United States Steel Corporation, 1964*)

several such mills being located within large integrated plants and having two or more 200-ton electric arc furnaces.

The modern mini-mill includes one or more electric furnaces with high-power energy sources; a 100ton electric furnace may have a 75-MW transformer. These furnaces are water-cooled, at least in the upper wall where cooled panels substitute for refractory above the slag line. Water-cooled roofs are in use, and water-cooled electrode holders are being developed to provide reduced electrode consumption.

The product line of the mini-mill includes specialquality bars and rods, as well as reinforcing bars and small structural shapes. Some mills are producing plate from electric furnace remelting operations. This technology, which may depend on direct-reduced iron in the future, is important in steel production.

Special processes. Remelting and refining of special alloys are carried out in duplex or secondary processes; the principal ones are argon-oxygen decarburization, electroslag refining, vacuum arc remelting, and vacuum induction melting.

In the argon-oxygen decarburization process, a liquid metal is refined in a refractory-lined vessel by injection of gas mixtures, in particular of varying argon-oxygen mixtures, through tuyeres in the side or bottom of the unit. Usually the liquid metal is provided from an electric arc furnace, or in smaller installations from an induction furnace. The argonoxygen decarburization process was developed originally for the production of stainless steels, but has been modified for the production of many other grades. The argon dilution of the gases in the process provides excellent degassing characteristics which allow achievement of low carbon contents in the liquid metal without chromium losses by oxidation to Cr₂O₃ into the slag; also hydrogen is removed from crack-sensitive grades of steel. Argon-oxygen decarburization provides good slag-metal mixing for removal of sulfur and separation of nonmetallic inclusions.

In the electroslag refining process, an impure electrode prepared from the material to be refined is remelted in a slag which is resistively heated. An electric current passed through the slag melts the electrode. Molten droplets of metal form on the tip of the electrode and are refined as they fall through the slag, forming an ingot in a water-cooled mold. *See* ELECTROMETALLURGY.

In the vacuum arc remelting process, a prepared electrode is arc-melted in a vacuum. The droplets of metal from the electrode are exposed to the vacuum and are solidified in a water-cooled mold. The process provides refining of volatile constituents and removal of dissolved gases, and promotes chemical homogeneity and improved structure.

Control of temperature, pressure, and stirring can be exercised over wide ranges and for extended periods of time in the vacuum induction melting process, allowing considerable flexibility for refining of an alloy. Vacuum induction melting is applied to small melts, as well as large ones ranging up to 60 tons (54 metric tons). The molds are usually enclosed in an adjoining vacuum chamber connected by largescale access doors and valves; this minimizes the time required to melt and prepare molds, since they can be done independently. Vacuum induction melting systems, essentially the same as the conventional induction melting systems, utilize high-frequency current in a water-cooled coil surrounding the crucible containing the melt; however, the processing is conducted under low vacuum pressures. *See* VACUUM METALLURGY.

Ladle metallurgy. Ladle metallurgy was used first to produce high-quality steels, but has been extended to producing many grades of steel because of the economic advantages of higher productivity. The purpose of these ladle treatments is to produce clean steel; introduce reactive additions, such as calcium or rare earths; add alloying additions, as for microalloyed steels, with high recovery; and increase furnace utilization, allowing higher-productivity smelting operations of the blast furnace, and melting and refining operations in steelmaking, followed by final ladle treatment of the steel to achieve controlled low levels of contained inclusions. Ladle treatments in steel production generally are classified as synthetic slag systems; gas stirring or purging; direct immersion of reactants, such as rare earths; lance injection of reactants; and wire feeding of reactants. These are often used in combination to produce synergistic effects, for example, synthetic slag and gas stirring for desulfurization followed by direct immersion, injection, or wire feeding for inclusion shape control.

A major aspect of steel development since the mid-1970s has been the effort directed at the improvement of physical and mechanical properties through steel cleanliness and reduced anisotropy, which are largely dependent upon the amount and form of oxygen and sulfur in the steel. Such improved properties are required by steels used for thin-sheet materials, heavy plate, high-strength lowalloy (HSLA), and for linepipe. Thin sheet steels require good formability; heavy plate requires guaranteed through-thickness properties; and linepipe steels require high-transverse-impact energy values. Removal of reaction products is most critical during the course of deoxidation of molten steel. Nonmetallic inclusions in the form of oxides of silicon, manganese, and aluminum may adversely influence final properties.

Sulfur in steel has always presented a problem in steelmaking. Improved surface quality can be achieved by lowering sulfur levels to 0.015% or below. For example, the achievement of low sulfur contents with two-slag practices in electric furnaces presents operating difficulties and production delays.

In the 1950s steel was produced primarily in the open-hearth furnace, where sulfur levels were normally in the range from 0.03 to 0.04%. Sulfur contents of 0.02 to 0.03% could be achieved with lower sulfur fuels and oxygen roof lances, permitting shorter heat times and improved mixing of metal and slag. The basic oxygen steelmaking process brought about an even further improvement within the range of 0.015–0.025% sulfur, and with the bottom-blown oxygen converter Q-BOP (Quelle-Basic oxygen process), sulfur contents in the 0.01–0.02% range can be achieved. However, even lower sulfur contents are required.

In killed steels with low oxygen content, such as when aluminum is used for deoxidation and grain size control, sulfur combines with manganese as highly deformable manganese sulfides. These manganese sulfides have a low melting point and, as the last liquid to solidify in the steel, collect as films at grain boundaries. During hot rolling, the manganese sulfides are plastically deformed into elongated stringers extending parallel to the rolling direction. This shape and distribution of sulfides can have a marked effect on the directional properties of steel products.

Two methods can be utilized to minimize the directionality problem. One is to reduce the sulfur content in the steel product to a minimum, lowering the total volume of manganese sulfide inclusions. This can be accomplished by decreasing the total sulfur input, treating the hot metal to remove sulfur, and controlling the scrap charge; by adjusting melting practices, including the use of additional quantities of flux; and by direct treatment of the liquid steel with desulfurizing agents, such as calcium or magnesium. A second approach involves modifying the shape or morphology of sulfide inclusions by producing a relatively nondeformable complex sulfide or oxysulfide by the addition of calcium or rare-earth metals.

Reoxidation and inclusion minimization. Since there is a strong affinity between oxygen and liquid steel, and the necessity to restrict pickup of oxygen during teeming operations, several steps to control reoxidation are advisable. The first is the protection of the steel pouring stream, whether from ladle to ingot mold or tundish, or from tundish to continuous casting mold. The separation of reoxidation products, or deoxidation products formed during cooling of the steel in the teeming operation, can be promoted. In particular, in continuous casting the appropriate design of tundish in terms of residence time of the steel in the tundish and the appropriate fluid flow patterns can enhance separation of oxides from the liquid steel. See METAL CASTING; PYROMETALLURGY; REFRACTORY; STAINLESS STEEL; STEEL.

Robert D. Pehlke Bibliography. Association of Iron and Steel Engineers (AISE), *The Making, Shaping, and Treating* of Steel, 10th ed., 1985; A. Ghosh, Secondary Steelmaking: Principles and Applications, 2000; Iron and Steel Society, Electric Furnace Steelmaking: Collective Work, 1985; T. Kawaguchi and K. Sugiyama, Steel Industry I: Manufacturing System 1989.

Stellar evolution

The large-scale, systematic, and irreversible changes with time of the structure and composition of a star.

Types of stars. Dozens of different kinds of stars populate the Milky Way Galaxy. The most common

are main-sequence dwarfs like the Sun that fuse hydrogen into helium within their cores (the core of the Sun occupies about half the mass). Dwarfs run the full gamut of stellar masses, from perhaps as much as 120 solar masses (120 M_{\odot}) down to the minimum of 0.075 solar mass (beneath which the full protonproton chain does not operate). They occupy the spectral sequence from class O (maximum effective temperature 50,000 K or 90,000°F, maximum luminosity 5 \times 10⁶ solar), through classes B, A, F, G, K, and M, to the new class L (2000 K or 3100° F and under, minimum luminosity 4×10^{-5} solar). Within the main sequence, they break into two broad groups, those under 1.3 solar masses (class F5) whose luminosities derive from the proton-proton chain, and higher-mass stars that are supported principally by the carbon cycle. Below the end of the main sequence (masses less than 0.075 M_{\odot}) lie, the brown dwarfs that occupy half of class L and all of class T (the latter around 1000 K or 1300°F). These shine both from gravitational energy and from fusion of their natural deuterium. Their low-mass limit is unknown. See CARBON-NITROGEN-OXYGEN CYCLES; PROTON-PROTON CHAIN; SPECTRAL TYPE; STAR.

Scattered among the dwarfs are different kinds of stars distinguishable (through comparison with the dwarfs) by their sizes and luminosities: bright giants (with maximum radii equaling that of the inner solar system), brilliant supergiants (maximum radii comparable to the orbits of Jupiter and Saturn), subgiants (which fall between giants and dwarfs), and dim white dwarfs (the size of Earth). All these classifications may be broken into subclasses. The list also includes the central stars of planetary nebulae, double stars, novae, supernovae, neutron stars, and black holes. The study of stellar evolution seeks to determine where they all came from. Do they stand alone, or are they somehow connected? If the latter, what kinds of stars precede and succeed what other kinds?

Stellar models. Stars age so slowly that it is impossible-except under rare circumstances-to see the transformations take place. Instead, stellar theorists, using the laws of physics, build numerical models of stars and then age them. The procedure is in principle simple, in practice difficult. At the start, a star is in hydrostatic equilibrium in which each layer supports the layers above it (the pressure difference across a layer equaling the weight of the layer), and is internally supported by thermonuclear fusion. The relation between pressure, density, and temperature is given by the perfect gas law, P =*NkT*, where *P* is pressure, *N* is the number density of atomic particles, k is Boltzmann's constant, and T is temperature. As fusion proceeds, and the rate of energy generation and the internal chemical composition change, the structure of the star changes as well. Stepwise calculations that predict luminosity, radius, and effective temperature then allow the star to be followed as it "moves" (changes effective temperature and luminosity) on the Hertzsprung-Russell (HR) diagram. When appropriate mass loss is taken into account, most of the different kinds of stars fall into place. *See* GAS; HERTZSPRUNG-RUSSELL DIAGRAM; HYDROSTATICS; THERMODYNAMIC PROCESSES.

Star formation. Stars are born from compact knots within dark molecular clouds that are refrigerated by dust that blocks heating starlight. If the random knots, compressed by supernovae or other means, are dense enough, they can contract under their own gravity. Conservation of angular momentum demands that as they collapse they must spin faster. Star formation requires that angular momentum be removed such that the new "young stellar objects" do not completely tear themselves apart before they can become fully developing protostars. *See* ANGULAR MOMENTUM; MOLECULAR CLOUD.

High-speed particles (cosmic rays) from exploding stars partially ionize the dusty knots. The ions grab onto the weak magnetic field of the Galaxy and, as a result of their physical interaction with neutral atoms and molecules, provide the initial means to slow the rotation. If the rotation is still too fast, the contracting body may split into a double (or more complex) star, though the origins of doubles are not clearly solved. A contracting protostar still indeed rotates progressively faster until the part of its mass not accreted by the star itself is spun out into a dusty disk, from which planets might later accumulate. From the disk shoot powerful molecular flows that slow the star still more (**Fig. 1**). *See* PROTOSTAR.

When the protostar's interior reaches about 10^6 K (1.8×10^6 °F), it can fuse its internal deuterium. That and convection, which brings in fresh deuterium from outside the nuclear-burning zone, bring some stability, and a star can now be said to be born. Stars like the Sun shrink at constant temperature until deuterium fusion dies down. Then they heat at roughly constant luminosity until the full proton-proton chain begins, which provides the stars' luminosity and stops the contraction. The stars settle onto the zero-age main sequence (from which they



Fig. 1. Focused, even created, by a circumstellar disk, a jet pours from the youthful star Herbig-Haro 30 (HH 30). (C. Burrows, Space Telescope Science Institute; WPFC2 Science Team; NASA)



Fig. 2. Hertzsprung-Russell diagram of star formation. Stars begin to stabilize at the birthline and then descend on evolutionary tracks as T Tauri stars (dots) toward the zero-age main sequence. (*After J. B. Kaler, Cosmic Clouds, W. H. Freeman, 1997, from work of S. W. Stahler and I. Iben, Jr.*)

will later evolve). At the same time, the surrounding dusty cloud is clearing, allowing new, accreting, and highly active T Tauri stars to be seen flocking around their birthclouds (**Fig. 2**). The whole process takes only 10 or so million years, the mature stars then leaving their birthplaces, destined both to age and to orbit the Galaxy. High-mass stars proceed similarly, but at such a great pace that the death process begins even as the birth process is ending. *See* T TAURI STAR.

Early evolution. The main sequence is that zone on the HR diagram in which stars are stabilized against gravitational contraction by fusion. The higher the stellar mass, the greater the internal compression and temperature, and the more luminous the star. Hydrogen fusion is highly sensitive to temperature, a small increase in stellar mass meaning a much higher fusion rate. Although greater mass means a greater nuclear-burning core mass and therefore a larger fuel supply, the increased fusion rate is more than offsetting and thereby shortens stellar life. A star's fuel supply is proportional to mass (M), while the rate of fusion is expressible by the luminosity (L). The average mass-luminosity relation found from binary stars is $L \propto M^{3.5}$. If a star's lifetime t is proportional to (fuel supply)/(rate of use), then $t \propto M/L \propto M/M^{3.5} =$ $1/M^{2.5}$. While the new Sun was destined to survive

on the main sequence for 10^{10} years, a 0.1-solarmass star will live there for 10^{13} years, while a 100-solar-mass star will exhaust its core hydrogen in only 2.5×10^6 years.

Main-sequence life, while stable, is not altogether quiet. Even there stars change and evolve. As hydrogen "burns" to helium, four particles (protons) are converted to one (a helium nucleus). The pressure of a gas depends on the number of particles per unit volume, not on their kind. The result is a slow shrinkage of the core, which increases the temperature, raises the fusion rate, and causes the core to eat into the surrounding hydrogen envelope (incorporating fresh fuel). As a result, main-sequence dwarfs slowly brighten and eventually expand and cool some at their surfaces. (The Sun will more than double in brightness, dooming life on Earth long before core hydrogen is gone.) As a result, the main sequence spreads itself into a band toward the cool side (to the right) of the zero-age main sequence, the band widening toward the top.

Clusters of stars are born together in both space and time, most (presumably) with fully intact main sequences. As a cluster ages, stars peel off the main sequence (to become giants and supergiants) from the top down. A star cluster can be dated by where its main sequence ends. Open clusters, which occupy the Galaxy's disk, range from just-forming to about 10^{10} years old, which gives the disk's age. The globular clusters of the Galaxy's halo, however, have burned their main sequences down to around 0.8 solar. This process takes roughly 12×10^9 years, which must be close to the age of the Galaxy. *See* MILKY WAY GALAXY; STAR CLUSTERS.

The main sequence is divided into three parts. Below 0.8 solar mass (roughly class G8), no star has ever had time to evolve. Low-mass (K and M dwarfs) evolution is therefore of only academic interest. Stellar evolution deals with less than 15% of the stars. Their fates depend again on mass. Between 0.8 solar mass and around 10 solar masses (classes G8 to B1), the stars die as white dwarfs. Above it (classes O and B0), they explode. Binary stars contribute further to the richness of stellar phenomena.

Intermediate-mass evolution. Main-sequence life lasts until the core hydrogen is almost gone, at which time hydrogen fusion rather suddenly shuts down. With no support, the now-quiet helium core can contract more rapidly under gravity's force. It heats, causing hydrogen fusion to spread into a thick enclosing shell that runs on the carbon cycle. With a new (though temporary) energy source, the star first dims some while it expands and cools at the surface, changing its spectrum to class K. The transition takes only a few hundred million years or less, leaving few stars in the middle of the HR diagram, the lower masses appearing as F, G, and K subgiants.

The consequences again depend on mass. As core contraction proceeds beyond the rightward transition in the HR diagram, stars from about 1 to 5 solar masses (still fusing hydrogen in a shell) suddenly and dramatically increase their luminosities. The future Sun will eventually grow 1000 times brighter than it is today, and a 5-solar-mass star (which begins at about 600 solar luminosities) will reach nearly 3000 solar. At the same time, the stars swell to become red giants. While the core (roughly half a solar mass) of the future Sun shrinks to the size of Earth, the radius will expand to that of Mercury's orbit, or even beyond. When the core temperature climbs to 10^8 K, helium nuclei (alpha particles) begin fusing to unstable beryllium (⁸Be), which quickly decays back to alpha particles, setting up an equilibrium. The tiny amount of ⁸Be present reacts with additional alpha particles to create helium via the triple alpha process:

> ${}^{4}\text{He} + {}^{4}\text{He} \leftrightarrow {}^{8}\text{Be}$ ${}^{8}\text{Be} + {}^{4}\text{He} \rightarrow {}^{12}\text{C} + \text{gamma ray}$

(The "skip" over lithium, beryllium, and boron renders these elements rare.) The initiation of the triplealpha process is explosive in stars like the Sun, since the core electrons become degenerate, like those that support white dwarfs. (The sudden burst of energy is absorbed and does not reach the stellar surface.) Fusion with additional helium nuclei creates oxygen and even neon. Above 2 solar masses, helium burning starts more quietly (**Fig. 3**). *See* NUCLEOSYN-THESIS.

The star, now stabilized by a helium-burning core that is surrounded by a hydrogen-fusing shell, retreats about halfway down the red giant branch. The numerous lower-mass stars reside in the class K "red giant clump." Low-mass metal-deficient globular cluster stars that have suffered different rates of mass loss spread out from the clump toward higher temperatures to create the distinctive horizontal branch. Energy-generating fusion reactions will try to proceed toward iron (the most stable of all nuclei). Some 80% of the energy is generated in hydrogen burning, so (discounting further burning modes) the helium-burning stage lasts only around 20% of the main-sequence lifetime. From 5 solar masses up, evolution proceeds similarly, but instead of settling into a distinct location on the HR diagram, the stars loop to the blue (higher temperatures) where they fuse helium as class G, F, and A giants. This change in evolutionary style is not abrupt, but gradual with increasing mass. See GIANT STAR.

Asymptotic giant branch (AGB). When the helium has fused to carbon and oxygen, the core again contracts. Helium fusion spreads outward into a shell, and for the second time the star climbs the HR diagram's giant branch. Since the second climb is roughly asymptotic to the first, the second climb creates the asymptotic giant branch. The shrinking carbon-oxygen core is now surrounded by a shell of fusing helium, while the old hydrogen-fusing shell expands, cools, and shuts down. Eventually, however, the helium shell runs out of fuel, and the hydrogen shell reignites. Hydrogen burning feeds fresh helium into the space between it and the carbon core, and when there is enough of it, helium burning reignites explosively in a helium flash (or thermal pulse) that can affect the star's surface. The



Fig. 3. Hertzsprung-Russell diagram of evolution of intermediate-mass stars. These stars evolve from the main sequence to become giants that stabilize during helium fusion in a "clump." When helium fusion is done, they brighten again as AGB stars, lose their envelopes, and evolve to the left to produce planetary nebulae, finally at lower left becoming white dwarfs. (*After J. B. Kaler, Stars, W. H. Freeman, 1992, 1993, from work of I. Iben, Jr.*)

flash squelches hydrogen fusion, and the whole process starts again, helium flashes coming at progressively shorter intervals. AGB stars become larger and brighter than before, passing into the cool end of class M, where they eventually become unstable enough to pulsate as long-period variables (Miras). The Sun will become 5000 times brighter than now and will reach out to the Earth's orbit, perhaps destroying the planet. More massive Miras can exceed the size of the Martian orbit. *See* MIRA; VARIABLE STAR.

During this activity, stars go through various stages of convective dredge-up, in which they can raise elements created in their nuclear-burning zones to the stellar surfaces. The first of these dredge-ups alters the nitrogen and carbon isotope ratios while the star is on the first ascent of the red giant branch. Above about 3 solar masses, a second dredge-up in the early AGB stage can increase both surface nitrogen and helium. In the helium-flashing state, even carbon from helium fusion can be brought upward. Moreover, a host of elements created by slow neutron capture are elevated as well, including zirconium, strontium, barium, and many others. When the carbon abundance equals that of oxygen, an S star is seen (loaded with fresh zirconium); when carbon exceeds oxygen, a genuine carbon star is seen. See CARBON STAR.

Mass loss and planetary nebulae. During the giant stages, stellar winds greatly increase. Mira pulsations cause shock waves that help drive mass from the stellar surfaces, where the cooled gas becomes ever richer in molecules, some even condensing into dust grains. High luminosity pushes the dust outward, and the dust couples with the gas, resulting in slow (10 km/s or 6 mi/s) thick winds tens of millions of times stronger than the solar wind (up to 10^{-5} solar mass per year). Because stars are in this state for hundreds of thousands of years, they will lose much of themselves back into space-the Sun about half of its mass, a 10-solar-mass star over 80%. Advanced giants and Miras lose so much that they can visually disappear within dust clouds. Carbon star winds contain various kinds of carbon dust as well as organic molecules, while oxygen-rich stars produce silicate dust. The dust blown into interstellar space helps make new generations of stars. The elements created in the nuclear-burning shells are added to the interstellar inventory, where they too find their way into new generations of both stars and planets.





Fig. 4. Thick molecule-filled cloud surrounding the carbon star IRC+10 216. (a) Image made in radiation from hydrocyanic acid (HCN). (b) Image in radiation from cyanoacetylene (HC₃N). Each map is 6000 astronomical units across. (*J. Bieging; Berkeley-Illinois-Maryland Association*)



Fig. 5. Hubble Space Telescope image of planetary nebula NGC 7662. Planetary nebulae are formed when giant stars slough their outer envelopes, leaving a white dwarf behind. (B. Balick et al.; Space Telescope Science Institute; NASA)

A good portion of the carbon, most of the nitrogen, and significant fractions of the Galaxy's other elements come from such winds (**Fig. 4**).

So much mass is lost that an evolving star becomes stripped nearly to its fusion zone, which is protected from the outside by a low-mass hydrogen envelope. As the inner region becomes exposed, the wind diminishes in mass but increases in speed and temperature. Hammering at the surrounding dusty, molecule-filled shroud of lost mass, the high-speed wind compresses the inner edge into a thick ring. Eaten away from the top by the wind and from below by fusion, the stellar envelope shrinks, slowly exposing the hot shell-core structure beneath. When the stripped star's surface reaches 25,000 K (45,000°F), the dense ring that its high-speed wind had previously created is ionized. Subsequent recapture of electrons by ions, along with collisional excitation of heavy atoms, causes the shell to glow, and a planetary nebula is born. The first of these was announced by William Herschel (who named the class on the basis of their disklike shapes) in 1785. Over a thousand are now known (Fig. 5). See PLANETARY NEBULA.

As the planetary nebula, expanding at about 20 km/s (12 mi/s), grows in size, it eventually becomes so tenuous that the ionizing radiation bursts through, allowing much of the full structure to be seen. After 50,000 years, the nebula invisibly merges with the interstellar medium. During this time, the star inside first heats at constant luminosity to over 100,000 K (180,000°F), the luminosity and final temperature depending on the old core's mass (which ranges from around 0.5 M_{\odot} to nearly 1.4 M_{\odot} , the Chandrasekhar limit, above which white dwarfs cannot exist). As residual nuclear fusion shuts down, the star cools and dims at constant radius to become a white dwarf. The luminosity-mass relation is now



Fig. 6. Hertzsprung-Russell diagram for evolution of high-mass stars. These stars evolve to red supergiants, some evolving back to blue supergiants. Above about 60 solar masses, evolution is stalled by extreme mass loss. (*From work of A. Maeder and G. Meynet; and R. Humphreys and K. Davidson*)

reversed, as higher mass means higher gravitational compression and smaller radius.

Made of carbon and oxygen surrounded by a thin hydrogen or helium atmosphere, the white dwarf is supported forever by degenerate electron pressure. Since the evolutionary and cooling times of white dwarfs can be calculated theoretically, the oldest of them provides an age measure for the galactic disk, which accords well with that found from open clusters. *See* WHITE DWARF STAR.

High-mass evolution. As the mass of a star increases, so does the mass of the core. The Sun will turn into a white dwarf of around 0.6 solar mass. At 10 solar masses, the core reaches the Chandrasekhar limit, and the star cannot become a white dwarf. At first, high-mass evolution proceeds similarly to that of stars of lower mass. As high-mass stars use their core hydrogen, they too migrate to the right on the HR diagram, becoming not so much brighter but larger, cooling at their surfaces and turning into supergiants. Below about 40 solar masses, they become class M red supergiants, losing huge amounts of mass through immense winds. Some (depending on mass) stabilize there by the start of helium fusion; others loop back to become blue supergiants. Above 60 solar masses, so much mass is lost through winds that the stars do not make it much past class B, stalling there as helium fusion begins (Fig. 6). As the stars-as well as bright giants-loop through the middle of the diagram, some become Cepheid variables. See CEPHEIDS; SUPERGIANT STAR.

Though almost all these supergiants vary to some

extent, the most massive become unstable and undergo huge eruptions. In 1846, Eta Carinae, near 100 solar masses (and a probable double star), brightened to match Sirius and Canopus, lost a solar mass, and then faded to near the edge of naked-eye vision (**Fig. 7**). Such luminous blue variables (LBVs), plus similar lesser lights, are losing their hydrogen envelopes and may be turning into helium-rich Wolf-Rayet stars, highly luminous stars that have been



Fig. 7. Hubble Space Telescope image of a luminous blue variable, the supergiant Eta Carinae, buried in the middle of a giant expanding cloud of its own making. (*J. Morse, University of Colorado; Space Telescope Science Institute; NASA*)


Fig. 8. Field around the type II core-collapse supernova SN 1987A. (a) Two weeks after the supernova exploded in the Large Magellanic Cloud in 1987. The supernova far outshines its neighboring stars, and was easily visible 150,000 light-years away. (b) Before the explosion it was a rather ordinary B1 blue supergiant (indicated by arrow). (© Anglo-Australian Observatory, photograph by David Malin)

almost completely stripped of their outer hydrogen. Though the sequence is not clear, the luminous blue variables probably first turn into the nitrogen-rich WN variety (whose nitrogen has been enriched from hydrogen fusion by the carbon cycle) and then into the carbon-rich WC species (the carbon from more advanced helium fusion).

While intermediate-star fusion stops at carbon and oxygen, supergiants continue onward. The carbonoxygen core shrinks and heats to the point that carbon fusion can begin, and then carbon and oxygen convert to a more complex mix dominated by oxygen, neon, and magnesium. Helium fusion now continues in a surrounding shell that is nested in one that is fusing hydrogen. Once carbon burning has run its course, the unsupported oxygen-neonmagnesium core shrinks and heats, and now it is carbon burning's turn to move outward into a shell. When hot enough, the oxygen-neon-magnesium mix fires up to burn to one dominated by silicon and sulfur. Continuing the process, the developed silicon and sulfur core finally becomes hot enough to fuse to iron, the silicon-burning core wrapped in oxygenneon-magnesium, helium, and hydrogen-burning shells.

Supernova. Each nuclear fusion stage generates less energy, and since each takes on the role of supporting the star, each lasts a shorter period of time. While hydrogen fusion takes millions of years, the iron core is created from silicon fusion in a matter of weeks. Iron cannot fuse and produce energy. The core, about 1.5 solar masses and the size of Earth, suddenly collapses at a speed a good fraction that of light. The iron atoms are broken back to neutrons and protons. Under crushing densities, free

electrons merge with protons to make yet more neutrons. When the collapsing neutron core hits nuclear densities of 10^{14} g/cm³ (10^{14} times the density of water), it violently bounces, and the rebound tears away all the outer layers.

From the outside, the observer sees the explosion as a type II supernova that can reach an absolute magnitude of -18 (**Fig. 8**). Temperatures within the exploding layers are high enough that a great variety of nuclear reactions, from equilibrium fusion of one element into another to rapid neutron capture, create all the chemical elements up to and beyond uranium. The optical glow eventually comes mainly from the manufacture of a tenth or so of a solar mass of radioactive nickel, ⁵⁶Ni, which quickly decays to radioactive cobalt and then to stable iron, ⁵⁶Fe, which gets blasted into the cosmos along with everything else. *See* RADIOACTIVITY.

The nature of the rebound is far from understood. By itself, it does not account for the explosion of a supernova. However, the merger of core protons and electrons generates vast numbers of near-massless neutrinos. Nearly everything is transparent to them. Those created in the solar core by the proton-proton chain go right through the Earth. The densities around the rebounding neutron core, however, are so high that the neutrinos are trapped, and may aid in shoving the surrounding shells and envelope outward. Other ideas invoke violent convection within the neutron core to provide the energy. However it is done, the outflowing neutrinos carry away 100 times more energy than does the blast itself. The power output of a single supernova for a brief moment equals that of the rest of the stars in the visible universe. At the highest progenitor masses, supernovae are believed to eject their energy in bipolar jets, leading to longer-duration gamma-ray bursts (GRBs). The progeny of O stars, supernovae are rare, occurring in the Milky Way Galaxy at a rate of only one or two per century, so the odds of one being close to Earth are fortunately small. See GAMMA-RAY BURSTS; NEUTRINO; SUPERNOVA.

Supernova remnants and compact remains. The debris of the explosion, the supernova remnant, highly enriched in iron and other heavy elements, expands for centuries into space. Its shock wave energizes interstellar space, heats vast bubbles of gas to hundreds of thousands of degrees, and helps compress interstellar matter and create new stars. Its heavy elements become incorporated into molecular clouds. Most of the Earth came from previous generations of supernovae, each one contributing perhaps a mountain mass. Accelerated by the shock, ambient interstellar electrons, protons, and heavier nuclei approach the speed of light and speed through the Galaxy as cosmic rays, which in their turn also help initiate star formation (Fig. 9). See CRAB NEBULA; IN-TERSTELLAR MATTER; NEBULA; SHOCK WAVE.

Exposed by the explosion is the compact hot neutron star. Only 25 km (15 mi) across, it is stabilized by the pressure of neutron degeneracy. Conservation of angular momentum makes the little star spin dozens of times per second, its magnetic field compacted to



Fig. 9. Crab Nebula, the remnant of the supernova of 1054, as seen in radio radiation. (*National Radio Astronomy Observatories*)

10¹² or more times stronger than Earth's. Radiation beams out along a tilted, wobbling magnetic axis, and if Earth is in its path, a "pulse" of radiation is observed, the neutron star now a "pulsar." As the neutron star radiates away its energy, it spins more slowly and finally disappears from view. The most highly magnetized neutron stars, the magnetars, can have magnetic field strengths greater than 10¹⁴ times that of Earth, and are related to anomalous x-ray pulsars (AXPs) and to soft gamma-ray repeaters (SGRs), the latter occasionally releasing bursts so powerful that they can affect the Earth. The Galaxy must contain over 108 quiet, near-invisible neutron stars. Offcenter detonation in the supernovae that created them may give them kicks that can send them off at speeds far higher than those of most of the Galaxy's stars. See NEUTRON STAR; PULSAR.

Beyond the Chandrasekhar limit of $1.4 M_{\odot}$, electron degeneracy pressure cannot stabilize a white dwarf, and it must collapse. Neutron stars are close to this limit, and are themselves similarly limited to around 3 solar masses. Stars at the upper end of the main sequence are expected to create iron cores that exceed even this limiting mass. The dense remains cannot stabilize, and they too must therefore collapse. When such a collapsing star passes a critical radius at which the escape velocity is about that of light, it become invisible, and a black hole is born. About half a dozen black hole candidates are recognized in binary systems in which the black hole affects the companion. *See* BLACK HOLE.

Double-star evolution. Double (binary) stars are a common fact of stellar life. Within wide margins, about half the Galaxy's stars have companions. If the components of a binary are well separated, by tens or hundreds of astronomical units, the stars evolve separately as if each were single. If the stars are close enough together, however (the critical limits dependent on mass), they can profoundly affect each other and their courses of evolution.

A binary is first made of a pair of dwarf stars. The more massive evolves first into a giant or even a supergiant. If the two are close enough to start with, the growing giant can be affected by the smaller star. As it swells, it approaches a teardrop-shaped tidal surface in which the gravitational pulls from the two stars (including the forces resulting from orbital motion) are effectively equal. If the giant or supergiant actually reaches the zero-gravity surface, mass can flow through the point of the teardrop toward the smaller star. If the two still have sufficient separation, the incoming matter will first flow into a disk around the dwarf, from which the dwarf will accrete mass. If the two are close enough, matter will impact the dwarf directly without forming a disk. In either case, the result is that mass will be transferred, resulting both in interesting spectral activity and in the more massive star eventually becoming the less massive. Among the most dramatic examples of this kind of behavior is the eclipsing binary Algol, in which a class B dwarf accretes mass from a K giant. Mass loss can also wrap the pair in a common envelope. That and angular momentum losses resulting from the close interaction as well as from interacting magnetic fields can draw the two closer together and perhaps even make them merge.

The giant eventually produces a planetary nebula and then becomes a white dwarf. Mass loss has left the white dwarf with almost no hydrogen, perhaps only a thin skin on its surface. If the two stars have been drawn close enough, the white dwarf can tidally stretch the remaining dwarf (usually a G, K, or M star) to the zero-gravity surface, which can then transfer mass, mostly hydrogen, back to the white dwarf (Fig. 10). Instabilities in the accretion disk, as well as ionization and illumination of the flowing stream by a hot spot created by interacting gases, make the pair a flickering cataclysmic variable (CV). When the fresh hydrogen layer is thick, compressed, and hot enough, it explosively burns via the carbon cycle, the thermonuclear runaway creating a nova that brightens to absolute visual magnitude $M_V =$ -10 or so. Several are seen each year, one of first magnitude every few decades. Since a nova is only a surface phenomenon and does not much affect the stars of the binary, the system settles back into its prior CV behavior, and after 10⁵ years may produce



Fig. 10. A star fills its tidal surface and passes mass onto a white dwarf in preparation for a nova or even a type la supernova. (From J. B. Kaler, The Little Book of Stars, Copernicus Books, New York, 2000)

a nova explosion again. *See* CATACLYSMIC VARIABLE; NOVA.

In wider dwarf-white dwarf combinations, the ordinary dwarf will eventually begin its own evolution and become a giant. If it had received mass from its companion when it was a dwarf (perhaps through a wind), the giant may have been contaminated with by-products of nuclear fusion and appear as a barium star, anomalously rich in barium and other elements. As the giant grows, it passes its mass to the white dwarf through its vigorous wind or via an accretion disk. The hot white dwarf, or a hot boundary layer between an accretion disk and the white dwarf's surface, now gives the binary dual characteristics, those of both a hot and a cool star at the same time. These symbiotic stars can vary dramatically, and the transferred hydrogen can erupt into long-term fusion reactions. See SYMBIOTIC STAR.

White dwarf supernovae. There are two quite different kinds of supernovae. Type II supernovae have hydrogen in their spectra. They occur in galactic disks, where young high-mass stars reside, and are produced by iron-core collapse. Type I events, however, do not exhibit spectral hydrogen. They are divided into three categories: Type Ia supernovae display an absorption line of ionized silicon, whereas type Ib exhibits helium instead of silicon, and Ic has neither (or weak helium). While Ib and Ic supernovae are also confined to the disks of host galaxies, Ia supernovae occur in both disks and galaxy haloes where there are no high-mass stars. Tycho's Star of 1572 and possibly Kepler's Star of 1604 were type Ia (though some argue that the latter was type II).

The Ib variety is probably produced by core collapse in massive hydrogen-poor Wolf-Rayet stars. White dwarfs in double systems are obvious candidates for type Ia. The more massive the receiving white dwarf, the more quickly the infalling fresh hydrogen reaches the nova flash point. If massive enough, the interval between successive novae becomes less than the time over which they have been observed, and the novae become recurrent. For example, RS Ophiuchi and its cohort produce novae every 20 years or so. Each nova leaves a little residual matter, so the white dwarf increases in mass. If it is pushed over the Chandrasekhar limit, the star must catastrophically collapse and burn, producing a type Ia supernova. A second scenario that might produce type Ia supernovae is the merger of the stars in a double white dwarf system in which the sum of the masses exceeds the Chandrasekhar limit. Type Ia supernovae are even brighter than core-collapse type II events, and can reach $M_V = -19$. They also deposit more iron into the cosmos than do the type II, typically three-tenths of a solar mass. All the iron in the universe is made by supernovae. See WOLF-RAYET STAR.

Though the origins of type Ia explosions are not very well understood, the supernovae have such predictable absolute magnitudes that they can be used to measure distances to distant galaxies and determine not just the Hubble constant but also more subtle expansion characteristics of the universe.

Heavy-element enrichment. Many of the above scenarios are still shrouded in mystery. Not all the doubles have all the behavior patterns described, and there are great numbers of categories of each kind of phenomenon. The overall result of stellar evolution is clear, however. As stars become giants that turn into white dwarfs, or as they become supergiants that create neutron stars or black holes, they feed huge quantities of enriched matter back into the star-generating clouds of interstellar space. Later generations of stars therefore have more heavy elements than do earlier generations. Computer modeling of the chemical composition of the Galaxy, starting with the hydrogen and helium of the big bang, can closely replicate the chemical composition of the Sun, providing powerful evidence that the theories are correct. The Earth, made almost entirely of heavy elements, is a distillate of solar gases, and exists only as a result of the combined action of the James B. Kaler stars.

Bibliography. L. H. Aller, *Atoms, Stars, and Nebulae*, 3d ed., Cambridge University Press, 1991; E. Böhm-Vitense, *Introduction to Stellar Astrophysics*, 3 vols., Cambridge University Press, 1989, 1992; J. B. Kaler, *Stars*, Freeman, New York, 1992, paperback, revised 1998; J. B. Kaler, *The Cambridge Encyclopedia of Stars*, Cambridge University Press, 2006; R. J. Tayler, *The Stars: Their Structure and Evolution*, 2d ed., Cambridge University Press, 1994.

Stellar magnetic field

A magnetic field, far stronger than the Earth's magnetic field, which is possessed by many stars. Magnetic fields are important throughout the life cycle of a star. Initially, magnetic fields regulate how quickly interstellar clouds collapse into protostars. Later in the star formation process, circumstellar disk material flows along magnetic field lines, either accreting onto the star or flowing rapidly out along the rotation axis. Outflowing material (stellar winds) carries away angular momentum, slowing rotation at a rate that depends on stellar magnetic field strength. On the Sun, dark sunspots, prominences, flares, and other forms of surface activity are seen in regions where there are strong magnetic fields. There is some evidence that long-term variations in solar activity may affect the Earth's climate. Turbulence in the solar atmosphere drives magnetic waves which heat a tenuous corona (seen during eclipses) to millions of degrees. About 10% of hotter stars (with temperatures of about 10,000 K) with stable atmospheres are Ap stars, which have stronger magnetic fields that control the surface distribution of exotic elements. Even after stars end their internal fusion cycle and become compact remnants, magnetic fields channel accreting material from binary companions, occasionally producing spectacular novae. Despite the enormous gravity around pulsars, magnetic forces far exceed gravitational forces, creating intense electromagnetic beams that spin down (slow the rotation of) the pulsar. See BINARY STAR; CATACLYSMIC

VARIABLE; NOVA; PROTOSTAR; PULSAR; STELLAR EVO-LUTION.

Measurement techniques. Stellar magnetic fields are measured remotely using optical, infrared, and radio spectra. Magnetic fields alter atomic and molecular energy levels via the Zeeman effect, causing spectral lines to split and shift relative to their nonmagnetic locations. The size of the shift depends directly on wavelength and magnetic field strength. When the splitting is large compared to line width (as in sunspots, some Ap stars, and magnetic white dwarfs), the location and relative strengths of individual components yield both field strength and orientation. When the splitting is too small to resolve, measurements of spectral line polarization (polarimetry) can sometimes put limits on magnetic field strength, but only if the field geometry is simple (as in Ap stars). For distant stars that resemble the Sun, the polarization signal is small because the surface as a whole contains nearly equal amounts of oppositely directed magnetic field. In this case, the only recourse is to look for subtle enhancements of spectral line width or strength due to the Zeeman effect. See ASTRONOMICAL SPECTROSCOPY; POLARIMETRY; ZEE-MAN EFFECT.

Observations. Broadening of spectral lines in sunspots was first observed by J. Norman Lockyer in 1866, but the magnetic origin of this broadening was not understood until after the theoretical work of Pieter Zeeman in 1897 and the first polarimetry by George Ellery Hale in 1908. Field strengths are typically 0.3 tesla (3 kilogauss) in sunspots. Zeeman broadening implies similar field strengths for several other cool stars, but current stellar measurement techniques are effective only if at least 5% of the surface contains strong magnetic fields (compared to 0.5% for the Sun). For cool stars, polarimetry is useful only if specific areas with uniform magnetic field orientation can be localized (through spatial maps of the Sun, Doppler images of rapidly rotating stars, and emission from flares or accretion sites). Ap stars have magnetic fields as strong as 3.3 T (33 kG), organized into global dipoles that are often offset from the center of the star and tilted with respect to the rotation axis. For reference, the Earth has a dipole magnetic field that is only 6×10^{-5} T (0.6 G) at the pole. Finally, dense stellar remnants can have tremendous fields, reaching up to 10^4 T (10^8 G) for white dwarfs and 108 T (1012 G) for neutron stars. See GEO-MAGNETISM; NEUTRON STAR; SOLAR MAGNETIC FIELD; WHITE DWARF STAR.

Origin and dynamics. Large magnetic structures on the Sun vary in location, number, and magnetic polarity with an average cycle period of 22 years. Similar cycles have been detected on several other cool stars with turbulent (convective) atmospheres. In these stars, magnetic heating of coronae increases with rotation rate. Such behavior is expected of hydromagnetic dynamos, but details are sketchy and more than one type of dynamo may operate in the Sun. In warmer or more compact stars with convectively stable envelopes, magnetic fields are assumed to be concentrated fossils from earlier stages of evolution. Ultimately, the magnetic fields in these stable stars are attributed to the original interstellar gas out of which the stars formed. *See* MAGNETOHYDRODY-NAMICS; STAR; SUN. Jeff A. Valenti

Bibliography. G. Chanmugam, Magnetic fields of degenerate stars, *Annu. Rev. Astron. Astrophys.*, 30:143–184, 1992; I. Rüedi et al., Magnetic field measurements on moderately active cool dwarfs, *Astron. Astrophys.*, 318:429–442, 1997; J. C. del Toro Iniesta, On the discovery of the Zeeman effect on the Sun and in the laboratory, *Vistas Astron.*, 40:241–256, 1996.

Stellar population

One of the categories into which stars may be classified, based on their based on their spatial distribution and dynamics within a galaxy and their epoch of formation. The stellar component of the Milky Way Galaxy consists of three populations: the thin disk, the thick disk, and the halo. The thin disk, originally referred to as population I, is the youngest, and includes stars currently forming as well as older stars such as the Sun. It is located amid most of the molecular and cold atomic gas, and confined to a height of order 1 kiloparsec $(3.3 \times 10^3 \text{ light-years})$ or 3.1×10^{16} km or 1.9×10^{16} mi) from the galactic plane. The bulge and halo, roughly corresponding to the original population II, is a much older, far more extended structure, having an approximately spherical distribution with a scale length of order 3.5 kpc.

Development of concept. The existence of different populations in the stellar component of external galaxies was announced by W. Baade in 1944. During the wartime blackouts of Los Angeles, he was able to take the first deep multiwavelength photographs of the Andromeda Galaxy, M31. He found that the blue stars are confined to the disk of that galaxy, while the redder stars whose colors are more like the Sun or the globular clusters are more concentrated toward the galactic center and nuclear bulge. Baade referred to these stellar components as populations and labeled the younger blue disk stars population I. Those in the halo and nuclear spheroid, identified with the RR Lyrae variables (also known as cluster variables because of their preponderance in globular clusters), were called population II. A hypothetical population III has also been proposed to represent the first stars formed during the collapse of a galaxy. Some evidence for these stars is found in the extremely high abundances of r-process elements found in stars in which the overall abundance of heavy elements (elements other than hydrogen and helium) is less than 10⁻⁴ that of the Sun. Further kinematic studies of space motions demonstrated the distinctness of the two principal populations and accounted for their different distributions but left open the interpretation of these effects. See ANDROMEDA GALAXY; NUCLEOSYNTHESIS; VARIABLE STAR.

The population concept is now central to the understanding of chemical evolution of the Milky Way Galaxy. The globular clusters, the prototypical old stellar objects, have main-sequence turn-off ages of



The metallicity distribution of stars as a function of velocity V perpendicular to the galactic plane. (Negative values of V indicate motion toward the plane.) (a) $-30 \text{ km/s} \le V \le -10 \text{ km/s}$. (b) $-60 \text{ km/s} \le V \le -30 \text{ km/s}$. (c) $-100 \text{ km/s} \le V \le -60 \text{ km/s}$. (d) $-150 \text{ km/s} \le V \le -100 \text{ km/s}$. (e) $V \le -150 \text{ km/s}$. (After B.W. Carney, D. W. Latham, and J. R. Laird, A survey of proper-motion stars, VIII. On the galaxy's third population, Astron. J., 97:423-430, 1989)

at least 1010 years and are among the most metaldeficient stars in the Milky Way Galaxy. These are widely distributed throughout the system and serve as markers in the halo to the mass distribution. On the other hand, the OB associations contain many massive stars with ages not greater than 10⁷ years and have the highest metal abundances, often as much as a factor of 2 higher than the Sun. (At an age of about 4.6×10^9 years, the Sun was formed as a thin disk star as well.) The subdwarfs, the metal-deficient mainsequence stars, are a systematically older group of stars than those found confined to the thick galactic disk. The illustration shows how the metallicity distribution of stars decreases with increasing magnitude of velocity perpendicular to the galactic plane. Here, the metallicity [m/H] is the logarithmic difference in the overall abundance of heavy elements with respect to the Sun. For example, in a star with [m/H] = 0, this abundance is equal to that of the Sun, while in a star with [m/H] = -3 it is 10^{-3} that of the Sun. See STAR CLUSTERS.

Populations in galaxies. Fundamentally, the presence of these distinct stellar populations demonstrates that the Milky Way Galaxy has undergone an extended period of active star formation. The spheroid population is quite similar to elliptical galaxies, which are assumed to have ended active star formation many billions of years ago. Starburst galax-

ies give a counterexample of the fixity of the population distinction, however, since they may have completed their primary star-forming phase long ago, but the injection of a new mass of gas through collisions or accretion in a large-scale flow can produce a dramatic turn-on of the star formation. There is also increasing evidence that the Galactic populations may be contaminated by accretion of other, lower-mass satellite galaxies within the Local Group whose star-forming histories differed from that of the Milky Way. *See* GALAXY, EXTERNAL; GALAXY FORMA-TION AND EVOLUTION; STARBURST GALAXY.

Characteristics of populations. The thin disk is the youngest component of the Galactic stellar population. Still actively forming massive stars from molecular clouds, it is confined to within about 0.35 kpc of the plane. All of the stars have metallicities lying between about one-fifth and twice the solar value (illus. a), similar to that of the interstellar medium diffuse gas, and star formation appears to have remained relatively constant in this population for about the past 8×10^9 years, although there have been some excursions of a factor of order 50% during that time. One reason for the relatively small thickness of the disk is the low velocity dispersion of the component stars, about 22 km/s (14 mi/s); their systematic motion is completely dominated by the differential rotation of the disk. Since the most massive stars are also the youngest, having the shortest nuclear lifetimes, these hot stars are found associated with H II regions and OB associations as well as in open clusters. Massive stars are uniquely formed in the thin disk, supporting the attribution of youth to the population. Finally, type II supernovae, the death throes of massive stars, show a thin-disk distribution throughout the Milky Way Galaxy, and there is an indication that the remnants of massive stars, the pulsars (neutron stars with strong magnetic fields), are also confined to the thin disk. See INTERSTELLAR MATTER; MOLECU-LAR CLOUD; PULSAR; SUPERNOVA.

The thick disk is an older population, approximately $9-10 \times 10^9$ years, roughly corresponding to the range between what was once called population II and population I. Its metallicity lies between about one-tenth and one-third of the solar value. The stars in this population are distributed over greater distances from the plane, up to 1.5 kpc, and have correspondingly larger velocity dispersion, about 45 km/s (28 mi/s). This population also includes globular clusters and subdwarfs that overlap at the lowest end of the abundances with the properties of the halo globulars, although the system of old disk globulars is distributed differently than those of the halo. Type I supernovae are associated with this population. In the Milky Way Galaxy the spatial structure of this population roughly resembles that of an E7 elliptical galaxy.

Finally, enclosing the disk and the nuclear spheroidal bulge, there is a halo that extends to considerable distances from the plane, with some stars as distant as 30 kpc. This population has an age of order $10-15 \times 10^9$ years and a scale height of order 3.5 kpc or greater. The stars in this region have very large velocity dispersions, about 130 km/s (80 mi/s),

and do not appear to participate in the differential rotation as much as other stars. Their metallicities are all lower than about one-twentieth that of the Sun and may extend down to 10^{-3} of the solar value (illus. *e*). The most metal-poor globular clusters belong to this population. This stellar halo is not the same as the so-called dark matter halo, but is probably embedded within it. *See* DARK MATTER.

Delineation of populations. The population paradigm has become progressively more blurred with time, as the complex history of Galactic star formation has become clearer, and there has been considerable debate over whether any of these populations are truly distinct. Indeed, the separation based on elemental abundances renders this question somewhat circular since these abundances are already embedded in the definition of the membership in the different populations. Nonetheless, there are some distinct differences in these populations that permit their delineation. For instance, the globular clusters, although separated into halo and disk systems, are distinctly older than any of the clusters found in the disk. The oldest disk open clusters, NGC 188 and M69, show abundances nearly as low as the most metal-rich globulars like 47 Tucanae, but are younger by nearly 3×10^9 years than the youngest of the disk globulars. The most metal-deficient globulars are considerably older than any of the stars in the disk and much lower in abundance. See MILKY WAY GALAXY; STAR. Steven N. Shore

Bibliography. R. Cayrel et al., First stars V: Abundance patterns from C to Zn and supernova yields in the early Galaxy, *Astron. Astrophys.*, 416:1117-1138, 2004; K. Freeman and J. Bland-Hawthorn, The new Galaxy: Signatures of its formation, *Annu. Rev. Astron. Astrophys.*, 40:487-537, 2002; A. Helmi et al., Pieces of the puzzle: Ancient substructure in the galactic disc, *Mon. Not. Roy. Astron. Soc.*, 365:1309-1323, 2006; B. Nordström et al., The Geneva-Copenhagen survey of the solar neighbourhood: Ages, metallicities, and kinematic prpperties of ~14,000 F and G dwarfs, *Astrom. Astrophys.*, 418:989-1019; A. Renzine, Stellar population diagonostics of elliptical galaxy formation, *Annu. Rev. Astron. Astrophys.*, 44:141-192, 2006.

Stellar rotation

The spinning of stars, due to their angular momentum. Stars do not necessarily rotate as solid bodies, and their angular momentum may be distributed nonuniformly, depending on radius or latitude. However, it is nearly impossible to resolve the surfaces of stars, or see their interiors, and the limited ability to observe them means that their rotation is generally expressed as a single number, $v_{eq} \sin i$. In this measured quantity, v_{eq} is the star's equatorial rotational velocity (in kilometers per second), and *i* is the angle between the star's rotation axis and the line of sight to the star. In other words, $v_{eq} \sin i$ is the component of a star's rotation that is projected onto the line of sight between the star and the observer. Measurements of $v_{eq} \sin i$ in stars range from as little as 1 km s⁻¹ (0.6 mi s⁻¹) up to 400 km s⁻¹ (250 mi s⁻¹) or more. A more physically useful measure of rotation is Ω , a star's angular velocity, or $P_{\rm rot}$, the rotation period (the inverse of Ω). In some cases it is possible to measure $P_{\rm rot}$ directly. *See* ANGULAR MOMENTUM; STAR.

A $v_{eq} \sin i$ value is determined from the breadth of absorption lines in the star's spectrum. The Doppler effect causes the lines to be broadened because one limb of the star is receding and the other approaching. As a reference, the Sun is a very slowly rotating star ($v_{eq} \sin i = 1.8 \text{ km s}^{-1}$ or 1.1 mi s⁻¹), and solar absorption lines have a full width at half maximum (FWHM) of about 7 km s⁻¹ (4 mi s⁻¹), which arises from thermal motions, the physics of line formation, and turbulent motions in the Sun's atmosphere. This means that if a star's rotation rate exceeds 15-20 km s^{-1} (9-12 mi s^{-1}), then rotation dominates the broadening of the lines and other effects need not be considered, in general. To determine smaller values of $v_{eq} \sin i$ reliably, it is necessary to observe absorption lines carefully and to then fit them with models that take into account all the broadening mechanisms. By doing this it is possible to measure $v_{eq} \sin i$ down to the Sun's level and below. See DOPPLER EFFECT; SUN.

Late-type stars (cooler than about 6500 K or 11,200°F) exhibit spots on their surfaces analogous to sunspots, and these can be large enough to produce observable changes in the light of the star. In such cases it is possible to measure $P_{\rm rot}$ directly and, in a few instances, changes in $P_{\rm rot}$ have led to estimates of differential rotation (the dependence of Ω on latitude). Differential rotation is critical for the mechanisms that generate magnetic fields in lowmass stars. Rotation periods are independent of aspect and so provide a true indication of the intrinsic spread of rotation within star clusters, for instance. The high quality of $P_{\rm rot}$ measurements means that more is now known about the rotation in low-mass stars than is known for massive stars, the opposite of the case a generation ago. See STAR CLUSTERS; STEL-LAR MAGNETIC FIELD.

Stars get their angular momentum from the material from which they form, and one of the key problems in star formation is understanding how a cloud of material in the interstellar medium can contract enough to become a star. This is because a tiny degree of spin in the natal cloud leads to huge rotation rates by the time the material is dense enough to be a star. Stars appear to regulate their rotation in these early phases by interacting with the disk of material that surrounds them (the debris left over from star formation) and through jets, although this is an active area in current research. Stars are observed to have their rotation axes distributed randomly, which means that they do not acquire their rotation from the rotation of the Milky Way Galaxy but from turbulent motions in the parent cloud. This is confirmed by noting that the angular momentum axes of binary stars are also oriented randomly and that there is no tendency for stars in clusters to have rotation axes that align with the Galaxy's. See BINARY STAR; PRO-TOSTAR; STELLAR EVOLUTION.

Each type of star has its own rotational characteristics, and there is a broad trend in which typical rotation rates decline with the mass of the star in going down the main sequence. There is an especially prominent break in rotation such that stars with masses below about 1.2 times the solar mass have very low rotation rates, due to their intrinsic structure, as explained below.

Early-type stars. The early-type stars (types O, B, A, and early-F) typically have $v_{eq} \sin i$ values of 50 km s⁻¹ (30 mi s⁻¹) or more, and so their rotation is easily discerned on moderate-resolution spectrograms. For this reason, and because the early-type stars were easier to study with photographic plates, they were the first to have their rotational properties examined in detail, and most of the extant data dates from the photographic era. The most massive stars (of types O and B) sometimes exhibit nonradial pulsations and strong winds. They also can have surface abundance anomalies that result from internal mixing. All these effects are related to rotation, and they can complicate the interpretation of the observations in very massive stars. At the same time, rotation rates in O and B stars can approach the breakup velocity, reducing the effective gravity in the star and thereby affecting the star's evolution. See ASTRONOM-ICAL SPECTROSCOPY; SPECTRAL TYPE.

For the O stars, the distribution of $\nu_{eq} \sin i$ has two peaks, centered at about 100 and 300 km s⁻¹ (60 and 190 mi s⁻¹). The highest velocities seen are near 400 km s⁻¹ (250 mi s⁻¹), and the high $\nu_{eq} \sin i$ values (200 km s⁻¹ or 125 mi s⁻¹ and more) are seen only in the main-sequence O stars. These very high rates extend into the Be stars. O stars can also have significant turbulence (20 km s⁻¹ or 12 mi s⁻¹), and so there are none with sharp-lined spectra. O stars have strong radiation-driven winds and significant mass loss so that these stars can lose angular momentum during their main-sequence lifetimes.

The highest rotation rates can reduce the central pressure within a star; this can lead to slightly lower luminosity (by as much as 3%). The shape of the star can also change so that the surface temperature changes with latitude. A more significant effect is that rapid rotation influences mixing within the star and transports angular momentum. In particular, higher rotation leads to a higher degree of turbulent diffusion that can double a star's luminosity in some cases.

The B stars have an average $v_{eq} \sin i$ of about 140 km s⁻¹ (87 mi s⁻¹), but the Be stars (B stars with emission lines, comprising about 15–20% of the B stars) rotate at significantly higher rates (the average v_{eq} for Be stars, corrected for inclination, is 265 km s⁻¹ or 165 mi s⁻¹). Even this high rate is still about a factor of 2 below the breakup velocity (500–550 km s⁻¹ or 310–340 mi s⁻¹).

The A and early-F stars have moderate rotation rates (50-100 km s⁻¹ or 30-60 mi s⁻¹) except for those exhibiting peculiar chemical abundances. The Am, or metallic-line A stars, and the Ap, or magnetic A stars, rotate at an average rate of about 40 km s⁻¹ (25 mi s⁻¹), a factor of 3-4 lower than normal stars of similar colors. Only a very few chemically peculiar stars have $\nu_{eq} \sin i$ exceeding 100 km s⁻¹ (60 mi s⁻¹).

Also, in these peculiar stars rotation declines with age, but the relationship is neither well delineated nor understood. The opposite is also true: very few normal A stars have low $v_{eq} \sin i$ values.

Late-type stars. Late-type stars are cooler than about 6500 K (11,200°F) and have spectral types of late-F, G, K, or M. The key property of late-type stars that determines much of their phenomenology is the presence of a convective envelope. In such stars, rotation (especially differential rotation) interacts with the convection to produce complex motions in the electrically conductive plasma of the star. These circulatory patterns enable the star to regenerate a seed magnetic field, the so-called dynamo mechanism. The magnetic field can grip an ionized wind beyond the star's surface, leading to gradual angular momentum loss, a process that is seen occurring on the Sun. More rapidly rotating stars produce stronger magnetic fields, as evidenced by spectroscopic indicators of magnetic-related activity, such as emission in the cores of the Ca II H and K lines. There is, therefore, a strong correlation between rotation and activity, in the various forms that activity is observed (Ca II H and K, $H\alpha$, x-rays, and so forth). These activity indices are often observed to measure $P_{\rm rot}$ for stars because they vary more pronouncedly than the star's light overall. See MAGNETISM; STELLAR MAGNETIC FIELD

In this scenario, late-type main-sequence stars spin down over their lifetimes, and that is observed to occur because steadily declining rotation rates can be seen among stars in clusters of increasing age. Moreover, stars that are born together but with different rotation rates will tend to end up with the same $P_{\rm rot}$ because the faster-rotating star will lose angular momentum more quickly. Young stars tend to rotate rapidly because they have not had time to lose the angular momentum with which they formed, while old low-mass stars rotate slowly.

Young late-type stars are mostly observed in clusters because there are few of them elsewhere, and because clusters provide rich samples of the same age. The Pleiades, for example, is 100 million years old, meaning that its low-mass stars have only recently reached the zero-age main sequence (ZAMS). Among the solar-type stars of the Pleiades, an intrinsic spread in rotation of more than a factor of 20 is seen at any one mass, with the most rapidly rotating stars spinning at 100 times the solar rate. These ultrafast rotators are a minority constituent of the cluster (about 20% of the stars), but their presence along with many slowly rotating stars is difficult to account for in models of the angular momentum evolution of low-mass stars. It has been speculated that some of these young stars may experience a decoupling of the radiative core from the convective envelope, with the core continuing to spin at a high rate for an extended time. The α Persei cluster is also at the zero-age main sequence, with an age of 50 million years, and its stars confirm that what is seen in the Pleiades is typical for young low-mass stars. By the time that stars become as old as the Hyades (age 600 million years), their rotation rates have largely converged and show little spread at any

one mass, although there is a strong mass dependence. *See* HYADES; PLEIADES.

The most rapidly rotating low-mass stars seen outside clusters spin at about 10 times the solar rate, but the vast majority, being old like the Sun, rotate very slowly (at the solar level or below). Evolved late-type stars (giants and supergiants) rotate at very low rates, in part because they have swelled so much since leaving the main sequence. A few rare exceptions spin rapidly, probably because they acquired the orbital angular momentum of their companion when a binary system coalesced. *See* GIANT STAR; SUPERGIANT STAR.

Very low mass stars and brown dwarfs. Significant numbers of very low mass stars and brown dwarfs have been found and studied at high spectroscopic resolution. These objects are fully convective from their surfaces all the way to their cores, in contrast to the Sun and similar stars, which have a convective envelope. This change in structure corresponds to a change in the observed rotation in that very low mass stars and brown dwarfs appear not to lose angular momentum and so continue to spin rapidly. These objects are very small in size, so the rotation rates seen—15 to 60 km s⁻¹ (10 to 40 mi s⁻¹)—imply high angular velocities of as much as 100 times that of the Sun. *See* BROWN DWARF.

Presence of planets. There is no relationship between the rotation of a star and the presence of planets around it. It is true that planets have been found primarily around old, slowly rotating stars like the Sun, but that is in part because such stars have the very sharp absorption lines needed to measure the reflex motion to the necessary high precision. Young solar-type stars have broader lines that prevent reaching the necessary precision, and they also exhibit effects related to their activity that add noise to the velocity measurements. *See* EXTRASOLAR PLAN-ETS.

Solar rotational history. The rotational history of the Sun is not known in detail, but the broad features of that history can be delineated from studying stars of solar mass of many ages. The Sun formed from interstellar material, most probably in concert with many other stars at the same time. For about the first 10 million years of its life, it kept a disk of debris material around it, and this formed into planets before about 50 million years was reached. At an age of 10 million years the Sun rotated at two or three times its present angular velocity, and it was well above the main sequence. The Sun reached the zero-age main sequence when it was about 50 million years old, at which point it could have rotated at a rate of 3 to 100 times the current solar rate, although it probably was at the low end of this range. Over its 4.5-billion-year main-sequence lifetime, the Sun has gradually lost angular momentum to get to the rate that is seen today. Helioseismological studies of the Sun's interior show it to rotate as a solid body, so if it ever had a rapidly rotating core, it does no longer. In about 5 billion years the Sun will exhaust most of the hydrogen in its core and leave the main sequence to become a giant star. It will bloat to 20 or more times its present size, and conservation of

angular momentum will mean that its rotation rate will become extremely small.

The Sun remains the standard of comparison for other stars. The rotation of the Sun was among Galileo's first discoveries as he traced the motions of spots across the surface, and only a little later the different rates of rotation at different solar latitudes led to the fundamental notion that the Sun is a fluid, not a glowing solid. Angular momentum is central to the formation and evolution of stars, and work now taking place will lead to a fuller picture of what happens in stars and how the Sun fits into that picture. David Soderblom

Bibliography. D. F. Gray, *The Observation and Analysis of Stellar Photospheres*, 3d ed., 2005; A. Maeder and P. Eenens (eds.), *Stellar Rotation*, IAU Symp. 215, Astronomical Society of the Pacific, San Francisco, 2004 ; R. Pallavicini, G. Michela, and S. Sciortino (eds.), *Stellar Clusters and Associations: Convection, Rotation, and Dynamos*, Astronomical Society of the Pacific, San Francisco, 2000; J.-L. Tassoul, *Stellar Rotation*, Cambridge University Press, 2000.

Stem

The organ of vascular plants that usually develops branches and bears leaves and flowers. On woody stems a branch that is the current season's growth from a bud is called a twig. The stems of some species produce adventitious roots. *See* ROOT (BOTANY).

General Characteristics

While most stems are erect, such as aerial structures (**Fig.** 1*a*), some remain underground (**Fig.** 2*a–b*), others creep over or lie prostrate on the surface of the ground (Fig. 1*c–e*), and still others are so short and inconspicuous that the plants are said to be stemless, or acaulescent (Fig. 1*b*). When stems lie flattened immediately above but not on the ground, with tips curved upward, they are said to be decumbent, as in juniper (Fig. 1*c*). If stems lie flat on the ground but do not root at the nodes (joints), the stem



Fig. 1. Kinds of stems. (a) Woody (plum). (b) Acaulescent (evening primrose). (c) Decumbent (juniper). (d) Procumbent (purslane). (e) Repent (ground ivy).



Fig. 2. Specialized stems. (a) Rhizome (Solomon's seal).
(b) Rhizome (grass). (c) Moniliform rhizome (pepperroot).
(c) Caudex (trillium). (e) Tuber (potato). (f) Corm (crocus).
(g) Section of corm. (h) Tunicated bulb (onion). (i) Scaly bulb (lily). (j) Scape (English plantain). (k) Thorn (hawthorn).
(h) Tendril (grape). (m) Cladode (asparagus).

is called procumbent or prostrate, as in purslane (Fig. 1*d*). If a stem creeps along the ground, rooting at the nodes, it is said to be repent or creeping, as in ground ivy (Fig. 1*e*).

Shape and texture. Most stems are cylindrical and tapering (terete), appearing circular in cross section; others may be quadrangular, as in mints, or triangular, as in some sedges.

Herbaceous stems (annuals and herbaceous perennials) die to the ground after blooming or at the end of the growing season. They usually contain little woody tissue. Woody stems (perennials) have considerable woody supporting tissue and live from year to year. A woody plant with no main stem or trunk, but usually with several stems developed from a common base at or near the ground, is known as a shrub. Suffrutescent stems are intermediate between herbaceous and shrubby and become partly woody and perennial at the base, as in teaberry.

Specialized stems. The stems of some plants are highly modified in various ways (Fig. 2). An underground horizontal stem is a rhizome or rootstock. It may be thickened, as in Solomon's seal; slender, as in Bermuda grass; or contracted at regular intervals, as in pepperroot, in which case it is a moniliform rhizome. A caudex is an upright perennial underground stem; it is much like a rhizome but grows vertically, as in trillium. The term caudex is applied also to the trunks of palm trees which grow in a similar manner but aboveground. Tubers are enlarged ends of rhizomes in which food accumulates, as in white potato. A corm is a short, erect, fleshy subterranean stem usually broader than high and often covered with dry membranous coats, as in crocus. A

bulb may be regarded as a short, subterranean stem with many overlapping fleshy leaf bases or scales, as in onion. A scape is a leafless flowering stem arising from the ground, as in plantain or dandelion. A thorn is a rigid, sharp-pointed modified branch of a woody plant, as in hawthorn. A tendril is a slender coiling structure capable of twining about a support to which the plant is then attached, as in grape. A cladophyll, or phylloclade, is a usually flattened stem resembling a leaf that arises in the axil of a minute leaf (scale), as in asparagus. In some instances the resemblance of a cladophyll to a leaf is noted only or mainly in its green color, as in certain species of cacti. *See* PLANT ORGANS. Nelle Ammons

External Features

A shoot or branch usually consists of a stem, or axis, and leafy appendages. Stems have several distinguishing features. They arise either from the epicotyl of the embryo in a seed or from buds. The stem bears both leaves and buds at nodes, which are separated by leafless regions or internodes, and sometimes roots and flowers. A woody twig (**Fig. 3**) shows leaf scars and lenticels on its surface. Eventually, secondary growth thickens the stem axis and obscures its division into nodes and internodes and the relation between buds and leaves. *See* LEAF.

Nodes. The nodes are the regions of the primary stem where leaves and buds arise. The number of leaves at a node is usually specific for each plant species. In deciduous plants which are leafless during winter, the place of former attachment of a leaf is marked by the leaf scar. The scar is formed in part by the abscission zone formed at the base of the leaf petiole. Small bundle scars left by the broken ends of the vascular (conducting) tissues between the stem and leaf are seen within the leaf scar. Flower and fruit scars, marking the points of flower and fruit attachment, may also be visible.

Internodes. The stem regions between nodes are called internodes. Internode length varies greatly among species, in different parts of the same stem, and under different growing conditions. One or many internodes may grow rapidly during stem elongation. Generally, an individual internode matures



Fig. 3. Winter woody twig (horse chestnut) showing apical dominance. (After E. W. Sinnott and K. S. Wilson, Botany: Principles and Problems, 5th ed., McGraw-Hill, 1955)



Fig. 4. Portion of a cross section of a young stem of elder (Sambucus canadensis) showing a lenticel. (G. H. Conant, Triarch Products)

from the base upward in dicotyledons and many monocotyledons. Conversely, internodes of some monocotyledons (grasses and sedges) and horsetails (*Equisetum*) have prolonged growth from intercalary meristems located at the base of each internode.

Lenticels. There are small, slightly raised or ridged regions of the stem surface that are composed of loosely arranged masses of cells in the bark (**Fig. 4**). Their intercellular spaces are continuous with those in the interior of the stem, therefore permitting gas exchange similar to the stomata that are present before bark initiation. The size and shape of lenticels vary among species and with age they are absent in a few species. *See* PERIDERM.

Bud scales. In most temperate, perennial plants, buds are protected by a series of modified leaves or bud scales. Wax or hairs on their surface further protect the apical meristem within the bud. The buds of many tropical trees are often protected by the bases of mature leaves or simply by the expanding, immature leaves. Such naked buds are also present in most herbaceous plants.

Most temperate trees form a terminal bud at the end of a season's growth. The cluster of scars that mark the position of the closely packed terminal bud scales also delimit the segment of stem produced during one growing season (Fig. 3). The age and yearly growth of a twig can be determined by observing the number of sets of these terminal bud scale scars and the distances between them.

Bud types. Buds are partly developed shoots or undifferentiated masses of meristematic cells. They may be classified as terminal, lateral, and adventitious. *See* BUD.

Terminal or apical buds. These are buds at the tips of branches, often the largest buds on the plant, which usually contribute to the growth in length of the stem bearing them.

Lateral buds. These are buds located on the sides of the stem and are initiated within the terminal bud from superficial tissues at the nodes. They are further classified according to their position. Most plants have a bud in the axil of each leaf directly above the leaf (or leaf scar). Such axillary buds usually remain dormant for some time. If the axillary bud develops into a branch, it becomes organized into a new terminal bud which itself produces new lateral buds. In some species, additional buds are produced above or on either side of the axillary bud. Such accessory or supernumerary buds may eventually develop as flower buds, thorns, or vegetative branches like the axillary bud.

Adventitious buds. These are buds which are not produced within the terminal bud and lack a definite relation to leaf position. They arise from differentiated (mature) tissues of the root, stem, or leaf, or from wound callus. The root suckers of many trees (oak, sweet gum, aspen) and troublesome weeds arise from adventitious buds. Such buds are also produced on leaf cuttings of African violet and begonia.

Bud development. Most buds of a plant remain undeveloped for indefinite periods and are known as dormant, latent, or potential buds in contrast to active or developing buds.

In many trees and shrubs, especially those with large, overwintering buds, the whole year's growth is laid down in rudimentary form in the bud during early summer. These buds expand by elongation of internodes the following spring. Some buds, especially flower buds, may form in the spring of the same season during which they expand. In temperate trees most lateral buds remain dormant for the first season, and only grow into branches the following season because of the apical dominance of the terminal bud. Such delayed bud outgrowth is characteristic of proleptic branches. In contrast, many tropical trees have sylleptic branches, in which lateral buds grow out immediately without a rest period. *See* APICAL DOMINANCE; PLANT GROWTH.

Branching. There are three major types of stem branching: dichotomous, monopodial, and sympodial. Dichotomy occurs by a division of the apical meristem to form two new axes. Such branching without axillary buds occurs in some lower vascular plants and in a few angiosperms (some palms and cacti, and *Strelitzia*). If the terminal bud of an axis continues to grow and lateral buds grow out as branches, the branching is called monopodial. If the apical bud terminates growth in a flower or dies back and one or more axillary buds grow out, the branching is called sympodial. Often only one bud develops so that what appears to be a single axis is in fact composed of a series of lateral branches that are arranged in linear sequence.

In many woody plants there is a distinction between long and short shoots, which results from the degree of internodal elongation. Short shoots are commonly the site of reproduction, and the spur shoots of apple, because of their congested nodes, bear a tight cluster of leaves.

Stem form. The large and conspicuous stems of trees and shrubs assume a wide variety of distinctive forms.

Columnar stems are basically unbranched and form a terminal leaf cluster, as in palms, or lack obvious leaves, as in cacti.

Branching stems have been classified either as excurrent, when there is a central trunk and a conical



Fig. 5. Tree architecture. The idealized models of tree architecture can account for most types of branching and growth in woody plants. (After F. Hallé and R. A. A. Oldeman, An essay on the architecture and dynamics of growth of tropical trees, Penerbit Universiti Malaya, 1975)

leaf crown, as in firs and other conifers, or as decurrent (or deliquescent), when the trunk quickly divides up into many separate axes so that the crown lacks a central trunk, as in elm.

Another approach to the study of stem form is the concept of tree architecture which uses models that can describe most living and fossil trees and shrubs (**Fig. 5**). Each architectural model is defined by the orientation, growth sequence, and branching pattern of the trunk and lateral branches of shoot. In at least some trees (*Terminalia*), the observed branching pattern approaches the theoretical ideal for the optimum filling of space and exposure of leaf surfaces, important adaptive characteristics for stem form that have probably been evolutionarily selected for. *See* TREE. Jack B. Fisher

Internal Features

The stem is composed of the three fundamental tissue systems that are found also in all other plant organs: the dermal (skin) system, consisting of epidermis in young stems and periderm in older stems of many species; the vascular (conducting) system, consisting of xylem (water conduction) and phloem (food conduction); and the fundamental or ground tissue system, consisting of parenchyma and sclerenchyma tissues in which the vascular tissues are embedded. The arrangement of the vascular tissues varies in stems of different groups of plants, but frequently these tissues form a hollow cylinder enclosing a region of ground tissue called pith and separated from the dermal tissue by another region of ground tissue called cortex. *See* CORTEX (PLANT); EPI-DERMIS (PLANT); PARENCHYMA; PHLOEM; PITH; SCLE-RENCHYMA; XYLEM.

Primary and secondary tissues. Part of the growth of the stem results from the activity of the apical meristem located at the tip of the shoot. The derivatives of this meristem are the primary tissues: epidermis, primary vascular tissues, and the ground tissues of the cortex and pith. In many species, especially those having woody stems, secondary tissues are added to the primary. These tissues are derived from the lateral meristems, oriented parallel with the sides of the stem: cork cambium (phellogen), which gives rise to the secondary protective tissue periderm, which consists of phellem (cork), phellogen (cork cambium), and phelloderm (secondary

cortex) and which replaces the epidermis and vascular cambium, which is inserted between the primary xylem and pholem and forms secondary xylem (wood) and phloem (**Fig. 6**). *See* APICAL MERISTEM; LATERAL MERISTEM.

Stele. The vascular tissues and the closely associated ground tissues—pericycle (on the outer boundary of vascular region), interfascicular regions (medullary or pith rays), and frequently also the pith—may be treated as a unit called the stele. The variations in the arrangement of the vascular tissues serve as a basis for distinguishing the stelar types. The word stele means column and thus characterizes the system of vascular and associated ground tissues as a column. This column is enclosed within the cortex, which is not part of the stele. *See* PERICYCLE.

Primary State

The leafy stem of the plant is initiated as a shoot primordium in the embryo. This primordium is called the plumule, or epicotyl, and is located above the insertion of the cotyledons (seed leaves) on the hypocotyl. The epicotyl has an apical meristem that gives rise to all subsequent primary parts of the leafy stems. In the primary growth of the stem, nodes and internodes are differentiated, and the leaves appear in a characteristic arrangement (phyllotaxis). The cells derived from the apical meristem compose the meristematic precursors of the primary tissues of the stem and are called primary meristems (protoderm, ground meristem, procambium) [Fig. 6]. These meristems differentiate into the primary tissue systems.

Shoot apex or apical meristem. This so-called growing point consists of the meristematic initials and their derivatives. The organization of the apical meristem varies in different groups of plants. In the angiosperms the initials are usually arranged in two or more layers (tiers), frequently three in gymnosperms, a large group of apical initials without a tunica and with some cell stratification; and in the ferns and still lower vascular plants the shoot tip may bear a single apical initial. The organization in the angiosperms is referred to as that of tunica and corpus. The outermost tier or tiers of initials and their immediate derivatives constitute the tunica; the innermost tier and its immediate derivations compose the corpus. The tunica is a mantlelike region enclosing the corpus or core. The tunica increases only in surface by anticlinal (at right angles to the surface) divisions; the corpus shows volume growth, for its cells divide in various planes. All the initials and their first derivatives together are often designated as a promeristem. Below this promeristem are partially differentiated tissues, the primary meristems, named from the periphery inward: the protoderm; the outer ground meristem; the procambium, or provascular tissue; and the inner ground meristem. As stated previously, the primary peristems differentiate into primary tissue systems: the protoderm into the epidermis, the ground meristem into tissues of cortex and pith, and the procambium into primary vascular tissues (Fig. 6).



Fig. 6. Diagrammatic representation of a woody, dicotyledonous stem in transverse and longitudinal sections. It illustrates the gradual transition (arrow) from primary growth (growth in length) to secondary growth (growth in diameter). The primary plant body is derived from the apical meristem; the secondary plant body is produced by the vascular cambium and the cork cambium.

Differentiation. The complex of physiological and morphological changes that occur in a cell as it develops from a meristematic to a mature state is called differentiation. Common evidences of differentiation of cells are vacuolation (water uptake), accumulation of ergastic substances (metabolic products), formation of plastids, and increase in thickness of the cell wall and its impregnation with new wall substances. In highly specialized cells the nucleus may break down (sieve elements in the phloem), or even the entire protoplasm may be lost (water-conducting cells in the xylem). *See* PLANT CELL.

Primary tissue systems. As seen in the stem as a whole, the primary tissues form integrated systems and are referred to as the primary tissue systems: epidermis (or epidermal system), ground tissue system, and primary vascular system.

Epidermis. The epidermal system differentiates from the protoderm. The epidermis is usually composed of a single layer of cells. Occasionally, the protoderm divides periclinally (parallel to the surface) and produces a multiple epidermis. Epidermal cells may be variously modified into guard cells with stomata or the diverse epidermal appendages called hairs, or trichomes. The walls of the epidermis are cutinized, that is, impregnated with a waxy substance, and the outer walls are covered with a continuous layer of the same substance, the cuticle. *See* EPIDERMIS (PLANT).

Cortex. Between the epidermis and the vascular system is the cortex, formed by the outer ground tissue. However, there is no distinguishable cortex in plants in which the vascular bundles have a scattered arrangement, as in corn. The outer layers of the cortex frequently form collenchyma, a mechanical, or strengthening, tissue. The cortex may also include chlorenchyma (tissue containing chlorophyll). Still other cells may develop as sclereids, fibers, or laticifers. Secretory canals may be present. In water plants the cortical parenchyma may contain large air spaces and is then called aerenchyma. The outermost layers of the cortex are sometimes cutinized, like the epidermis, and are referred to as the hypodermis. Cortical parenchymal cells contain various inclusions such as starch, tannins, and crystals. See COLLENCHYMA; CORTEX (PLANT).

Endodermis. The innermost layer of the cortex (or possibly, in the lower vascular plants, the outermost layer of the stele) forms a histochemically distinct and sometimes a morphologically specialized layer, the endodermis. If an endodermis is morphologically distinct, it forms a compact layer and the anticlinal walls of its cells bear a suberized and lignified band, or Casparian strip. The endodermis is found in the stems and roots of lower vascular plants, the rhizomes and roots of seed plants, and a few herbaceous stems of angiosperms. It is not differentiated in the aerial stems of gymnosperms and most angiosperms. In young stems of angiosperms a homologous layer contains abundant starch and is called the starch sheath. *See* ENDODERMIS.

Pericycle. If a distinct layer or layers of cells are present between the ground-tissue endodermis and the vascular tissue, the tissue constitutes the pericycle. This tissue is interpreted as part of the ground tissue of the stele. The pericycle may consist of parenchyma or sclerenchyma. It may be seen in the stems and roots of lower vascular plants, but is usually absent in the stems of gymnosperms and angiosperms. The fibers located on the periphery of the vascular systems in many dicotyledons are often called pericycle fibers. In many species, however, these fibers originate in the phloem. *See* PERI-CYCLE.

Pith. The ground tissue enclosed by the vascular tissues is the medulla or pith. In stems with no cortex the pith is not distinguishable. The vascular bundles may be scattered throughout the central ground tissue, as in some monocotyledons, or there may be strands of internal phloem or complete vascular bundles located within the pith, as in some dicotyledons. The pith may consist of parenchyma or may become sclerotic (hard-walled), especially at the nodes. The outer zone of the pith often consists of small, persistent cells with more or less sclerified walls. Many herbaceous plants have hollow stems (internodes) because the growth of the pith fails to keep pace with stem expansion. Some woody plants, such as the walnut and Liriodendron, have a diaphragmed pith, that is, pith with thick-walled parenchyma cells and sclereids arranged in disks, or diaphragms, of firm tissue.

Primary vascular system. In the ferns and the seed plants the primary vascular system, derived from the procambium, is composed of anastomosing (interconnected) vascular bundles. At each node some of the vascular bundles diverge into the leaves and axillary branches. A bundle in the stem which leads directly into a leaf is called a leaf trace; one leading into a branch is a branch trace. One or more leaf traces may be associated with each leaf. Where the leaf trace diverges from the stem into a leaf at a node, the vascular cylinder commonly has a more or less circumscribed region of parenchyma. This region is called a leaf gap or lacuna. The lateral branches of most dicotyledons and gymnosperms have two branch traces. Some species have one trace per branch, others more than two. In monocotyledons, for example, the axillary shoot connection consists of many vascular strands. At the node the branch traces may have a single branch gap which, moreover, is commonly confluent with the leaf gap.

Separate parts of the primary vascular system of the stem constitute vascular strands or vascular bundles. The phloem and xylem within a vascular bundle may be arranged in various ways. The most common type in gymosperms and angiosperms is the collateral bundle in which the phloem occurs only on one side of the xylem. If phloem occurs on both sides of the bundle, as in the Cucurbitaceae, the bundle is bicollateral. If one kind of vascular tissue completely surrounds the other, the bundles are concentric. The concentric bundle is amphicribral if the phloem surrounds the xylem, as in ferns, or amphivasal if the xylem surrounds the phloem, as in some monocotyledons and dicotyledons.

The primary vascular system with associated fundamental tissue is termed the stele. The simplest type of stele, and perhaps also the most primitive, is the protostele, in which a solid central core of xylem is surrounded by phloem, as found generally in the roots, in the stems of lower vascular plants such as the fern Gleichenia, and the stems of some water plants (Fig. 7a). In the plant group Lycopsida the lycopods or "ground pines" have a modified protostele with alternating bands of xylem and phloem (Fig. 7b), and the "club mosses" (Selaginella) have separate protosteles supported in air spaces by radially elongated endodermal cells designated trabeculae. In the stems of vascular plants located higher than Lycopsida on the evolutionary scale, the center of the stele is occupied by a pith, and the vascular system has a tubular structure. The stele, therefore, is called a siphonostele (tubular stele). If the phloem occurs only on the outer side of the xylem of a siphonostele (gymnosperms and angiosperms), the stele is called an ectophloic siphonostele (Fig. 7c). If the phloem differentiates also on the inner side of the xylem (in such ferns as Adiantum, Dicksonia, and Marsilia), the stele is designated an amphiphloic siphonostele or a solenostele (Fig. 7d). In the ferns, in which the stem is short and the leaves are set close together, the leaf gaps overlap, and the tubular stele is dissected into a network of distinct strands. Such a stele is called dictyostele, as in Polypodium



Fig. 7. Types of steles as seen in photomicrographs of cross sections of stems. (a) Protostele (*Gleichenia*). (b) Radial (modified) protostele (*Lycopodium*). (c) Ectophloic siphonostele (*Tilia*). (d) Amphiphloic siphonostele (*Dicksonia*). (e) Dictyostele (*Polypodium*).

and *Dryopteris* (Fig. 7*e*). Another modification of the siphonostele is the eustele, in which the vascular system consists of collateral or bicollateral strands, and both interfascicular regions and leaf gaps are present. Eustele is characteristic of gymnosperms and angiosperms. The most complex stele, the atactostele, has a dispersed arrangement of strands, as in many monocotyledons.

Primary development. The origin of the stem from the apical meristem is closely associated with the development of leaves. At initiation the leaf primordium is merely a small protuberance on the flank of the shoot tip and cannot be delimited from the stem part to which it is attached. Only later, as growth continues, does the primordium become an organ distinguishable externally and internally from the stem. The stem itself shows at first no division into nodes and internodes. It appears like a complex of superimposed disks (future nodes), each bearing one or more leaf primordia, depending on leaf arrangement. Later, during elongation, the disklike zones become separated from one another by cell division and enlargement between the levels at which the leaf primordia are inserted. Thus, the internodal elongation is initiated, and the stem becomes differentiated into nodes and internodes. The elongation of internodes is the phenomenon that brings about the characteristic rapid extension of new shoots during the spring growth of trees and shrubs.

Internally, the differentiation of stem tissues is more or less closely correlated with the differentiation of leaves. In lower vascular plants with protosteles in their stems, as the lycopods, the vascular system differentiates beneath the apical meristem as a column of provascular tissue (procambium), and the leaf traces, which connect the small leaves with this column, differentiate later through the cortex. In its differentiation the provascular tissue becomes distinct from the surrounding ground tissue of the cortex because the cells of the latter show an increase in vacuolation and an overall enlargement, whereas the cells of the provascular tissue retain their dense cytoplasm and they increase in length but remain narrow.

In the ferns and seed plants, in which the leaves are more prominent parts of the shoot than in the lycopods and the vascular system of the stem consists of a cylinder of strands (siphonostele structure) connected with the leaves at regular intervals (at the nodes), the vascular system is initiated in relation to the leaves. Vacuolation of the differentiating cortex and pith delimits the future vascular region as a still highly meristematic cylinder of cells. Then within this cylinder, procambial strands differentiate beneath the emerging leaf primordia chiefly by longitudinal cell division and cell elongation; that is, the leaf traces are initiated. The cells between the bundles that do not become procambial cells differentiate as parenchyma cells of the interfascicular regions and the leaf gaps. If followed downward in the stem, the young leaf traces will be found joined with one another and with older traces. Thus, the procambial strands form an interconnected system similar to that of the mature primary vascular system, except that all its parts are still short. The procambial system elongates in coordination with the other tissues of the internodes by further division and elongation of cells. Eventually, primary phloem and primary xylem differentiate in the procambial strands, the pholem in continuity with that of the older parts of the stem (acropetal differentiation, that is, from the base of the stem toward its apex), the xylem first appearing at the bases of leaves then differentiating in two directions: downward (basipetally) toward a connection with the xylem of older stem regions and upward in the leaves. See LYCOPODIALES; MAGNOLIOPHYTA; POLYPODIALES.

Differentiation of lateral buds. The lateral or axillary buds arise more or less close to the apical meristem of the terminal bud, depending on the type of branching characteristic of the plant. The buds are initiated by cell division in the stem to produce a lens-shaped group of cells in the outer cortex near the axils of subtending leaves. The divisions occur in such a manner that eventually an apical meristem of the bud is organized. The bud meristem may develop immediately into a side shoot by producing leaves and increments of stem, or it may interrupt growth after variable degrees of initial development. The lateral bud may also develop into an inflorescence, or flower, by appropriate changes in growth pattern as compared with that of a vegetative shoot. The bud and parent stem establish a vascular connection through the bud traces, the differentiation of which is more or less similar to that of the leaf traces of the same plant.

Secondary State

The stems of gymnosperms, most dicotyledons, and some monocotyledons show an increase in thickness by secondary growth. A lateral meristem, the vascular cambium, produces secondary vascular tissues which constitute the secondary body. Also, a cork cambium, or phellogen, produces a secondary protective tissue called periderm, a tissue which replaces the primary protective tissue, the epidermis.

The vascular cambium originates from the procambial and certain parenchyma cells of the primary body. Procambial cells that remain meristematic after the stem completes its elongation and the primary vascular tissues become mature proceed to divide periclinally (parallel with the surface). These cells are designated as cambial initials. In addition, periclinal divisions are initiated by parenchymal cells between the vascular bundles; these cells also are cambial initials. Collectively, all these initials constitute the vascular cambium. The part of cambium



Fig. 8. Secondary tissue in the stem of sunflower (*Helianthus annuus*). (a) Transverse section of the entire stem showing the thicker-walled, dark, secondary xylem cells as they are added to the primary, more in the fascicular than in the interfascicular areas. (b) Enlarged view of a portion of a. (*From J. B. Hill, L. O. Overholts, and H. W. Popp, Botany: A Textbook for Colleges, 2d ed., McGraw-Hill,* 1950)

originating within the vascular bundles is the fascicular cambium, that arising in the parenchyma between vascular bundles, the interfascicular cambium (**Fig. 8**).

The arrangement of the secondary vascular tissues produced by the cambium varies in different plants. This tissue may appear as (1) a continuous cylinder; (2) individual strands, with secondary activity limited to the bundles (no interfascicular cambium is formed); (3) individual strands, with secondary vascular tissues produced in the fascicular regions and secondary ray parenchyma in the interfascicular regions; or (4) anomalous (atypical) arrangements characterized by uneven growth of xylem or phloem or by formation of successive cambia, one farther outward than the preceding.

The cork cambium or phellogen originates in the epidermis or in the cortex from parenchyma or collenchyma cells. It produces phellem or cork



Fig. 9. Diagram of a segment of an oak log showing the grain in three dimensions, that is, the pattern formed by the annual rings and the wood rays. At the right, the block has been cut radially, and above, transversely. At the left (the surface of the log), a portion of the phloem and bark has been removed, showing a tangential view of the wood underneath. (After E. W. Sinnott and K. S. Wilson, Botany: Principles and Problems, 5th ed., McGraw-Hill, 1955)

outwardly and phelloderm or secondary cortex inwardly.

Woody dicotyledons. The cambium here is a continuous cylinder. The fascicular cambium produces secondary xylem and phloem; the interfascicular cambium forms either the same kind of tissue as the fascicular cambium or only vascular ray tissue.

As seen in cross sections, the secondary xylem, or wood, is formed in distinct concentric rings called annual, or growth, rings (Fig. 9). Each annual ring is composed of an inner portion called the spring, or early, wood, in which the vessels and tracheids are relatively large, and an outer portion called the summer, or late, wood, in which the cells have much smaller diameters and thicker walls than those of the early wood. Usually only the outermost growth rings function in conduction of water. The outer, lightcolored, relatively soft, functioning part of the wood is called sapwood. The nonconducting older (early formed) wood, called heartwood, is often filled with gums, resins, tannins, and mineral salts, and is dark in color. Growth rings in oblique or horizontal branches become asymmetrical, the rings being wider in the upper half (in most angiosperms) or in the lower half (in conifers). The wood with the wider rings has a different histologic and chemical structure from that with the narrower and is called tension wood in the angiosperms and compression wood in the conifers.

The wood varies in structure, chiefly depending on (1) the relative amounts of vessels, tracheids, and wood fibers, (2) the distribution of wood parenchyma, and (3) the presence and character of vascular rays.

Two patterns of vessel arrangement are recognized. If the early wood contains wide vessels and the late wood narrow vessels, so that each growth ring begins with a zone of large vessels, the wood is called ring-porous (**Fig. 10***a*). If vessels of more or less uniform widths are formed throughout the year, the wood is called diffuse-porous (Fig. 10*b*).

Wood texture, coarse or fine, refers to the size and

number of the xylem cells, especially vessels. The term grain refers to the arrangement of the cells, especially fibers, making up the wood. A wood is hard or soft, depending upon the number of lignified fibers. *See* WOOD ANATOMY.

Among the many wood products, plywood and veneer are particularly important. In making plywood, the wood is softened and the log rotated against a heavy blade to produce a continuous sheet of very thin wood which is cut to suitable lengths; pieces alternately stacked with grain at right angles are bonded together in varying odd numbers of layers, pressed, and dried. Plywood, or other wood bases, may be covered by thin sheets of fine-grained wood or veneer. Characteristic and interesting grain patterns for fine woods or veneers are obtained by cutting wood in transverse, radial, and tangential section (**Fig. 11**). Veneer wood with the most interesting pattern or grain comes from the base of a tree,



Fig. 10. Transverse sections, magnified, of two kinds of dicotyledonous wood. (a) White oak, ring-porous wood with large spring vessels and narrow (uniseriate) and wide (multiseriate) rays. (b) Yellow birch, a diffuse-porous wood with narrow (mostly uniseriate) rays. (U. S. Forest Products Laboratory)



Fig. 11. Silver maple (*Acer saccharinum*) wood in different sections. Dark and light colors and grain irregularities result from cutting through fibers and vessels of annual rings in different planes. Also, the rays, particularly in radial section, give the wood a pleasing and interesting appearance. (a) Transverse section. (b) Radial section. (c) Tangential section. (*From J. B. Hill, L. O. Overholts, and H. W. Popp, Botany: A Textbook for Colleges, 2d ed., McGraw-Hill, 1950*)



Fig. 12. Block of white-pine wood, in three dimensions, showing relationships of tracheids and ray elements. (After J. B. Hill, L. O. Overholts, and H. W. Popp, Botany: A Textbook for Colleges, 2d ed., McGraw-Hill, 1950)

from burls, or from crotch wood, where the vascular arrangements may be wavy or spiral. Some of the choice curly grains, such as curly maple, result from the development of adventitious buds. Injury to the cambium may produce a similar pattern. *See* WOOD PRODUCTS.

Gymnosperms. Most gymnosperms are similar to the woody dicotyledons in that a vascular cambium forms a continuous cylinder of secondary xylem and phloem. Commercial gymnosperm wood is usually derived from conifers. Conifer wood is composed of tracheids, fiber tracheids, and uniseriate wood rays, but no vessels (**Fig. 12**). Because of the absence of vessels, the conifer wood is more homogeneous in sections than the angiosperm wood (**Fig. 13**).

As stems grow and form new branches, those near the base are likely to be crowded, shaded, and die, particularly in pine. The dead branches fall off, and their bases are covered by successive layers of wood, forming knots. Knots weaken the wood and are not desirable unless for ornamental paneling.

The secondary growth of other gymnosperms is similar to that of the conifers. The cycads are interesting because growth rings are added at intervals of 2-20 years and are difficult to distinguish, except in *Dioon*. Many of the cycads form only one ring of wood; but others, such as *Cycas*, *Dioon*, *Macrozamia*, and *Encephalartos*, produce several rings.

Monocotyledons. Some monocotyledons, as the bamboos and palms, may form rather woody trunks by primary thickening. The thickening growth may occur rather generally throughout the stem, or there may be a primary thickening meristem, a cambiumlike zone of cells originating from the ground tissue beneath young leaf bases. It increases the width of the stem directly beneath the apex. Additional division of cells throughout the stem and the enlargement of cells complete the growth in thickness. In certain monocotyledons, such as Aloe, Yucca, Agave, and Dracaena, secondary growth also occurs. However, the cambium does not originate in the bundles, which are scattered in the ground tissue. It appears in the ground parenchyma outside the vascular bundles and by tangential divisions forms parenchyma on the outside and both vascular strands and parenchyma on the inside. The basic structure of the primary and secondary bodies is the same-in both, discrete vascular strands are embedded in ground tissue, usually sclerified (hardened) parenchyma.

Effect of secondary growth. The secondary xylem covers the primary xylem and pith, usually without changing them; whereas the primary phloem and cortex are pushed outward and are more or less compressed by the centrifugal growth of the secondary vascular tissues. The epidermis and cortex may keep pace with secondary vascular growth, or they may be ruptured and replaced by periderm.

Cambium develops in the parenchyma of the leaf gap, and its derivatives close the leaf gaps and break and bury the leaf traces.



Fig. 13. Enlarged photomicrographs of pine wood. (a) Transverse section. The wood cells are cut across at right angles to their length. Note the thin-walled cells (tracheids) in the springwood and the thick-walled cells in the summerwood (fiber tracheids). The uniseriate wood rays run at right angles to the annual rings. The large openings are resin canals. (b) Radial section. (c) Tangential section. In the last two sections the tracheids are cut lengthwise. In the radial section the rays run in a horizontal direction across the tracheids. In the tangential section the rays appear as lens-shaped clusters of cells. (U. S. Forest Products Laboratory)



Fig. 14. Cross section of 4-year-old pine stem, showing annual xylem rings and other tissues. (From J. B. Hill, L. O. Overholts, and H. W. Popp, Botany: A Textbook for Colleges, 2d ed., McGraw-Hill, 1950)

Types

The stems of vascular plants, or Tracheophyta, show an endless variety of form and structure. In this article only a few are considered as type examples.

Gymnosperm stems. This group of Tracheophyta contains some of the largest plants in the world (especially in Pinales), as well as *Welwitschia*, in which there is no aerial stem at all. No annual or perennial herbs are recognized in this group. *See* PINOPSIDA.

Pinales. The gymnosperms of this order are muchbranched, small-leaved, woody tree forms, except some shrubby junipers. The stem has typically an ectophloic siphonostele. Most conifers produce a conspicuous amount of secondary tissues which form a solid continuous cylinder (**Fig. 14**). Resin ducts occur in the cortex and the vascular tissues, except in the yews. The wood is dense and massive, composed of tracheids, fiber tracheids, and uniseriate wood rays. *See* SECRETORY STRUCTURES (PLANT).

Cycadales. The stem types range from tuberous and partly or wholly subterranean to relatively tall, aerial stems. The stem is usually unbranched and bears a crown of large leaves. The cortex and pith are large, loose, and parenchymatous. The vascular cylinder is narrow and has broad medullary rays and a loose-

textured wood. The stele is a siphonostele. The primary xylem usually contains scalariform tracheids, the secondary xylem tracheids with bordered pits. Leaf traces girdle the stem that is, they have a nearly horizontal course in the cortex. The stem is rigid mainly because of persistent leaf bases.

Ginkgoales. The only living representative, *Ginkgo biloba*, is a large profusely branched tree with both long shoots (normal vegetative) and short shoots (restricted vegetative or reproductive). Secondary growth is vigorous, and the pith and cortex are relatively small.

Gnetales. The three genera included in this order are highly specialized. *Ephedra* is usually a much branched shrub, *Gnetum* species are mostly lianas, and in *Welwitschia* most of the stem is buried in the ground. A feature which distinguishes these plants from other gymnosperms is the presence of vessels among the xylem elements.

Angiosperms. The angiosperms (flowering plants) include two main groups: the dicotyledons (plants with two seed leaves) and the monocotyledons (plants with one seed leaf). The stems of these two groups of plants show some differences. It is customary also to divide the stems of angiosperms into woody and herbaceous types. Although the various stem types intergrade in their character, some approximate generalizations regarding their differences may be made, as shown in the **table**.



Fig. 15. Woody dicotyledon; 1-year-old woody stem of *Liriodendron*, showing characteristics of ectophloic siphonostele; epidermis not yet replaced by periderm.

Comparison of dicotyledon and monocotyledon stems			
Dicotyledons	Monocotyledons		
Stems woody, herbaceous, vines	Stems herbaceous, few woody		
Secondary growth common; phloem, vascular cambium, and xylem in concentric layers; separate vascular bundles common in the primary body, uncommon in secondary	Some woody types with secondary growth; vascular cambium not between xylem and phloem; in the primary body vascular bundles often scattered through ground tissue, sometimes in two circles; in the secondary body also separate vascular bundles		
Distinct pith and cortex	No definable cortex and pith		
Endodermis and pericycle usually absent; sometimes pericycle fibers on outer limit of vascular cylinder; usually phloem fibers	Endodermis and pericycle usually absent; vascular bundles with sheaths of sclerenchyma		
Number of leaf traces to each leaf is variable, rarely large	Number of leaf traces to each leaf is large; sheathing leaf bases enclose the stem		
Periderm frequently present	Periderm not common		



Fig. 16. Transverse sections of woody vines. (a) Clematis. (b) Serjania mexicana; mature stem shows irregular development (unequal amounts) of secondary xylem and phloem. (c) Serjania subdentata; mature stem shows five well-developed cortical steles.

Woody dicotyledons. The primary vascular cylinder in the woody dicotyledons is an ectophloic siphonostele, and the pith and cortex are well defined. *Liriodendron* illustrates this type of structure (**Fig. 15**). The stem from the inside out consists of (1) a central pith of large, loose, parenchyma cells; (2) a vascular system containing closely arranged vascular bundles, with the leaf traces often recognizable as rather large bundles protruding into the cortex; (3) a cortex; and (4) an epidermis. When secondary growth occurs, a continuous vascular cambium and continuous layers of secondary xylem and phloem are formed. The epidermis is eventually replaced by a periderm. There is no endodermis and no pericycle. Primary phloem fibers form the outer limit of the vascular region.

Woody vines or lianas. Vines are generally characterized by a primary and secondary vascular system dissected into strands by broad rays (**Fig. 16***a*).



Fig. 17. The primary tissues of the stem of a typical herbaceous dicotyledonous plant, sunflower (*Helianthus annuus*), showing the characteristics of an ectophloic siphonostele. (a) Transverse section of entire stem showing the arrangement of vascular tissues in one circle. (b) Enlarged portion of a showing details of the structure, including one vascular bundle. (From J. B. Hill, L. O. Overholts, and H. W. Popp, Botany: A Textbook for Colleges, 2d ed., McGraw-Hill, 1950)

Also, they often have a ringbark which results from concentric successively formed periderms. Lenticels may be absent in ringbark.

Tropical lianas (climbing woody vines) often have a more or less unusual structure. There may be, for example, a highly irregular development of secondary xylem and phloem, so that the outline of the xylem is uneven (Fig. 16b). A still more striking structure results when the leaf traces develop complete cylinders of cambium, xylem, and phloem at levels where they have diverged from the central cylinder. The stem appears as though composed of several central cylinders (Fig. 16c).

Herbaceous dicotyledons. These forms differ from the arborescent, or woody, dicotyledons chiefly in having a smaller amount of cambial activity or none. They are often similar to 1-year-old woody stems. The primary vascular cylinder forms an ectophloic siphonostele with the vascular strands often more widely separated by interfascicular regions than those in the arborescent species (**Fig. 17**). Vascular cambium may be absent, as in *Ranunculus*; may be present but form only parenchyma and sclerenchyma in the interfascicular regions, as in *Medicago*, or may form a continuous vascular cylinder, as in *Vinca* or *Digitalis*.

Herbaceous monocotyledons. In these plants the vascular system is composed of widely spaced strands arranged in one of four ways. First, as in most grasses, the vascular bundles are arranged in two circles, with the outer smaller bundles embedded in a continuous sheath of sclerenchyma close to the epidermis (**Fig. 18***a*). The vascular bundles are collateral, each enclosed in a sheath of sclerenchyma. The pith may break down in the internodes but not in the nodes. Transverse bundles in the nodal region interconnect the leaf traces.

Second, in a few monocotyledons, for example *Clintonia* (Fig. 18*b*), a single, concentrically arranged series of bundles may occur. Each bundle has its own complete endodermis.

Third, in a few monocotyledons the bundles are grouped in the center of the stem, as in the rhizome



Fig. 18. Photomicrographs of transverse sections of monocotyledonous stems showing various arrangements of vascular bundles. (a) Wheat (grass). (b) Clintonia. (c) Corn. (d) Bamboo. (From J. B. Hill, L. O. Overholts, and H. W. Popp, Botany: A Textbook for Colleges, 2d ed., McGraw-Hill, 1950)

of *Acorus calamus*. The bundles are amphivasal, the phloem being surrounded by xylem.

Fourth and most commonly, vascular bundles are numerous and scattered in a ground tissue, as in corn or bamboo (Fig. 18c and d), with no pith or cortical regions being evident. The complex arrangement of the vascular bundles is related to the large number and variation in size of leaf traces. There may or may not be a sclerenchyma cylinder about each bundle, but subepidermal parenchyma is strongly sclerified. Endodermis and pericycle are absent.

Woody monocotyledons. Some monocotyledons, as the bamboos and palms, may form rather woody trunks by thickening of cell walls with age until nearly the whole ground tissue is heavily sclerified. Despite their large size these trunks are composed of primary tissues only. There are, however, woody monocotyledons with secondary growth, but the secondary tissues resemble the primary tissues in that they are composed of vascular bundles embedded in ground tissue, usually sclerified parenchyma.

Specialized erect stems. Two kinds of erect stems have special names. Culm is a name applied to hollow, but solid-jointed, stems of grasses, woody and herbaceous. Caudex means the axis of a plant, consisting of root and stem. The term is sometimes applied to short, enduring stems or stocks which throw up new stalks each year from persistent buds at ground level. Also, the name is used to describe the trunks of palms and tree ferns which consist of persistent leaf bases forming a column.

Perennial stems. Perennials which have woody stems are generally called trees or shrubs. Shrubs are woody perennials with several main stems branched from or near the ground and generally not more than 20–25 ft (6.0–7.5 m) high. Trees are woody perennials with a single main trunk or axis which rises some distance above the ground before branching. Trees generally exceed shrubs in height. The two words are convenient but not exact terms because trees and shrubs intergrade in height and form. Herbaceous plants also may be perennial. Their stems commonly die down to the ground each year.

Modified aerial stems. Stems may be variously modified for photosynthesis, storage, propagation, support, and protection. *Cladode.* This term is applied to a branch with a single internode which is flattened and serves as a leaf, as is found in asparagus.

Phylloclade. A flattened leaflike shoot that serves to replace a leaf as a photosynthetic organ, as found in the *Opuntia* cactus, is called a phylloclade.

Thom. This hard, pointed projection may be a modified branch, or leaf, or stipule. Thorns may be unbranched, as in the osage orange, or branched, as in the honey locust. A lateral branch may form leaves and flowers, then terminate growth by producing a thorn at the apex.

Creeping or prostrate stems. Stems that trail along the surface of the ground and that may take root at the nodes are called creeping or prostrate stems.

Runner. A runner, as in the strawberry, is a horizontally growing, sympodial stem system; that is, it appears single but actually is composed of a series of lateral branches arranged in lineal order. In this system the stem forms adventitious roots near the tip, then gives rise to a rosette of leaves from a bud. A new runner emerges from the axil of a reduced leaf at the base of this rosette.

Stolon. The term stolon is used for creeping stems, shorter than runners. They also root adventitiously (raspberry, currant). The term has no precise meaning.

Climbing stems or vines. These are long, slender stems which usually climb by special devices.

Rambler. The ramblers rest on the tops of other plants and some, as certain roses, have spines or prickles which help them to adhere to their support.

Root climbers. English ivy and posion ivy have stems which climb by means of adventitious roots.

Tendril climbers. These plants climb by means of modified leaf tendrils, as in the garden pea, or by stem tendrils, as in the grape.

Twiners. In the twiners the whole stem winds about its support, as pole beans or many tropical lianas.

Underground stems. Some stems grow underground and are often mistaken for roots. The main kinds of underground stems are rhizomes or rootstocks, tubers, corms, bulbs, and rhizomorphic droppers.

Rhizomes are usually quite varied, plagiotropic (growing horizontally), perennial, underground

stems found in a vast number of plants. Their cauline (stem) nature is evident by well-defined nodes and internodes. Roots, scale leaves, and axillary buds form at the nodes. Some buds form leafy, upright shoots while others form new, sympodially branched underground shoots. Among different species, rhizomes may be (1) thin, tough, and rapidly growing; (2) fleshy though rapidly growing; and (3) short and fleshy, serving more for storage than for spreading. The anatomy of rhizomes, except for some dorsiventrality, is much like that of the aerial stem of the same species. The other types of underground stems, such as tubers, corms, and bulbs, are essentially modified rhizomes. James E. Gunckel

Bibliography. J. D. Mauseth, Plant Anatomy, 1988.

Stem cells

Cells that have the ability to self-replicate and to give rise to mature cells. The concept of stem cells was originally based on renewing tissues. Many adult tissues, such as the skin, blood, and intestines, consist of mostly mature and short-lived cells that must be continuously replaced. Stem cells were postulated as the source of the self-renewal. In the early 1960s, Canadian scientists Ernest A. McCulloch and James E. Till provided the first experimental proof of the existence of stem cells in the blood system. They revealed that a type of cell in bone marrow possesses the capacity to replicate itself and to differentiate to various lineages of mature blood cells. Self-renewal, together with the capacity for differentiation, defined the properties of stem cells. This definition is generally used in stem cell biology today.

Stem cells can be found at different stages of fetal development and are present in a wide range of adult tissues. Many of the terms used to distinguish stem cells are based on their origins and the cell types of their progeny. There are three basic types of stem cells. Totipotent stem cells, meaning that their potential is total, have the capacity to give rise to every cell type of the body and to form an entire organism. Pluripotent stem cells, such as embryonic stem cells, are capable of generating virtually all cell types of the body but are unable to form a functioning organism. Multipotent stem cells can give rise only to a limited number of cell types. For example, adult stem cells, also called organ- or tissue-specific stem cells, are multipotent stem cells found in specialized organs and tissues after birth. Their primary function is to replenish cells lost from normal turnover or disease in the specific organs and tissues in which they are found. See CELL DIFFERENTIATION; HISTOGENESIS.

Totipotent and embryonic stem cells. Totipotent stem cells occur at the earliest stage of embryonic development. The union of sperm and egg creates a single totipotent cell. This cell divides into identical cells in the first hours after fertilization. All these cells have the potential to develop into a fetus when they are placed into the uterus. [To date, no such totipotent stem cell lines (primary cell cultures) have been developed.] The first differentiation

of totipotent cells forms a sphere of cells called the blastocyst, which has an outer layer of cells and an inner cell mass. The outer layer of cells will form the placenta and other supporting tissues during fetal development, whereas cells of the inner cell mass go on to form all three primary germ layers: ectoderm, mesoderm, and endoderm. The three germ layers are the embryonic source of all types of cells and tissues of the body. Embryonic stem cells are derived from the inner cell mass of the blastocyst. They retain the capacity to give rise to cells of all three germ layers. However, embryonic stem cells cannot form a complete organism because they are unable to generate the entire spectrum of cells and structures required for fetal development. Thus, embryonic stem cells are pluripotent, not totipotent, stem cells. See EMBRYOLOGY; EMBRYONIC DIFFERENTIA-TION; GERM LAYERS.

Embryonic germ cells. Embryonic germ cells differ from embryonic stem cells in the tissue sources from which they are derived, but appear to be similar to embryonic stem cells in their pluripotency. Human embryonic germ cell lines are established from the cultures of the primordial germ cells obtained from the gonadal ridge of late-stage embryos, a specific part that normally develops into the testes or the ovaries. Embryonic germ cells in culture, like cultured embryonic stem cells, form embryoid bodies, which are dense, multilayered cell aggregates consisting of partially differentiated cells. The embryoid body-derived cells have high growth potential. The cell lines generated from cultures of the embryoid body cells can give rise to cells of all three embryonic germ layers, indicating that embryonic germ cells may represent another source of pluripotent stem cells.

Growing mouse embryonic stem cells. Much of the knowledge about embryonic development and stem cells has been accumulated from basic research on mouse embryonic stem cells. The techniques for separating and culturing mouse embryonic stem cells from the inner cell mass of the blastocyst were first developed in the early 1980s. To maintain their growth potential and pluripotency, mouse embryonic stem cells can be grown on a feeder layer, usually consisting of mouse embryonic fibroblast cells. The feeder cells support embryonic stem cells by secreting a cytokine growth factor, the leukemia inhibitory factor, into the growth medium. Alternatively, purified leukemia inhibitory factor can be added to the growth medium without the use of a mouse embryonic feeder layer. (The leukemia inhibitory factor serves as an essential growth factor to maintain embryonic stem cells in culture.) A line of embryonic stem cells can be generated from a single cell under culture conditions that keep embryonic stem cells in a proliferative and undifferentiated state. Embryonic stem cell lines can produce indefinite numbers of identical stem cells. When mouse embryonic stem cells are integrated into an embryo at the blastocyst stage, the introduced embryonic stem cells can contribute to cells in all tissues of the resulting mouse. In the absence of feeder cells and the

leukemia inhibitory factor in cultures, embryonic stem cells undergo differentiation spontaneously. Many studies are focused on directing differentiation of embryonic stem cells in culture. The goal is to generate specific cell types. Formation of cell aggregates with three-dimensional structure during embryonic stem cell differentiation in culture may allow some of the cell-cell interaction to mimic that of in vivo development. The culture conditions can be designed to support and select specific cell types. With these experimental strategies, preliminary success has been achieved to generate some cell types, such as primitive types of vascular structures, blood cells, nerve cells, and pancreatic insulin-producing cells. *See* GENETIC ENGINEERING; SOMATIC CELL GENETICS.

Growing human embryonic stem cells. Since 1998, research teams have succeeded in growing human embryonic stem cells in culture. Human embryonic stem cell lines have been established from the inner cell mass of human blastocysts that were produced through in vitro fertilization procedures. The techniques for growing human embryonic stem cells are similar to those used for growth of mouse embryonic stem cells. However, human embryonic stem cells must be grown on a mouse embryonic fibroblast feeder layer or in media conditioned by mouse embryonic fibroblasts (see illustration). There are a number of human embryonic stem cell lines being generated and maintained in laboratories in the United States and other nations, including Australia, Sweden, India, South Korea, and Israel. The National Institutes of Health has created a Human Embryonic Stem Cell Registry, which lists stem cell lines that have been developed and can be used for research. Human embryonic stem cell lines can be maintained in culture to generate indefinite numbers of identical stem cells for research. As with mouse embryonic stem cells, culture conditions have been designed to direct differentiation into specific cell types (for example, neural and hematopoietic cells).

Adult stem cells. Adult stem cells, also referred to as somatic stem cells, occur in a wide variety of mature tissues in adults as well as in children. Like all stem cells, adult stem cells can self-replicate. Their ability to self-renew can last throughout the lifetime of individual organisms. Unlike embryonic stem cells, though, it is usually difficult to expand adult stem cells in culture. Adult stem cells reside in specific organs and tissues but account for a very small number of the cells in tissues. They are responsible for maintaining a stable state of the specialized tissues. To replace lost cells, stem cells typically generate intermediate cells called precursor or progenitor cells, which are no longer capable of selfrenewal. However, they continue undergoing cell division, coupled with maturation, to yield fully specialized cells. Such stem cells have been identified in many types of adult tissues, including bone marrow, blood, skin, gastrointestinal tract, dental pulp, retina of the eye, skeletal muscle, liver, pancreas, and brain. Adult stem cells are usually designated according to their source and their potential. Adult stem cells are multipotent because their potential



A colony of the human embryonic stem cell line H9 growing in culture. The elongated cells surrounding the colony are part of the mouse embryonic fibroblast feeder layer on which the stem cells are grown. The inset shows the enlarged area, with arrows pointing to the two nuclei of a single stem cell. The H9 cell line meets the criteria for the use of human embryonic stem cells by researchers funded by the U.S. government. (*Courtesy of Clay Glennon*)

is normally limited to one or more lineages of specialized cells. However, a special multipotent stem cell that can be found in bone marrow, called the mesenchymal stem cell, can produce all cell types of bone, cartilage, fat, blood, and connective tissues. *See* REGENERATION (BIOLOGY).

Blood stem cells. Blood stem cells, or hematopoietic stem cells, are the most studied type of adult stem cells. The concept of hematopoietic stem cells is not new, as it has been long realized that mature blood cells are constantly lost and destroyed. Billions of new blood cells are produced each day to make up the loss. This process of blood cell generation, called hematopoiesis, occurs largely in the bone marrow. The presence of hematopoietic stem cells in the bone marrow was first demonstrated by E. A. McCulloch and J. E. Till in a mouse model in the early 1960s. The first experimental work on stem cells was an unexpected outcome from their study for measuring the effects of radiation. They found that the blood system of a mouse that has been subjected to heavy radiation can be restored by infusion of bone marrow. The stem cells responsible for reconstituting the blood system generate visible cell colonies on the spleen of the recipient mouse. Each of the spleen colonies consists of one or more types of blood cells, and all the cells in a colony are derived from a single cell. Self-renewal capacity of the colony-forming cells is demonstrated by their ability to form secondary spleen colonies. Such blood stem cells, known as colony forming unit-spleen cells, qualify as pluripotent hematopoietic stem cells because they can replicate and give rise to multiple types of mature blood cells. A definitive proof of blood stem cells is their ability to reconstitute the blood system. Bone marrow transplantation demonstrates the restorative powers of blood stem cells in humans. *See* HEMATOPOIESIS; TRANS-PLANTATION BIOLOGY.

Isolating blood stem cells. Like other adult stem cells, blood stem cells are rare and difficult to isolate. Only about 1 in 100,000 cells in the bone marrow is a stem cell. Scientists have used cell-sorting methods to enrich and purify blood stem cells. Stem cells differ from mature cells in their surface markers, which are specific protein molecules on the cell membrane that can be tagged with monoclonal antibodies. By using a set of surface markers, some expressed mainly on stem cells and others on mature blood cells, nearly pure populations of stem cells can be separated from bone marrow. The stem cells purified by this approach can engraft (begin to grow and function) and reconstitute the blood system in the recipient. In animal studies, as few as 30 purified stem cells can rescue a mouse that has been subjected to heavy radiation. Besides the bone marrow, a small number of blood stem cells can be found in circulating blood. In addition, stem cells in the bone marrow can be mobilized into the bloodstream by injecting the donor with certain growth factors or cytokines. This approach can result in a large number of stem cells circulating in peripheral blood, from which they can be collected and used for transplant therapy.

Umbilical cord blood and cord blood banks. An alternative source of blood stem cells is human umbilical cord blood, a small amount of blood remaining in the placenta and blood vessels of the umbilical cord. It is traditionally treated as a waste material after delivery of the newborn. However, since the recognition of the presence of blood stem cells in umbilical cord blood in the late 1980s, its collection and banking has grown quickly. Similar to bone marrow, umbilical cord blood can be used as a source material of stem cells for transplant therapy. In 1989, the first successful cord blood transplant was reported for treating a 6-year-old boy suffering from Fanconi's anemia (an inherited disease that primarily affects the bone marrow, resulting in decreased production of blood cells) in Paris. Since then, over 6000 cord blood stem cell transplants have been performed worldwide, mainly in patients with blood conditions and in some cancer therapies. However, because of the limited number of stem cells in umbilical cord blood, most of the procedures are performed on young children of relatively low body weight. A current focus of study is to promote the growth of umbilical cord blood stem cells in culture in order to generate sufficient numbers of stem cells for adult recipients.

Many blood banks have been established to collect and cryopreserve cord blood cells. Commercial banks offer services of storing cord blood of healthy newborns for potential future use by themselves or their siblings. Although it is considered a biological insurance, the chance of a child using his or her own cord blood is estimated at 1 per 20,000 collections. Of the estimated 6000 cord blood transplants, only 14 were performed using autologous sources.

Public banks encourage donation of cord blood for unrelated transplants. The Stem Cell Research and Therapeutic Act of 2005 (H.R. 2520) established a national umbilical cord blood program, providing federal funding to collect and store cord blood for blood cell transplants. The program functions to provide a national inventory of 150,000 cord blood units for public use and to establish a registry network integrated with the national marrow donor registry administered by the National Marrow Donor Program (NMDP).

Mesenchymal stem cells. Mesenchymal stem cells (MSCs) are a type of multipotent adult stem cells, and they are defined by the capacity to give rise to a variety of connective tissue lineages, including bone, cartilage, tendon, muscle, and fat cells. Classic studies found a type of cells in bone marrow stroma capable of generating fibroblast-like cell colonies. These clonogenic cells were termed colony forming unitfibroblasts (CFU-F). CFU-F share some characteristics of MSCs. MSCs appear as fibroblast-like spindleshaped cells. CFU-F assay is still used to evaluate MSCs in cell cultures. MSCs can be distinguished and isolated from other cells based on phenotypic characteristics. Typically, MSCs express specific surface antigens SH2, SH4, and STRO-1 and lack blood cell markers CD45 and CD34. MSCs can replicate as multipotent cells. The mesenchymal cell lineage potential can be demonstrated in vitro with appropriate culture conditions. Differentiation can be induced to osteocytes by dexamethasone and ascorbate, to chrondrocytes by transforming growth factor- β 3, or to adipocytes by dexamethasone and insulin.

MSCs are primarily obtained from bone marrow stromal cells. They are also found in small numbers in umbilical cord blood. In addition, adipose-derived stem cells (ASCs) have been shown to be similar to MSCs. Fat tissue is of mesenchymal origin and contains stromal components. ASCs can be isolated from fat tissue by the method of liposuction. Human ASCs have been shown to exhibit the capacity to give rise to fat, bone, cartilage, muscle, and possibly neurons. Thus, ASCs may provide a potential source of multipotent adult stem cells.

Neural stem cells. Neural stem cells, the multipotent stem cells that generate nerve cells, are a new focus in stem cell research. Active cellular turnover does not occur in the adult nervous system as it does in renewing tissues such as blood or skin. Because of this observation, it had been dogma that the adult brain and spinal cord were unable to regenerate new nerve cells. However, since the early 1990s, neural stem cells have been isolated from the adult brain as well as from fetal brain tissues. Stem cells in the adult brain are found in the areas called the subventricular zone and the ventricle zone. Brain ventricles are small cavities filled with cerebrospinal fluid. Another location of brain stem cells occurs in the hippocampus, a special structure of the cerebral cortex related to memory function. Stem cells isolated from these areas are able to divide and to give rise to nerve cells (neurons) and neuron-supporting cell types in culture.

Plasticity. Stem cell plasticity refers to the phenomenon of adult stem cells from one tissue generating the specialized cells of another tissue. The longstanding concept of adult organ-specific stem cells is that they are restricted to producing the cell types of their specific tissues. However, a series of recent studies have challenged the concept of tissue restriction of adult stem cells. Much of the experimental evidence is derived from transplant studies with blood stem cells. Bone marrow stem cells have been shown to contribute to liver, skeletal muscle, and cardiac cells in human recipients. In mouse models, purified blood stem cells have been demonstrated to generate cells in nonblood tissues, including the liver, gut, and skin. Although the stem cells appear able to cross their tissue-specific boundaries, crossing occurs generally at a low frequency and mostly only under conditions of host organ damage. The finding of stem cell plasticity is unorthodox and unexpected (since adult stem cells are considered to be organ/tissuespecific), but it carries significant implications for potential cell therapy. For example, if differentiation can be redirected, stem cells of abundant source and easy access, such as blood stem cells in bone marrow or umbilical cord blood, could be used to substitute stem cells in tissues that are difficult to isolate, such as heart and nervous system tissue. However, the concept of plasticity has been the subject of controversy. The observed frequency of lineage conversion is generally low. An alternative explanation to plasticity is the phenomenon of fusion of host and donor cells. Recent findings suggest that blood cells contribute to other tissues by fusing with preexisting cells rather than by converting to other cell lineages

Potential clinical applications. Stem cells hold great potential for developing cell therapies to treat a wide range of human diseases. Already in clinical use is blood stem cell transplant therapy, well known as bone marrow transplant therapy for the treatment of patients with certain types of blood diseases and cancers. The discovery of stem cells in various adult tissues, stem cell plasticity, and human embryonic stem cells brings new excitement and opportunities. Stem cells offer the possibility of cell replacement therapy for many human diseases, such as Parkinson's and Alzheimer's diseases, spinal cord injury, diabetes, heart disease, and arthritis, that result from loss or damage of cells in a specialized tissue of the body. Stem cell therapy might revolutionize the treatment and outcome of these diseases. Stem cell science is still in the very early stage. Much more research is required to understand the biological mechanisms that govern cell differentiation and to identify factors that direct cell specialization. Future cell therapy will depend largely on advances in the understanding of stem cell biology and the ability to harness the process of stem cell growth and differentiation.

Somatic cell nuclear transfer (SCNT) stem cells. SCNT involves a micromanipulation procedure in which the nucleus of an egg is removed and replaced by a nucleus taken from somatic cells, typically skin cells. Successful nuclear transfer requires reprogramming of the donor nucleus. The cells so created may divide in cultures to generate embryonic stem cells that can initiate embryogenesis. This is the technique being used in cloning animals, such as the first cloned mammal, Dolly the sheep. However, cloning by nuclear transfer is observed with extremely low efficiency, probably due to faulty and incomplete reprogramming of the donor nucleus. The mechanisms governing the transition from a differentiated genome to a totipotent state remain largely unknown.

A major interest in SCNT is the prospect of creating patient-specific embryonic stem cells. These cells would be genetically identical to the nuclear donor except for maternal mitochondrial deoxyribonucleic acid (mtDNA) of the oocyte. Therefore, the problem of graft rejection would be avoided if the cells could be used for transplant therapy for the donor patients. The concept of using SCNT to generate customized stem cells for cell therapy is also referred to as therapeutic cloning. However, there are hurdles and limitations to using embryonic stem cells in clinical applications. A major challenge is to achieve the directed differentiation and controlled growth before stem cells can be used for transplant therapy. Another issue on SCNT in human stem cells is the sourcing of human eggs. The procedure requires a large number of eggs from women, and poses an ethical and technical challenge. See HISTOCOMPATI-BILITY.

The success in producing embryonic stem cell lines by the SCNT technique has been demonstrated in mice. In an article published in *Science* in 2005, a team led by Hwang Woo Suk of South Korea claimed the establishment of patient-specific stem cell lines by using the SCNT technique. However, the paper was later retracted as the results were fabricated and the claim a fraud. The field is still left uncertain if somatic nuclear replacement is feasible in humans.

Ethical and regulatory issues. The use of human embryonic stem cells raises ethical, social, and legal issues. The major concern centers on the source of stem cells. Human embryonic stem cell lines are made from the inner cell mass of a blastocyst stage embryo. Most embryos used to produce stem cells are left over from in vitro fertilization (IVF) treatment. The embryos are destroyed by the procedure of extracting stem cells. The early embryo has the biological potential to develop into a person. However, society has not reached consensus on when human life begins. The attention on stem cell research and cloning calls for regulation and legislation from governments. In the United States, current policy allows federal funds to be used for research only on existing human embryonic stem lines. The human embryonic stem cell lines that meet the eligibility criteria are listed in the Human Embryonic Stem Cell Registry by the National Institutes of Health (NIH).

One concern about SCNT is that it may lead to the reproductive cloning of humans. In theory, the embryo created via SCNT could be used to clone a human if it were implanted into a womans uterus. In the United States, the legislators in the House of Representatives and the Senate have introduced bills proposing a ban of all forms of cloning, including research cloning, or inhibiting reproductive cloning while preserving therapeutic cloning research. However, these bills have not been passed, and no federal law has been established on human embryonic stem cell research. The Canadian Parliament has passed Bill C-6 that prohibits creation of a human clone, sale of sperm or ova, and commercial surrogacy. The bill permits the use of stem cells obtained from discarded products of in vitro fertilization, that is, excess and unused embryos. In the United Kingdom, a law permits the use of embryos in research and therapeutic cloning research but bans reproductive cloning, and implanting a cloned embryo in a human uterus is liable to criminal prosecution. Cloning research must be licensed from the Human Fertilization and Embryology Authority that governs embryonic and stem cell research in the United Kingdom. Chen Wang

Bibliography. D. Kennedy, Editorial retraction [retraction of W. S. Hwang et al., Science. 308(5729):1777-1783, 2005], Science. 311(5759):335, 2006; J. Kurtzberg, A. D. Lyerly, and J. Sugarman, Untying the Gordian knot: Policies, practices, and ethical issues related to banking of umbilical cord blood, J. Clin. Investig., 115(10):2592-2597, 2005; E. A. McCulloch and J. E. Till, Perspectives on the properties of stem cells, Nat. Med., 11(10):v-vii, 2005; J. M. Polak and A. E. Bishop, Stem cells and tissue engineering: Past, present, and future, Ann. N. Y. Acad. Sci., 1068:352-366, 2006; J. Rossant, Stem cells from the mammalian blastocyst, Stem Cells, 19:447-482, 2001; G. J. Spangrude, S. Heimfeld, and I. L. Weissman, Purification and characterization of mouse hematopoietic stem cells, Science, 241:58-62, 1988; J. A. Thomson et al., Embryonic stem cell lines derived from human blastocysts, Science, 282:1145-1147, 1998; I. L. Weissman, D. J. Anderson, and F. Gage, Stem and progenitor cells: Origins, phenotypes, lineage commitment, and transdifferentiations, Annu. Rev. Cell Dev. Biol., 17:387-403, 2001.

Stenolaemata

A class of Bryozoa confined to fully marine water. The Stenolaemata include several thousand species distributed among four orders: Cystoporata, Trepostomata, Cryptostomata, and Cyclostomata (=Tubuliporata). The commonly recognized order Fenestrata likely is part of the Cryptostomata. First appearing late in the Early Ordovician, stenolaemates expanded quickly to dominate bryozoan assemblages through the Early Cretaceous. By mid-Cretaceous, the calcified gymnolaemate order Cheilostomata, which originated in the Late Jurassic, had increased in diversity and abundance to equal that of stenolaemates, eventually surpassing the stenolaemates, which are now a minor part of the bryozoan fauna in most marine environments. *See* BRYOZOA; CRYPTOSTOMATA; CYCLOSTOMATA (BRYOZOA); CYSTOPORATA; TREPOSTOMATA.

Colony morphology. Stenolaemate colonies vary greatly in size and shape. Many Paleozoic through mid-Mesozoic forms were large and massive, but more recent representatives are commonly small and delicate. Colonies are encrusting, erect, or freeliving. Some erect forms have single or regularly spaced, flexible cuticular joints; segments between joints and all other entire colonies are rigidly calcified and enclosed within a thin cuticular membrane. The outer cuticle is secreted by a thin layer of epithelial cells that on colony basal (underside) surfaces switch secretion from cuticle to calcite crystals bound within ultrathin cuticular sheets. This may also happen locally on other surfaces of colonies, generating fixed-wall regions. In actively growing zones and over entire nonbasal surfaces of many colonies, all skeleton above the basal surface is secreted by single layers of epithelial cells that are ultimately folded in from the basal wall, with a fluid-filled gap between the skeleton and the outer cuticle, generating free-walled colonies or regions.

Zooidal morphology. Living stenolaemate colonies are made up of individual feeding units called zooids with lophophores (rings of tentacles) that are circular in basal outline (see **illustration**). Lophophores of feeding zooids bear 8 to fewer than 20 tentacles that flare into usually simple conical or bell shapes when in the feeding position. The lophophore barely clears the skeletal aperture (terminal opening) when fully protruded, with the mouth most commonly at or just below the skeletal aperture. Tentacles of living stenolaemates (Cyclostomata) have lateral cilia tracts, but unlike other bryozoans lack frontal cilia.

Zooidal body cavities are enclosed within tubular or prismatic zooecia (skeletons of individual zooids), gradually expanding from their proximal ends. The outer cuticle folds into each zooid as a vestibular wall and generally attaches by ligaments to the skeleton



Some morphologic features of autozooids of Cyclostomata, the only living order of stenolaemates.

at some distance below the zooecial aperture. In all but one Mesozoic group, the zooidal orifice (wall opening through which lophophore and tentacles protrude) is closed by a ring of muscles, instead of a covering flap (epistome or operculum). Two membranes continue inward from the ligament, one extending as a sleeve-shaped sheet of ectoderm and mesoderm attached at its inner end around the base of tentacles. The other inward-extending membrane (the membranous sac) divides the body cavity and surrounds the gut, and consists only of mesodermal cells, interpolated annular muscle cells, and a basement membrane. Contraction of the muscle cells of the membranous sac appears to be critical in protrusion of the lophophore to its feeding position.

Zooecia typically are relatively thin-walled in colony interiors (endozone) and relatively thickwalled in outer regions (exozone). Inner surfaces of zooecia are lined by a single layer of epithelium, and it has been hypothesized that as a prerequisite for calcification of the tubular zooecia, the outer body wall of the stenolaemate ancestor split into the inner, mesodermal membranous sac and the outer, calcite-secreting ectoderm. Most Paleozoic stenolaemates had imperforate zooecia, but skeletal walls of most post-Paleozoic forms have pores (except basal walls). Ancient stenolaemates variously had planar transverse partitions (diaphragms), variously curved partitions or blisters, and spines subdividing or projecting into zooecial cavities. Undivided zooecial cavities function as living chambers, otherwise only the portion of the zooecial cavity on the outer side of the last-formed basal diaphragm functions as the living chamber. Living chambers vary across the class from about 2×10^{-3} mm³ to 1 mm³.

Polymorphism. Many stenolaemate colonies are constituted of skeletally monomorphic zooids, but polymorphs are present in some. The most common polymorphs are structural (various kenozooids), or females (gonozooids) with brooding chambers. More rarely a few other types of polymorphs occur, such as nanozooids (reduced zooids with single tentacles) and eleozooids (avicularium-like zooids).

Reproduction and growth. Individual colonies are hermaphroditic, some only with skeletally monomorphic, apparently hermaphroditic feeding zooids, and others with female zooids distinguished by chambers that brood polyembryonic embryos (multiple embryos developed by fission from a single fertilized egg). Larvae do not feed and are short-lived. They settle and metamorphose into a hemispherical protoecium from which a cylindrical tube extends to complete the ancestrular zooid (initial zooid of the colony). Colonies grow by asexual budding, and the polypide (organs and tissues) of each zooid is formed by proliferation of cells on the underside of the cuticle that overlies the skeletal orifice. Polypides usually live only a few weeks, after which they degenerate into a roughly spherical, brown-stained mass (brown body). A series of polypides may bud during the life of each zooid. Frank K. McKinney

Bibliography. R. S. Boardman, General features of the Class Stenolaemata, in R. A. Robison (ed.), *Treatise on Invertebrate Paleontology*, pt. G: *Bryozoa* (revised), vol. 1, Geological Society of America and University of Kansas, Boulder, 1989; R. S. Boardman, Reflections on the morphology, anatomy, evolution, and classification of the Class Stenolaemata (Bryozoa), *Smithson. Contrib. Paleobiol.*, 86:1–59, 1998; R. S. Boardman, A. H. Cheetham, and A. J. Rowell (eds.), *Fossil Invertebrates*, Blackwell Scientific Publications, Palo Alto, 1987; A. Ernst, Bryozoans, in B. D. Webby et al. (eds.), *The Great Ordovician Biodiversification Event*, pp. 147–156, Columbia University Press, New York, 2004; C. Nielsen and K. J. Pedersen, Cystid structure and protrusion of the polypide in *Crisia* (Bryozoa, Cyclostomata), *Acta Zool.* (Stockholm), 60:65–88, 1979.

Stephanoberyciformes

An order of fishes containing the pricklefishes and whalefishes. This order is recognized by G. D. Johnson and C. Patterson, but interrelationships of the taxa placed within it are debated among ichthyologists. Stephanoberyciformes is recognized as the sister group of all remaining acanthomorphs. It is defined by the following characters: pelvic girdle attached to cleithrum or coracoid of pectoral girdle; skull bone usually exceptionally thin (fragile bones often characteristic of pelagic fishes); subocular shelf absent; supramaxilla absent or reduced; and body shape variable, from elongate to short and rounded. The biology of stephanoberyciformes as a group is poorly known, and many species are known from only a few specimens. These are small deep-water fishes, ranging in total length from 2.1 to 43 cm (0.8 to 17 in.), but most are less than 10 cm (4 in.), and occupying depths to 5300 m (17,400 ft or 3.3 mi). Several species are known to be oviparous (the assumed mode of reproduction of the entire order), with planktonic eggs and larvae. The order comprises nine families, 28 genera, and about 75 described species, plus many others known but yet undescribed. Included in the order are families taken from the Beryciformes, Lampriformes, and the Trachichthyiformes taxa as described by J. A. Moore. See DEEP-SEA FAUNA; OSTEICHTHYES; TELEOSTEI.

Melamphaidae (bigscale fishes). Melamphaidae, comprising five genera and 36 species, can be distinguished by the following features: dorsal fin with 1–3 weak spines; pelvic fins thoracic or subthoracic, with 1 spine and 6 to 8 soft rays; caudal fin preceded dorsally and ventrally by 3 or 4 posteriorly directed spines; cycloid scales, often large and deciduous; lateral line absent, or limited to 1 or 2 pored scales; and lengths of 2.1 to 18 cm (0.8 to 7 in.), but most less than 10 cm (4 in.). Some species make daily vertical migrations from 700 m (2300 ft), and others are bathypelagic to depths exceeding 3500 m (11,500 ft). They occur in all oceans, but are unknown in Arctic seas and the Mediterranean Sea.

Stephanoberycidae (pricklefishes). Stephanoberycidae (**Fig. 1**), comprising three monotypic genera, can be identified by the following characteristics: 8 to 11 spines preceding the principal caudal fin rays



Fig. 1. Pricklefish (Stephanoberyx monae). (From G. Brown Goode and T. H. Bean, Oceanic Ichthyology, plate 55, 1896)

dorsally and ventrally; dorsal and anal fin spines, if present, weak; 10 to 14 soft rays in each dorsal and anal fin; pelvic fins abdominal or subabdominal, each with 5 soft rays; scales smooth or spiny; and lateral line faint. The three species represent the family in the eastern and western Atlantic (including the Gulf of Mexico), the Indian Ocean off South Africa, and the western and central Pacific. They are 8.1 to 13 cm (3.2 to 5.1 in.) in total length and have a depth range of 1655 to 5300 m (5430 to 17,400 ft).

Hispidoberycidae. Hispidoberycidae is a monotypic family known from the northeastern Indian Ocean and the South China Sea. The sole species, *Hispidoberyx ambagiosus*, has scales covered with small spines, an operculum with a stout spine, a dorsal fin with 4 or 5 spines and 10 soft rays, an anal fin with 3 spines and 9 soft rays, a total length of 18 cm (7 in.), and a depth range of 560 to 1000 m (1840 to 3280 ft).

Gibberichthyidae (gibberfishes). Gibberichthyidae comprises one genus and two species of rare deepsea fishes. Gibberichthys latifrons occurs in the Indo-West Pacific, where adults occupy depths between 750 and 2000 m (2460 and 6560 ft), and G. pumilus occurs in the western Central Atlantic (including the Gulf of Mexico) and the western Pacific at depths between 320 and 1100 m (1050 and 3600 ft). Both species are about 13 cm(5.1 in.) in total length. Larvae are usually found at depths less than 50 m (164 ft). The family is distinguishable by the following features: subabdominal pelvic fins, each with 1 spine and 5 or 6 soft rays; a series of 5 to 8 and 4 or 5 partly isolated short spines preceding the soft dorsal and anal fins, respectively; 7 to 9 dorsal and anal soft rays; cycloid scales; flanks with vertical rows of papillae over the vertical lateral line tubes; and swim bladder partially filled with fat. Small crustaceans are known foods of both prejuveniles and adults. See SWIM BLADDER.

Whalefishes. The following three families are called whalefishes; however, they are far from whale size, with the species ranging from only 11 to 39 cm (4 to 15 in.). The whalefishes share the following characteristics: whale-shaped body; very large mouth and highly distensible stomach; eyes well developed or degenerate; lateral line of large hollow tubes; luminous tissue on body; dorsal and anal fins far posteriorly and opposite one another; swim bladder absent; and black body with red and orange-tipped fins. Whalefishes are bathypelagic.

Rondeletiidae (redmouth whalefishes). Rondeletiidae is distinguished by a lateral line of 14 to 26 pores in a vertical series, no fin spines, and the presence

of pelvic fins. It consists of two species: *Rondeletia bicolor* (**Fig. 2**), from the western Atlantic off North Carolina and Suriname, which has a forward-directed spine above the eye and occurs at depths to 3000 m (9840 ft); and *R. loricata*, which occurs in temperate and tropical waters of the Atlantic, Indian, and Pacific occans to depths of 1130 m (3700 ft), makes diurnal migrations from 750 m (2460 ft) up to 100 m (328 ft), and is known to feed on isopods and crustacean remains. Both species are only about 14 cm (5.5 in.) in total length.

Barboursiidae (red whalefish). Barboursiidae is a family of only one species, *Barbourisia rufa*, which occurs in tropical and temperate latitudes throughout the world's oceans, including the western Atlantic (Gulf of Mexico) and eastern Pacific (off north-central California). Identifying features are as follows: fin spines absent; large mouth, with the maxilla extending far beyond the eye; lower jaw projecting beyond the upper; jaw teeth in broad villiform bands; body and fins with minute spinules, resulting in a velvety feel; lateral line with distinct pores; maximum total length of about 43 cm (17 in.); and juveniles mesopelagic and adults benthopelagic to depths of 2000 m (6560 ft).

Cetomimidae (flabby whalefishes). Cetomimidae (**Fig. 3**), with nine genera and 35 species (15 of which are known but yet undescribed), plus 36 species in the family Melamphaidae account for 75% of the order Stephanoberyciformes. The following features identify the family: skin loose and without scales; eyes very small or vestigial; no pelvic fins; 3 or 4 pairs of gill arches; photophores absent, but luminous organs often present around anus and dorsal and anal fin bases; and in live specimens brown or orange color with bright orange or red jaws and fins. The very few males that are known are only 3 to 5 cm (1.2 to 2 in.) in length and were previously classified as juveniles. The maximum total length is about 43 cm (17 in.). These fishes are oceanic in the



Fig. 2. Redmouth whalefish (*Rondeletia bicolor*). (*From G. Brown Goode and T. H. Bean, Oceanic Ichthyology, plate* 77, 1896)



Fig. 3. Flabby whalefish (Cetomimidae sp.). (Photo from Alaska Fisheries Science Center, National Oceanic and Atmospheric Administration, National Marine Fisheries Service)

Atlantic, Indian, and Pacific oceans, and are one of the deepest-dwelling families of the order, with one species, *Cetichthys parini*, at 5000 m (16,400 ft). *See* BIOLUMINESCENCE; PHOTOPHORE GLAND.

Mirapinnidae (tapertails). Mirapinnidae, comprising three genera and five species, occurs in tropical and subtropical zones of the Atlantic and Indian oceans and the western Pacific. The following characteristics identify the family: no scales; fins without spines; dorsal and anal fins opposite one another; dorsal fin with 16 to 33 soft rays and anal fin with 14 to 29 rays; pectoral fin with 13 to 24 rays; and pelvic fin jugular with 4 to 10 rays. Two subfamilies are recognized.

Mirapinninae, with one species, *Mirapinna esau*, is called hairyfish because of the short hairlike pile covering the body. It is further identified by pectoral fins inserted very high on the body, broad pelvic fins inserted in the jugular position well above the ventral profile, and overlapping halves of the caudal fin.

Eutaeniophorinae has two genera and four species. The body is eel-like, with the dorsal and anal fins placed far posteriorly and opposite one another.

Megalomycteridae (longnose fishes). Megalomycteridae comprises four genera and eight species, five of which are yet undescribed. The longnoses occupy depths between 1500 and 4000 m (4900 and 13,000 ft) in the Atlantic and Pacific oceans. They are identified by the following: exceptionally large olfactory organs; lack of fin spines; usually lacking pelvic fin, but if present thoracic with 1 to 3 rays; and less than 6 cm (2.4 in.) in total length. Herbert Boschung

Bibliography. A. W. Ebeling, Melamphaidae, pp. 427-432, in M. M. Smith and P. C. Heemstra (eds.), Smiths' Sea Fishes, Springer-Verlag, Berlin, 1986; A. W. Ebeling and W. H. Weed, Order Xenoberyces (Stephanoberyciformes), in Fishes of the Western North Atlantic, Mem. Sears Found. Mar. Res., Mem. 1 (pt. 6):397-478, 1973; R. Froese and D. Pauly (eds.), FishBase, World Wide Web electronic publication, version (05/2005); G. D. Johnson and C. Patterson, Percomorph phylogeny: A survey of acanthomorphs and a new proposal, Bull. Mar. Sci., 52(1):554-626, 1993; J. A. Moore, Phylogeny of the Trachichthyiformes (Teleostei: Percomorpha), Bull. Mar. Sci., 52(1):114-136, 1993; J. S. Nelson, Fishes of the World, 4th ed., Wiley, New York, 2006; J. R. Paxton, Barbourisiidae (p. 1740), Cetomimidae (pp. 1171-1173), Mirapinnidae (pp. 1174-1175), in K. E. Carpenter (ed.), The Living Marine Resources of the Western Central Atlantic, FAO Species Identification Guide for Fishery Purposes, vol. 2, FAO, Rome, 2003; J. R. Paxton, G. D. Johnson, and T. Trnski, Larvae and juveniles of the deepsea "whalefishes" Barbourisia and Rondeletia (Stephanoberyciformes: Barbourisiidae, Rondeletiidae), with comments on family relationships, Rec. Aust. Mus., 53:407-425, 2001; J. R. Paxton and T. Trnski, Rondeletiidae (pp. 1168-1169), Megalomycteridae (pp. 1176-1177), in K. E. Carpenter (ed.), The Living Marine Resources of the Western Central Atlantic, FAO Species Identification Guide for Fishery Purposes, vol. 2, FAO, Rome, 2003.

Stepping motor

An electromagnetic incremental-motion actuator which converts digital pulse inputs to analog output motion. The device is also termed a step motor. When energized in a programmed manner by a voltage and current input, usually dc, a step motor can index in angular or linear increments. With proper control, the output steps are always equal in number to the input command pulses. Each pulse advances the rotor shaft one step increment and latches it magnetically at the precise point to which it is stepped. Advances in digital computers and particularly microcomputers revolutionized the controls of step motors. These motors are found in many industrial control systems, and large numbers are used in computer peripheral equipment, such as printers, tape drives, capstan drives, and memoryaccess mechanisms. Step motors are also used in numerical control systems, machine-tool controls, process control, and many systems in the aerospace industry. See COMPUTER GRAPHICS; COMPUTER NU-MERICAL CONTROL; COMPUTER PERIPHERAL DEVICES; CONTROL SYSTEMS; DIGITAL-TO-ANALOG CONVERTER; PROCESS CONTROL.

Major types. There are many types of step motors. Most of the widely used ones can be classified as variable-reluctance, permanent-magnet, or hybrid permanent-magnet types. These three categories can be compared with regard to various properties (see **table**).

Variable-reluctance motor. A variable-reluctance step motor is simple to construct. The stator and rotor have different numbers of teeth. For example, a three-phase, single-stack motor can have 12 stator teeth and 8 rotor teeth, and be wound in a four-pole configuration (**Fig. 1**). When phase A is energized



Fig. 1. Three-phase, single-stack, variable-reluctance step motor with 12 stator teeth and 8 rotor teeth, wound in a four-pole configuration. Only the windings on phase A are shown. A, B, C are phases; N and S are north and south magnetic poles.

Comparison of major types of step motors			
Property	Variable-reluctance	Permanent-magnet	Hybrid permanent-magnet
Complexity	Simple construction		Relatively complex construction
Performance and efficiency	Low efficiency; requires relatively high current per unit torque	Low performance due to high inductance and stamped motor parts	Efficient; low current per unit torque requirement
Windings and drives	Unifilar windings and unipolar driver	Bifilar windings and unipolar driver; or unifilar windings and bipolar driver	Bifilar windings and unipolar driver; or unifilar windings and bipolar driver
Detent Rotor inertia	No inherent detent Low rotor inertia due to lack of a permanent magnet	Inherent detent High rotor inertia due to permanent magnet	Inherent detent High rotor inertia due to permanent magnet

by a dc voltage across the terminals, four of the rotor teeth will line up with the four stator teeth of the phase. When phase B is energized and phase A is deenergized, the rotor will rotate clockwise by 15° so that the other four rotor teeth will line up with the four stator teeth of phase B. Energizing phase C will cause the rotor to rotate counterclockwise by 15° . The principle of operation of the single-stack variable-reluctance step motor requires that some of the rotor teeth seek the minimum-reluctance path with the stator teeth of the energized phase. *See* RE-LUCTANCE.

Permanent-magnet motor. The permanent-magnet step motor (**Fig. 2**) has a rotor which is radially magnetized. Each stator section may have a single winding (unifilar winding) or a winding with a center tap, in which case the winding is called bifilar. A unifilar-wound motor must be driven by a bipolar drive; that is, the drive current must be reversed back and forth to advance the rotor. In contrast, a bifilar-wound permanent-magnet motor can be driven with a unipolar drive.

Hybrid permanent-magnet motor. This type of step motor also has a permanent-magnet rotor. Hybrid refers to the motor's operation under the combined principles of permanent-magnet and variable-reluctance



Fig. 2. Two-phase permanent-magnet step motor. (a) Rotor, magnetized with six radial poles. N and S are north and south magnetic poles. (b) Cutaway view.

step motors. The motor was originally designed as a two-phase ac synchronous motor for low-speed applications. It was found that the motor could also step incrementally if the phase windings are excited with dc pulses. The rotor of this step motor (**Fig. 3**) has a permanent magnet positioned in the axial direction to produce a north-south unidirectional magnetic field. The teeth on the two sections of the rotor are offset by one-half of a rotor tooth pitch. In general, the stator teeth are all uniformly distributed with a tooth pitch which is different than that of the rotor. The stator may have unifilar two-phase windings or bifilar windings. In the bifilar case, each phase has two windings, and the motor is often referred to as a four-phase motor.

Winding configurations. Most variable-reluctance step motors are wound with bifilar windings. Permanent-magnet and hybrid permanent-magnet step motors can be wound with either bifilar or unifilar windings.

Unifilar windings. A unifilar winding is a single coil wound on a pole of a motor. When a direct current is sent through the winding, a flux is produced in the pole if the magnetic path is closed. This produces a magnetic polarity at the pole. If the direction of the current is reversed, the direction of the flux will also be reversed. Thus, in order to reverse the polarity of the magnetic pole, with a unifilar winding, the direction of the current must be reversed. This requires a bipolar driver, that is, a driver that can reverse polarity between positive and negative.

Bifilar windings. A bifilar winding consists of two windings in opposite sense on the same pole. When a driver sends a current into one coil, a magnetic polarity is induced on the pole. By applying the same current to the other coil instead, the magnetic polarity of the pole is reversed. Thus, with a bifilar winding the magnetic pole polarity can be reversed without reversing the polarity of the drive voltage. The advantage is that only a single-ended, or unipolar, power supply is needed. A disadvantage of the bifilar winding is that only one-half of the total



Fig. 3. Hybrid permanent-magnet step motor with eight-pole stator. N and S are north and south magnetic poles. (a) Axial cross section. (b) Radial cross sections at A-A' and B-B' on part a.

windings are used at any time so that for a given frame size the motor is derated. However, bifilar windings can be connected in series or in parallel so that the entire windings can be utilized. Since the operations of permanent-magnet and hybrid permanent-magnet step motors depend on the reversal of the magnetic polarities of the stator poles, these motors must have either bifilar windings with a unipolar driver or unifilar windings with a bipolar driver. *See* WINDINGS IN ELECTRIC MACHINERY.

Torque. Step motors are torque-generating devices. The torque can be classified as static and dynamic. When a phase of a variable-reluctance step motor is energized with a direct current, a static holding torque is developed. As the rotor is forced to rotate, ideally a sinusoidal static holding torque curve is generated for this phase. Similarly, torque is generated when the other two phases of a three-phase motor are energized in turn. The equilibrium positions are also the detent positions of the individual phases. In practice, the static holding torque may not be sinusoidal, depending on the design of the motor. A variable-reluctance motor must have at least three phases to rotate continuously.

The torque components of a hybrid permanentmagnet motor are complex. The permanent magnet in the rotor generates a fourth-harmonic detent torque when the stator windings are not energized relative to the torque generated when the stator windings are energized.

The pull-out torque of a step motor represents the dynamic performance of the motor. It is defined as the maximum frictional torque that can be driven by the motor at a given speed. The pull-out torque (torque-speed) characteristic of a step motor is similar in shape to that of a dc motor. *See* DIRECT-CURRENT MOTOR; TORQUE.

Drivers. A driver stage for a variable-reluctance step motor has a power transistor for each motor phase. To energize a certain motor phase, a logic sequencer simply sends out a signal to turn on the corresponding power transistor so that a current supplied by the power source will flow through the motor phase windings. To drive the motor in a particular sequence, the phases are energized accordingly one at a time. The phases can also be energized two at a time.

The unipolar driver for a variable-reluctance step motor can also be used for permanent-magnet and hybrid permanent-magnet step motors with bipolar windings. As noted above, a bipolar driver is required for unifilar motor windings.

The power transistors of the step motor driver can be controlled by a chopper driver to control the average current to the motor windings. Typically, the motor is switched to a high supply voltage, several times the rated value, causing the current to build up rapidly. A current-sensing network turns the high voltage to a low voltage when the current reaches some preset threshold. The frequency of chopping can range from 1 to 30 kHz. *See* CHOPPING; MOTOR; SERVOMECHANISM. Benjamin C. Kuo

Bibliography. P. P. Acarnley, *Stepping Motors: A Guide to Modern Theory and Practice*, 2d ed., 1984; T. Kenjo and A. Sugawara, *Stepping Motors and Their Microprocessor Controls*, 2d ed., 1995; B. C. Kuo (ed.), *Proceedings of the Incremental Motion Control Systems and Devices Symposium*, 1972-1995; B. C. Kuo (ed.), *Step Motors and Control Systems*, 1979.

Stereochemistry

The study of the three-dimensional arrangement of atoms or groups within molecules and the properties which follow from such arrangement. Molecules that have identical molecular structures (that is, the same kind, number, and sequential arrangement of atoms) but differ in the relative spatial arrangement of component parts are stereoisomers. Inorganic and organic compounds exhibit stereoisomerism. Examples are structures (1) to (8). In the example, pairs





of stereoisomers are related by permutation (transposition) of bonded atoms or groups. In structures (3) and (4), a single permutation of Cl and H about either C atom produces two stereoisomers. Atom C_2 may be called a stereogenic atom (or a stereocenter).

Significance. Stereochemistry has played a significant role in the historical development of theories of molecular structure. The optical activity of substances such as sugar and turpentine was discovered by J. B. Biot in 1813, and L. Pasteur was the first to separate or resolve enantiomers (nonsuperimposable mirror images) from one another (1848). While Pasteur recognized that enantiomers must differ in symmetry, it was not until 1874 that J. H. van't Hoff and J. A. LeBel were able to relate symmetry, structure, and properties into the hypothesis of the tetrahedral carbon which formed the cornerstone of modern stereochemistry. The foundation of inorganic stereochemistry rests on the coordination theory of A. Werner (1893).

Since the latter part of the nineteenth century, stereochemical studies have been prominent in the evolution of both inorganic and organic chemistry. These studies have concerned themselves with the determination of configuration, the interconversion of diastereomers and the assessment of their energy differences, and conformational analysis.

Methods for the analysis and separation of stereoisomers have been devised, and stereochemical principles have been applied to the elucidation of reaction mechanisms, to the development of asymmetric syntheses, and to attempts to understand biological processes at the molecular level. *See* ASYMMETRIC SYNTHESIS; COORDINATION CHEM-ISTRY; RACEMIZATION.

Symmetry. The nature and the number of stereoisomers of a molecule are determined by the permutation of atoms or groups (called ligands) at stereocenters as well as by the symmetry of the molecule. The symmetry elements to be considered are: planes of symmetry, axes of symmetry, centers of symmetry, and rotation-reflection (or alternating) axes of symmetry. Two types of stereoisomers are known. Those such as (7) and (8), which are devoid of reflection symmetry—which cannot be su-

perposed on their image in a mirror—are called enantiomers. All other stereoisomers, such as the pairs (1)-(2), (3)-(4), and (5)-(6), are called diastereomers. The configuration of a stereoisomer designates the relative position of the atoms associated with a specific structure. The structures of stereoisomers (1) and (2) differ only in configuration. The same is true for (3) and (4), (5) and (7), and (7) and (8).

Enantiomers are related to one another as a left hand is to a right hand. Such structures are said to be chiral (Greek *cheir* = hand) or dissymmetric. Unlike diastereomers, which differ in most physical and chemical properties, enantiomers have identical properties other than the sign (+ or -) of their optical activity. This identical behavior exists toward all agents and processes which are themselves achiral. Chiral agents act differently toward enantiomers, however. Biological specificity toward enantiomers is dramatic; for example, the enantiomers of the amino acid leucine have different tastes: one is bitter, while the other is sweet. *See* BIOLOGICAL SPECI-FICITY; MOLECULAR ISOMERISM.

Configuration. In order to understand chemical processes involving stereoisomers, it is necessary to know the configurations of reactants and products. The configuration of stereoisomers (1) and (2) are designated cis (chlorines on the same side) and trans (chlorines on opposite sides), respectively; (3) and (5) are trans; (4) and (6) are cis. Configurations of diastereomers can often be determined from their physical properties. For example, the isomer of $C_2H_2Cl_2$ with zero dipole moment is (3), while (4) has a finite dipole moment. *See* DIPOLE; DIPOLE MOMENT.

The absolute configuration of chiral molecules such as (7) and (8) is the actual order of atoms about the stereogenic atom (tetrahedral carbon atom) which defines the specific enantiomer. While the three-dimensional picture of a molecule defines its stereochemistry just as the picture of a hand defines its handedness (right or left), it is not convenient or useful to rely on full three-dimensional representations in most cases, and a shorthand notation to designate configuration is widely accepted. Groups and atoms which surround the stereocenter are assigned priorities related by unambiguous rules to atomic number, for example, Br > Cl > F > H, and to substitution pattern. The molecule is viewed from the tetrahedral face opposite the carbon substituent with the lowest priority, that is, from the face opposite hydrogen in the case of (7). The counterclockwise descent (7a) from high- to low-priority substituent (Br followed by Cl followed by F) is designated S (Latin *sinister* = left). Enantiomer (8) exhibits a similar but clockwise order (8a) designated R (Latin



rectus = right). When the order of groups immediately bound to the stereocenter is equivocal, the priority is determined by atoms or groups attached to the atoms whose priorities are equivocal.

The same priorities may be used in the designation of the configurations of diastereomers such as (3) and (4). When the atoms or groups of highest priority lie on opposite sides of the double bond, as in (3), the stereoisomer is designated E (German *entgegen* = opposite). Isomer (4) in which the reference atoms are on the same side of the double bond are termed Z (German *zusammen* = together). This system is less ambiguous than the cis-trans system which is still applicable in simple cases.

Enantiomers are characterized by the sign of their optical activity at a given wavelength, temperature, and specified solvent. The assignment of configuration to specific enantiomers is carried out by chemical transformations of known stereochemistry, by spectroscopic means such as circular dichroism and optical rotatory dispersion, and by diffraction of single crystals employing anomalous x-ray scattering. From such studies it is known that the lactic acid isomer that is experimentally found to be levorotatory, that is, whose optical activity is (–)- or counterclockwise-rotating, has structure (**9**) corresponding



to the R configuration (*9a*). Configuration has no simple connection to the sign of the optical rotation. *See* COTTON EFFECT; DICHROISM; OPTICAL ACTIV-ITY; OPTICAL ROTATORY DISPERSION; X-RAY DIFFRAC-TION.

Two-dimensional (planar) projections of threedimensional structures are very commonly used to represent configurations. Projection formulas compress the tetrahedral geometry into the plane of the paper. In (9) the horizontal atoms or groups are forward of the plane (wedges) of the stereogenic atom, and the vertical groups are behind (solid or broken lines). Tipping formula (9) forward places the COOH group in the plane of the paper and the H atom behind, corresponding to structure (9a). The planar projection of (9) may be written as (9b). To avoid confusion in their use, the following conventions apply to planar projections: they must not be rotated in this plane through any but integral multiples of 180°; an odd number of exchanges of any pair of substituents is equivalent to transformation into its mirror image, and an even number of exchanges leaves the configuration unchanged.

The relative configurations of stereoisomers applicable to diastereomers are illustrated by the tartaric acids (10) and (11). Structure (10) represents (+)-tartaric acid (both stereocenters * have R configurations), and (11) represents *meso*-tartaric acid, the diastereomer which is optically inactive



(the two stereocenters, one R and one S, are mirror images of one another. Older configurational symbolism (D and L) involving intermolecular comparisons to reference substances such as glyceraldehyde and serine persists, but it is giving way to the R-S designations.

Stereoregular polymers with defined configurations at stereogenic atoms or at carbon-carbon double bond stereocenters differ substantially in properties from their diastereomers. Polymeric carbohydrates such as starch and cellulose and hydrocarbons such as rubber and gutta-percha exemplify natural diastereomer pairs. Synthetic polystyrene diastereomers (12) [isotactic] and (13) [syndiotactic] differ only in the configuration at stereogenic atoms. The high-melting, fiber-forming isomer is isotactic polystyrene (12).



Configurational studies are powerful probes in the elucidation of reaction mechanisms. Replacement of substituent atoms or groups in a reaction is often attended by a change in configuration at the stereocenter (Walden inversion), or it may result in the loss of optical activity (racemization). In some cases retention of configuration prevails. Such results provide evidence for the geometry of transition states or for the intervention of intermediates (ions or radicals).

Resolution. Most chiral substances occur in nature as only one enantiomer. Few natural products are represented in nature by both enantiomers, and few also exist as mixtures with equal proportions of enantiomers called racemates. On the other hand, synthesis of the chiral substances in the laboratory or in industry in the absence of chiral agents or conditions results in racemates which exhibit no measurable optical activity, since the activity of one enantiomer cancels that of the other.

The separation of one or both enantiomers from a racemate, called resolution, requires the intervention of an optically active reagent, catalyst, or other chiral influence. The most common resolution procedure involves the reversible conversion of an enantiomer mixture into a pair of diastereomers, with as widely different physical properties as practicable, by reaction with a single pure enantiomer. For example, a (\pm)-amine reacts with a (+)-acid to give two diastereomeric salts: (+)-RNH₃[⊕](+)-R'COO^{\ominus} and (–)-R-NH₃[⊕](+)-R'COO^{\ominus}.

These diastereomeric salts may be separated by fractional crystallization and may then be converted to the individual enantiomeric amines by cleavage with a strong base. The (+)-acid resolving agent may be recovered and reused. Covalent diastereomers may also be separated by crystallization or increasingly by gas, liquid, or thin-layer chromatography.

Kinetic resolution takes advantage of differences in rates of reaction of enantiomers with chiral reagents. Enzymes are chiral catalysts which can preferentially catalyze reactions of one enantiomer. Enzymatic resolutions play a major role in the syntheses of substances of biochemical importance. Chromatographic resolution on chiral stationary phases, which bind or dissolve one enantiomer more strongly than the other, constitutes another useful type of resolution.

In resolution by preferential crystallization or entrainment, one enantiomer preferentially crystallizes upon seeding of supersaturated solutions with seed crystals of that enantiomer. Though relatively few racemates are amenable to this type of resolution, it is nevertheless of considerable importance on an industrial scale. *See* CHROMATOGRAPHY; CRYSTALLIZA-TION.

Conformational analysis. Molecules are not rigid collections of atoms. Torsional stereoisomerism arises as a consequence of the rotation (torsion) of atoms about bonds within molecules. This gives rise to stereoisomers called conformers which are interconvertible by rotation about carbon-to-carbon single bonds. The existence of two extreme forms of ethane (in terms of energy), (14) [staggered] and (15) [eclipsed], was predicted by K. S. Pitzer in



1936 and later experimentally verified. An infinite number of forms (conformations) intermediate between the equilibrium (low-energy) conformation (14) and the higher-energy conformation (15) is possible. At room temperature most ethane molecules resemble (14). Though their presence is experimentally readily demonstrated, separation of conformers is normally not possible since the small energy difference between them [3 kcal/mole; 12.6 kilojoules/mole in the case of (14), and (15) makes their interconversion facile]. The ease of interconversion explains why some enantiomeric structures, such as the *gauche* conformers (large groups, CH₃, 60° apart) of *n*-butane (16), are incapable of resolution.



The relationship between the physical and chemical properties of substances and their preferred conformations (conformational analysis) has been the subject of many studies since 1950. Cyclohexane (17) exists overwhelmingly in the form of conformer (18; the chair form), which is the structural subunit



from which the carbon atom lattice of diamond is constructed. Substituents on six-membered rings are designated as either axial or equatorial, as in (**18**). Conformational analysis of synthetic hydrocarbon polymers such as polypropylene and biopolymers such as deoxyribonucleic acid (DNA) has led to an understanding of their properties. Much of the biological activity of enzymes, for example, is made possible by the specific conformations adopted by these macromolecules. *See* CONFORMATIONAL ANAL-YSIS; ENZYME; POLYMER.

Stereocontrolled synthesis. An understanding of the stereochemical consequences of chemical reactions and the determination of configuration of complex natural products has permitted the total synthesis of many stereochemically complex substances since around 1940. For example, cholesterol and gibberellic acid each contain eight stereogenic carbon atoms, which makes possible, in principle, $2^8 = 256$ stereoisomers (128 racemates). Disregard for stereochemical consequences during synthesis would lead to extremely complex mixtures requiring repeated separation and to very low yields.

Living organisms unerringly synthesize (biosynthesis) just one of these isomers in fully stereoselective processes. Laboratory and commercial syntheses of substances which have desirable biological properties have increasingly been devised so as to mimic biosyntheses by taking advantage of stereocontrolled reactions. These are of two types: stereospecific reactions and stereoselective reactions. Stereospecific reactions are those in which different stereoisomers are transformed into stereochemically different products. For example, Z-2-butene (19) [cis] is epoxidized to the *cis*-epoxide (20) only.



The diastereomeric *E*-2-butene [trans] forms only the *trans*-epoxide.

Stereoselective reactions are those in which a single reactant gives rise to a mixture of diastereomers in which one predominates. Those stereoselective reactions in which new stereocenters are formed in unequal amounts are called asymmetric syntheses. Transformations of optically active compounds into more complex substances which incorporate additional stereocenters are also asymmetric syntheses.

Syntheses of natural products which originate in achiral starting materials require at least one resolution step if nonracemic products are desired. Alternatively, a total synthesis may well originate in a single enantiomer of a compound having but one or two stereogenic atoms. Additional stereocenters are introduced during the synthesis through stereospecific and especially through stereoselective steps. Separation of diastereomers at various stages may be required. *See* STEREOSPECIFIC CATA-LYST. Samuel H. Wilen

Bibliography. E. L. Eliel, S. H. Wilen, and M. P. Doyle, *Basic Organic Stereochemistry*, 2001; E. Eliel, S. Wilen, and L. Mander, *Stereochemistry of Organic Compounds*, 1994; R. E. Gawley and J. Aube, *Principles of Asymmetric Synthesis*, 1996; M. M. Green et al. (eds.), *Topics in Stereochemistry: Materials-Chirality*, 2003; K. Mislow, *Introduction to Stereochemistry*, 2003; D. G. Morris, *Stereochemistry*, 2002; M. North, *Principles and Applications of Stereochemistry*, 1998.

Stereophonic radio transmission

The transmission of stereophonic audio signals over broadcast radio systems, including terrestrial analog amplitude-modulation (AM), frequency-modulation (FM) systems, and digital radio systems (terrestrial and satellite). *See* STEREOPHONIC SOUND.

Analog systems. Two-channel stereo broadcasting of audio was first implemented for FM radio signals in the 88-108-MHz frequency band (allocated for FM radio broadcasting) in the 1960s, and became the dominant mode of transmission, in particular for music programming, in the 1970s. A multiplexing method was used to accomplish this and allowed the signal transmitted by a stereo broadcast station to be decipherable by either a stereo receiver or a monophonic receiver. For the monophonic receiver to receive a stereo broadcast, some component of the broadcast must be a single signal that includes information from both the left and the right channels. This is called the L + R signal. A monophonic receiver simply demodulates the L + R signal and delivers it to the listener.



Baseband spectrum of an FM stereo signal. This audio and supersonic spectrum is frequency-modulated onto a radio-frequency carrier.

The stereo receiver must use the L + R signal and other transmitted information to demodulate separate L and R signals. For this, stereo broadcasts include an L - R signal. In the stereo receiver, after detection of the L + R and L - R signals, the L and R signals are separated as shown in Eqs. (1) and (2).

$$(L+R) + (L-R) = 2L$$
 (1)

$$(L+R) - (L-R) = 2R$$
 (2)

The subtraction used in Eq. (2) is realized through phase shifting and addition. In FM stereo transmissions, the L + R signal is directly frequencymodulated onto the radio-frequency carrier, assuring compatibility with monophonic systems (see **illustration**). The L - R signal is used to amplitudemodulate a 38-kHz hypersonic subcarrier. The 38kHz carrier is suppressed, creating a double-sideband (DSB), suppressed-carrier AM signal at 38 kHz. This hypersonic signal is then frequency-modulated onto the radio-frequency carrier. A 19-kHz pilot signal is also transmitted to indicate that the broadcast is stereo and to provide an accurate frequency reference for demodulation of the 38-kHz L - R signal. See AMPLITUDE MODULATION; FREQUENCY MODULA-TION.

In the FM stereo receiver, the entire baseband signal that has been frequency-modulated onto a radio-frequency carrier (L + R, pilot, and doublesideband L - R) is first detected with an FM discriminator. The L + R signal is isolated by a lowpass filter. The pilot tone is doubled and mixed with the double-sideband suppressed-carrier L – R signal, so that the L - R signal can be demodulated by an AM envelope detector. The L + Rand L - R audio signals are then passed through a network that implements the functions expressed in Eqs. (1) and (2). The resulting left- and rightchannel audio signals are amplified and presented to the listener. See AMPLITUDE-MODULATION DETEC-TOR; ELECTRIC FILTER; FREQUENCY-MODULATION DE-TECTOR; FREQUENCY-MODULATION RADIO.

It is also possible to send stereo audio signals using more than two channels over FM radio transmissions, although this technology is not standardized for use in the United States by the Federal Communications Commission (FCC), as is the two-channel technology just described. SRS Circle Surround, Neural Audio, and Dolby Laboratories have all developed proprietary surround sound technologies which are compatible with two-channel stereo FM broadcasts and which provide up to "6.1" channels of surround sound information (the ".1" represents a seventh, low-frequency channel with frequency response limited to approximately 100 Hz).

Stereo broadcasting for the AM band (550-1701 kHz) was first authorized by the FCC in 1982, when five different competing (and incompatible) systems were authorized because the FCC was unwilling to select just one and chose instead to let the marketplace decide. This proved to be a fateful decision since over the next 11 years the various AM stereo systems competed for market share. In the end, due in part to the confusion created by so many systems being available, AM stereo never became a popular service. In 1993, the FCC revisited their decision and finally selected a single system, called compatible quadrature amplitude modulation (C-QUAM), but receiver manufacturers and broadcasters never fully embraced this technology. In 2002 the FCC authorized a new digital radio service for the AM band, called in-band/on-channel (IBOC) digital radio, which supports stereo audio but is incompatible with C-QUAM, effectively ending the analog AM stereo saga in the United States.

Digital systems. A number of digital radio systems are being implemented around the world, and in most cases these systems support the transmission of not only two-channel stereo but also "surround sound" stereo signals with five or more channels. In 2005, satellite digital audio radio service (SDARS) provider XM Radio began broadcasting some of its music channels in 5.1-channel surround using technology developed by Neural Audio, and a number of terrestrial FM radio broadcasters were also experimenting with surround sound using the IBOC digital radio broadcast system authorized for the FM band by the FCC in 2002. David H. Layer; Jeffrey Ward

Bibliography. Federal Communications Commission, *Rules and Regulations*, Parts 73 and 74, and *Amendments*, 2004; D. G. Fink and W. Beaty (eds.), *Standard Handbook for Electrical Engineering Tutorial for Non-Engineers*, NAB and Focal Press, 2005; *NAB Engineering Handbook*, 9th ed., 1999; M. E. Van Valkenburg (ed.), *Reference Data for Engineers: Radio, Electronics, Computer, and Communications*, 9th ed., 2002.

Stereophonic sound

A system of sound recording or transmission in which signals are captured, mixed, or synthesized using two or more audio channels in such a way as to deliver a spatial or three-dimensional auditory impression to a listener when these audio channels are connected to loudspeakers in a listening room. Binaural or "head related" stereo is a term used to describe the capture or synthesis of a pair of signals that closely resemble those present at a listener's ears in a natural spatial sound field. Such signals are intended for reproduction using a pair of headphones worn by the listener, although they can be reproduced using loudspeakers if suitable signal processing is employed. *See* BINAURAL SOUND SYSTEM.

Stereo recording techniques were patented in the 1930s, but commercial development was not undertaken at that time. In the late 1930s and early 1940s, a few motion picture films were produced using two optical tracks for stereo playback. A variety of multichannel film systems were introduced in the 1950s. The commercial recording industry introduced twotrack stereo tape recordings in 1954 and singlegroove, two-channel stereo disks in 1957. The first stereo broadcasts on FM radio were made in 1961. In the 1970s, several quadraphonic (four-channel) systems were introduced, but none achieved lasting commercial success. From the late 1990s, "surround sound" stereo systems involving five or more channels began to achieve commercial success, primarily as an adjunct to home television systems. Wavefield synthesis systems, which reached the market in the early years of the twenty-first century, use very large numbers of loudspeakers in an attempt to recreate acoustical wavefronts in a physically accurate fashion. See STEREOPHONIC RADIO TRANSMIS-SION

Auditory cues in stereophony. The aim of stereophonic recording and reproduction is to deliver consistent auditory spatial cues to the listener to give rise to an intended spatial impression. It is nonetheless common for stereophony to rely on a degree of auditory illusion for this purpose, as it is not usually possible to reconstruct acoustical sound fields in a physically accurate manner over a wide listening area when using a limited number of loudspeakers.

In order to determine the origin (location) of a sound, the brain considers the differences in sound pressure level and arrival time at each ear. The external ear (pinna) also gives rise to modifications to the frequency spectrum of sound signals, which are unique to each source location and can represent height information as well as horizontal cues. Fluctuations in these cues contribute to a sense of spaciousness or width. *See* HEARING (HUMAN).

Stereophonic signals are created either by capturing the relevant cues using two or more microphones, or by artificially introducing the necessary interchannel differences using audio signal processing. The perceived source images that result from these cues during reproduction are known as phantom images because they are not real sources. In binaural stereo or wavefield synthesis, they are sometimes known as virtual sources.

Stereophonic sound pickup. There are three basic microphone techniques for two-channel stereophonic pickup: coincident, spaced, and individual instrument (also called close miking). The techniques are sometimes combined.

The coincident technique uses two directional microphones located very close together, often enclosed in the same case. The coincident technique relies only on capturing differences in sound level between the channels to encode different source positions. This can give rise to very precise stereo imaging. *See* MICROPHONE.

Spaced microphone techniques encode spatial cues using both time and level differences between the channels. These differences lead to directional perception based either on the "precedence effect," whereby the sound is localized toward the earlier loudspeaker signal, or on "summing localization," in which the perceived source position is the result of a trade-off between the two cues. Summing localization tends to work with interchannel time delays up to just over 1 millisecond, combined with interchannel level differences of up to 18 dB. Techniques can rely on large time differences and small level differences, or vice versa, but the quality of the stereo image resulting from different combinations is not identical. Techniques relying on summing localization tend to be known as "near-coincident" techniques because of the relatively small microphone spacings involved. Because of the time delay, spaced techniques can cause noticeable phase cancellations in monophonic reproduction.

Since the late 1960s, it has become common for much stereo program material to be recorded using a large number of microphones, each located close to one source or a small group of sources. By proportionally mixing the microphone outputs between the stereo channels, it is possible to provide the interchannel level differences necessary for perceiving phantom images. Artificial effects and reverberation may be added subsequently to create a spatial impression, or signals from microphones located in the reverberant field of a studio may be mixed with the dry signals.

Surround sound recording extends these concepts to more than two channels to create an immersive sound field. However, it is common to find that accurate stereophonic imaging is attempted only in front of the listener, using techniques similar to those described above, while the rear channels are used to enhance the experience of spaciousness and envelopment. This is particularly true of systems based on the International Telecommunications Union-Radiocommunication Sector document ITU-R BS.775 layout, known as 3-2 stereo or 5.1 surround, described below. *See* SOUND RECORDING; SOUND-RECORDING STUDIO.

Stereophonic reproduction. The ability to create convincing stereophonic reproduction is highly dependent upon the performance of the loudspeakers, the acoustics of the listening room, and the hearing process of the listener.

The loudspeakers should be identical and capable of broad frequency response with low distortion. Imaging is degraded by narrow or irregular dispersion of frequencies above about 500 Hz. Most high-quality loudspeaker systems have a horizontal coverage angle of 90 to 120° , above 500 Hz. The vertical angle is usually less, and often is 40 to 60° . The uniformity of the coverage pattern is an important measure of the loudspeakers' quality. All loudspeakers have wider coverage at low frequencies. *See* LOUD-SPEAKER.

For two-channel stereo, the ideal arrangement is one where the loudspeakers subtend an angle of $\pm 30^{\circ}$ at the listening position. This arrangement is maintained in the ITU BS.775 configuration for 3-2 Stereo. (5.1 surround), in which a center loudspeaker is added between the left and right front loudspeakers, as well as two rear-side loudspeakers at $\pm 110^{\circ}$. The rear-side locations were chosen as a compromise between the requirements for movie applications and those for music reproduction. In movie applications, the capacity to deliver effects from the rear is desirable, whereas for music reproduction the research evidence suggests a need to recreate lateral reflections to optimize envelopment.

The loudspeakers should be located in similar acoustical surroundings. The surfaces near each loudspeaker should be similar in acoustical quality in order that sound reflected from the area of each loudspeaker will be the same for the listener. The reverberation time of the listening room should ideally be less than about 0.5 s. Hard surfaces near the loudspeakers can reflect sound energy, which when delayed only a few milliseconds by its slightly longer path will combine with the direct sound in a frequency-dependent manner to produce irregularities in what should be a smooth frequency response at the listener's position. This "comb filter" effect can be reduced or avoided by making all surfaces around the loudspeakers acoustically very absorbent. In some professional listening rooms, such as recording or broadcast control rooms, this surface treatment extends from the loudspeakers to the listener's position; all wall surfaces behind the listener are made reflective and diffusive.

A modern high-quality loudspeaker system consists of two or more radiating units in the same enclosure, each reproducing only a part of the sound spectrum. These radiators cannot coexist in the same location; therefore, there can be path length differences (hence, time differences) between them and the listener. Because the hearing process is sensitive to short time delays, phantom images of sound sources being partly reproduced by two or more radiators in each loudspeaker will tend to smear or split in space. However, even among trained listeners, not all people can hear, or are distracted by, this effect. For those who are annoyed, several manufacturers have introduced time- (or phase-) aligned loudspeakers. Time alignment requires that path lengths be made equal by physical location of the radiators or by use of an electronic delay in order to synchronize arrival times at the listener's position. See SOUND-REPRODUCING SYSTEMS.

Rollins Brook; Francis Rumsey Bibliography. B. and J. Barlett, *Practical Recording Techniques*, 4th ed, Focal Press, 2005; J. Eargle (ed.), *Stereophonic Techniques: An Anthology*, Audio Engineering Society, 1986; D. Howard and J. Angus, *Acoustics and Psychoacoustics*, 3d ed, Focal Press, 2006; D. M. Huber, *Modern Recording Techniques*, 6th ed., 2005.
Stereoscopy

The phenomenon of simultaneous vision with two eyes, producing a visual experience of the third dimension, that is, a vivid perception of the relative distances of objects in space. In this experience the observer seems to see the space between the objects located at different distances from the eyes. The stereoscopic effect is unique and cannot be easily described to one who does not possess it. Stereopsis, or stereoscopic vision, provides the individual with the most acute sense of relative depth and is of vital importance in visual tasks requiring the precise location of objects.

Stereopsis is believed to have an innate origin in the anatomic and physiologic structures of the retinas of the eyes and the visual cortex. It is present in normal binocular vision because the two eyes view objects in space from two points, so that the retinal image patterns of the same object points in space are slightly different in the two eyes. The stereoscope, with which different pictures can be presented to each eye, demonstrates the fundamental difference between stereoscopic perception of depth and the conception of depth and distance from the monocular view. In the **illustration**, each of the two eyes views a pair of vertical lines A and B drawn on cards. The separation of these lines for the right eye is greater than that for the left eye. If the difference in separation is not too great, the images of the lines fuse when the two targets are observed by the two eyes. There is almost an immediate stereoscopic spatial experience that the two lines are located in space, line B being definitely more distant than line A. No hint of this spatial experience occurs in the observation of each target alone. This difference between the images in the two eyes provides the stimulus for emergence of the stereoscopic experience.

Stereoscopic acuity and the presence of stereopsis may be tested by several methods or instruments such as the ordinary hand stereoscope and similar devices, vectograph targets, the Howard-Dolman peg test, the Verhoeff stereopter, the Hering falling bead test, pins with colored heads stuck in a board to test for near vision, and afocal meridional magnifying lenses placed before one eye while the subject



Diagram of stereoscopic vision.

views a special field (leaf-room or tilting table). *See* VISION. Kenneth N. Ogle

Bibliography. C. H. Graham (ed.), Vision and Visual Perception, 1965; W. L. Gulick and R. B. Lawson, Human Stereopsis: A Psychological Approach, 1976; E. Hering, The Theory of Binocular Vision, 1977; K. N. Ogle, Optics: An Introduction for Ophthalmologists, 2d ed., 1976; K. N. Ogle, Research in Binocular Vision, 1950, reprint 1964; R. W. Reading, Binocular Vision: Foundations and Applications, 1983; H. Solomon, Binocular Vision, 1980.

Steric effect (chemistry)

The influence of the spatial configuration of reacting substances upon the rate, nature, and extent of reaction. The sizes and shapes of atoms and molecules, the electrical charge distribution, and the geometry of bond angles influence the courses of chemical reactions. The steric course of organochemical reactions is greatly dependent on the mode of bond cleavage and formation, the environment of the reaction site, and the nature of the reaction conditions (reagents, reaction time, and temperature). The effect of steric factors is best understood in ionic reactions in solution. The nucleophilic substitution reaction at a saturated carbon atom can serve as an illustration.

Saturated nucleophilic substitution. While the reacting carbon atom is in an electron-deficient state in the transition state (state of highest energy, somewhere between starting material and product) in a nucelophilic substitution process, the reaction can be varied from a two-step, unimolecular ionization to a one-step, bimolecular transformation. The former mode of reaction, a solvolysis or S_N 1 process, converts a tetrahedral carbon into a solvated planar carbonium ion intermediate, and hence leads to racemization (randomization of configuration); for example, reaction (1) where the R terms represent organic groups.



See REACTIVE INTERMEDIATES.

The second reaction path proceeds by a simultaneous rupture of the old bond and creation of a new one, either by inversion of configuration, a Walden inversion or $S_N 2$ process, for example, reaction (2);



or in a few cases by a front-side displacement of part of the substituent already present, an S_N *i* process, and hence, retention of configuration,

for example, reaction (3).

$$ROH \xrightarrow{SOCI_2} ROSOCI \longrightarrow R^{\odot}OSOCI^{\odot} \xrightarrow{-SO_2}$$
(ion pair)
$$R^{\odot}CI^{\odot} \longrightarrow RCI \quad (3)$$
(ion pair)

These substitution reactions are highly solventdependent; for example, tert-butyl chloride undergoes solvolysis close to 500,000 times faster in water, a solvent of high ionizing power, than in ethanol, a solvent of low dielectric properties. The nature of R, R', and R" is one of the factors determining which of the above pathways a compound prefers for its substitution. The larger the size of these three groups, the greater is the tendency to relieve steric strain by extrusion of X, and thus the more the need for an S_N 1 process. The smaller the size of the environment, the greater is the accessibility of the reagent from the backside, and thus the more tendency toward an $S_N 2$ process. As a consequence, tertiary compounds undergo racemization readily, whereas primary systems prefer inversion.

Both the rate and the steric course of an ionic displacement may depend often on the ability of groups adjacent to the reaction site to accommodate a positive charge. Substitution of α -halo ethers and allyl or benzyl halides occurs much faster than a similar reaction of unsubstituted halides because of the intermediacy of the stabilized cations or cationlike transition states in notation (4).

In view of the charge distribution over more than one atom in these cases, sometimes the incoming reagent forms a bond at a site different from that of the leaving group. As a consequence, an $S_N 2'$ process, reaction (5), or an $S_N i'$ path, reaction (6), may result. Both processes yield retention of configuration; that is, the orientation of the new substituent on its carbon atom is identical with that originally held by the former functional group on a site that is two carbon atoms removed.

In the presence of participating neighboring groups, even solvolyses can lead to retention of configuration. Certain rigidly held homoallyl systems undergo substitution at a site that is three carbons removed from the position of the leaving group, but with retention of configuration, as represented by reaction (7).

Double inversion is responsible for the retained configuration of the products of solvolysis of α -halo acid salts, reaction (8).





Solvolysis of *trans-\beta*-acetoxy systems in nonaqueous media leads to trans products; in the presence of water, cis compounds are obtained, reactions (9), where Ac = acetyl group (CH₃CO—).

Base treatment of *trans*-halohydrins leads to *trans*-vicinal glycols. The intermediate epoxide is isolable. Although ring opening of the latter may yield two different trans products, the diaxial one is formed preferentially in cycloalkane cases, reaction (10).

Rearrangement. Migration of neighboring groups toward the reaction site, resulting in skeletal rearrangements, is a common occurrence. Both the internal displacement of the leaving group by the migrating group and the subsequent external displacement of the migrating group by the solvent or added reagent proceed in a trans sense, that is, by back-side approach. Thus, the overall steric consequence of one migration sequence is retention of configuration.

There are several examples of 1,2-hydride migration, for example, reaction (11). Transannular hydride shifts are quite similar in nature, as in reaction (12).

The 1,2 migration of phenyl groups proceeds by way of fairly stable phenonium ions, reaction (13).

The Wagner-Meerwein rearrangement of saturated neighboring groups shows directional effects similar to those of the above migrations; for example, the conversion of camphene hydrochloride to bornyl chloride is stereospecific, with retention of configuration, reaction (14).

Neopentyl halides solvolyze to tertiary amyl derivatives, reaction (15).

Cyclohexane systems with equatorial leaving groups may undergo contraction to five-membered rings, reaction (16).

Organic compounds possessing potential leaving groups at the bridgehead of small bicyclic ring systems undergo substitution processes only sluggishly. In the absence of ready access at the backside of the reaction center, the $S_N 2$ pathway is excluded. The inability of the compounds to form planar carbonium ions precludes a $S_N 1$ route. However, displacements





do occur slowly at elevated temperatures, presumably via nonplanar cations.

Unsaturated nucleophilic substitution. Nucleophilic substitution reactions at unsaturated carbon atoms can take place by two possible mechanistic routes, an elimination-addition process and an additionelimination scheme. The former route is best illustrated by the transformation of aromatic halides into anilines, reaction (17).

The latter is encountered in the interconversion of carboxylic acids and their derivatives, for example, reaction (18). Because the central carbon atom has greater steric requirements in the reaction intermediate than in the starting material, the reaction velocity is strongly dependent on the size and number of neighboring groups; that is, an increase in the bulkiness of R is reflected in a decrease of the rate of the reaction.

The addition-elimination mechanism is portrayed also by the aromatic nucleophilic substitution reaction, for example, (19). In order to be able to stabilize the reaction intermediate, the all-important nitro group must be coplanar with the benzene ring. As a consequence, ortho substituents, which





may block this steric requirement, retard the reaction rate. Unusual aromatic nucleophilic substitutions have been observed in cases where steric hindrance by ortho substituents has prevented addition to aromatic ketones to occur, for example, reaction (20).

Addition reactions. The steric course of addition reactions at unsaturated sites depends largely on the reagent. Catalytic hydrogenation, a nonhomogeneous process of undetermined mechanism, occurs in a cis manner. In the absence of any steric interference, it leads to thermodynamically stable products. In the presence of steric hindrance, the two new hydrogen atoms are usually introduced on the least hindered side of the unsaturated compounds. However, sometimes some bulky polar groups actually aid, rather than retard, adsorption of the catalyst on their side of the reducing compound, thereby leading to products of opposite configuration.

The oxidation of olefins to vicinal glycols by per-

manganate salts or osmium tetroxide also proceeds in a cis fashion and involves the least-hindered side of the reacting substrate. The Diels-Alder reaction behaves similarly, for example, reaction (21). All addition processes, during which two new bonds are formed more or less simultaneously, yield cis adducts. *See* DIELS-ALDER REACTION.

Ionic addition reactions of olefins and acetylenes occur in a trans manner. The mode of addition is such as to lead to product via the most stable cations (Markovnikoff addition), for example, reaction (22).

Halogen addition to cyclic olefins leads to trans diaxial dihalides which, on standing, isomerize to the more stable trans diequatorial dihalides, reaction (23).

Ionic addition reactions of carbonyl compounds follow a steric course very similar to those of olefins. However, the reagents are mostly nucleophilic, and some reactions are equilibrium processes, for example, reaction (24). The orientation of attack and the



$$-C = C - C = 0 \xrightarrow{H^{\oplus}} \left[-C = C = C = C = 0 \xrightarrow{B^{\oplus}} 0 \xrightarrow{B^{\oplus} 0 \xrightarrow{B^{\oplus}} 0 \xrightarrow{B^{\oplus} 0 \xrightarrow{B^{\oplus}} 0 \xrightarrow{B^{\oplus}} 0 \xrightarrow{B^{\oplus} 0 \xrightarrow{B^{\oplus}} 0 \xrightarrow{B^{\oplus} 0 \xrightarrow{B^{\oplus}} 0$$



reaction rate are governed by the environment of the carbonyl group (C=O).

Addition reactions of conjugated carbonyl systems can occur through cation as well as anion intermediates, but they uniformly place the nucleophilic part of the reagents on the β -carbon atoms, for example, reactions (25) and (26).

The reaction of carbonyl compounds as enol or enolate anions with electrophilic reagents can take two different courses. If the process is kinetically controlled, the electrophile, a proton, halonium ion, or others, interacts with the substrate on its least-hindered side. However, if the reaction is thermodynamically controlled, then, independent of mechanism, the most stable product is obtained.

Elimination reactions. Elimination reactions can be carried out by pyrolysis of esters, halides, or amine oxides in the liquid or vapor phase. These eliminations always involve a rupture of vicinal cis bonds.

Alternatively, similar cleavage processes can be made to occur ionically in solution, in which case they proceed in a trans fashion. The direction of elimination depends greatly on the molecularity of the process, as well as on the sizes of the leaving group and the attacking base. The two-step, unimolecular cleavage, an E1 process, leads predominantly to the more substituted, hence more stable, olefins, for example, reaction (27).

The one-step bimolecular elimination (an E2 pro-

cess) of neutral compounds yields similar products, for example, reaction (28). However, an *E*2 reaction on positive ions, ammonium or sulfonium salts, affords the less substituted olefin in preponderant yield, reaction (29).

Reactions leading to the more stable products are said to follow the Saytzeff rule, whereas those yielding less stable olefins obey the Hofmann rule. Because the transition state in the *E*2 reaction is of lowest energy when all atoms involved in the elimination are in a plane, the fastest rates among cyclic compounds are encountered in the cases which permit a diaxial alignment of vicinal substituents.

Many ionic elimination reactions are known which involve the rupture of more than two bonds, for example, reaction (30).

$$\overrightarrow{\text{Br}-\text{CH}_2-\text{CH}_2-\text{CH}_2-\text{CH}_2-\text{Br}} \xrightarrow{:\text{Zn}} 2\text{CH}_2 = \text{CH}_2$$
(30)

The E2' processes, eliminations of two groups of carbon atoms separated by an olefinic linkage, appear to be cis in nature, reaction (31).



Electrophilic substitution. Steric factors have a fair control over the course of the aromatic electrophilic substitution reaction. In reactions of compounds containing ortho-para directing substituents, the *ortho/para* product ratio is usually greater than 1, and increases with the size of the substituent and that of the reacting species. The rate-accelerating participation of electrondonating groups, located ortho or para to the incoming substituent, in stabilizing the transition state is greatly diminished in the presence of bulky ortho neighbors which would prevent the groups from attaining coplanarity with the benzene, for example, reaction (32).



See CHELATION; CONFORMATIONAL ANALYSIS; OR-GANIC SYNTHESIS; ORGANIC REACTION MECHANISM; PHYSICAL ORGANIC CHEMISTRY; STEREOCHEMISTRY. Ernest Wenkert

Bibliography. A. T. Balaban et al. (eds.), *Steric Fit* in *Quantitative Structure-Activity Relations*, 1980; F. A. Carey and R. J. Sundberg, *Advanced Or ganic Chemistry*, pt. A: *Structures and Mechanisms*, pt. B: *Reactions and Synthesis*, 4th ed., 2000; E. L. Eliel and S. H. Wilen, *Stereochemistry of Organic Compounds*, 1994; J. E. McMurry and E. E. Simanek, *Fundamentals of Organic Chemistry*, 6th ed., 2006; G. Naray-Szabo and K. Simon (eds.), *Steric Aspects of Biomolecular Interactions*, 1987; S. D. Peyerimhoff (ed.), *Interactions in Molecules: Electronic and Steric Effects*, 2003.

Sterilization

An act of destroying all forms of life on and in an object. A substance is sterile, from a microbiological point of view, when it is free of all living microorganisms. Sterilization is used principally to prevent spoilage of food and other substances and to prevent the transmission of diseases by destroying microbes that may cause them in humans and animals.

Microorganisms can be killed either by physical agents, such as heat and irradiation, or by chemical substances. Regardless of the manner in which they are killed, they generally die at a constant rate under specified environmental conditions. If the logarithm of the number of survivors is plotted against time, the resulting curve is a straight line.

When testing a substance for sterility, care must be taken to employ appropriate techniques. A bacterial cell is considered to be killed when it is no longer capable of reproducing itself under suitable environmental conditions. If an inadequate medium is employed to subculture the treated bacteria, the substance being tested may be wrongly considered to be sterile.

By far the most resistant of all forms of life, to both physical and chemical killing agents, are some of the bacterial endospores. If they did not exist, sterilization of such materials as bacteriological media and equipment, hospital supplies, and canned foods would be much simpler.

Heat sterilization. This is the most common method of sterilizing bacteriological media, foods, hospital supplies, and many other substances. Either moist heat (hot water or steam) or dry heat can be employed, depending upon the nature of the substance to be sterilized. Moist heat is also used in pasteurization, which is not considered a true sterilization technique because all microorganisms are not killed; only certain pathogenic organisms and other undesirable bacteria are destroyed. *See* PASTEURIZA-TION.

Moist-heat sterilization. Some bacterial endospores are capable of surviving several hours at 212° F (100° C). Therefore, for moist-heat sterilization, an autoclave, pressure cooker, or retort, with steam under pressure, is required to achieve higher temperatures (**Fig. 1**).

Most bacteriological media are sterilized by autoclaving with steam at 250° F (121° C) with 15 lb pressure (10^{2} kilopascals) for 20 min or more, depending upon the volume of material being heated. Some spores are capable of surviving moist-heat sterilization equivalent to at least 7 min at 250° F (121° C).

Steam under atmospheric pressure in an Arnold sterilizer is sometimes employed for specialized bacteriological media that are easily heat damaged. Because many bacterial spores survive this treatment, it is obviously inadequate to ensure sterility.

Tyndallization. The food or medium is steamed for a few minutes at atmospheric pressure on three or four successive occasions, separated by 12- to 18-h intervals of incubation at a favorable growing temperature. In theory the intervals of incubation allow any surviving bacterial spores to germinate into more heat-sensitive, vegetative cells, which then would be killed during the next heat treatment. However, spores like vegetative cells may require special conditions such as an appropriate medium, proper oxygen tension, or proper temperature to germinate and reproduce. These conditions may not be realized during the intervals between heat treatment, and no matter how often the steaming is repeated ungerminated spores may survive and eventually germinate when conditions have been changed. The survival of ungerminated spores reduces the effectiveness of this method and it has been supplanted by other methods.

Hot-air sterilization. Glassware and other heat-resistant materials which need to be dry after treatment are usually sterilized in a hot-air sterilizer. Dry sterilization requires heating at higher temperatures and for longer periods of time than does sterilization by steam under pressure. Temperatures of $320-330^{\circ}$ F ($160-165^{\circ}$ C) for at least 2 h is generally employed in hot-air sterilization. Dry heat kills the germs through

denaturation of protein which may involve oxidative processes.

Radiation sterilization. Many kinds of radiations are lethal, not only to microorganisms but to other forms of life. These radiations include both high-energy particles as well as portions of the electromagnetic spectrum. The mechanism of the lethal action of these radiations is not entirely clear. It may involve a direct energy absorption at some vital part of the cell (direct target theory), or the production of highly reactive, ionized, free radicals near some vital part of the cell (indirect effect). Bacterial endospores are relatively resistant to all types of radiation. *See* RADI-ATION BIOLOGY.

Ultraviolet radiation. Radiant energy in the ultraviolet region of the spectrum is highly bactericidal, especially at wavelengths of approximately 265 nanometers. Lamps which generate ultraviolet radiation in this region are useful for the sterilization of air and smooth surfaces. Ultraviolet rays have very low penetrative capacity, since even a thin layer of glass absorbs a high percentage of the rays. Some irradiated cells that are presumably dead may be photoreactivated with visible light.

Gamma rays. These are high-energy, electromagnetic radiations similar to x-rays. They have great penetrative capacity, and their energy is dissipated in the production of ionized particles from the material being irradiated. Radioactive isotopes, such as cobalt-60, are a common source of gamma rays (**Fig. 2**). Gamma irradiation of foods has received much attention as a means of sterilizing foods without cooking them. Sterilization requires a radiation dose of approximately 5,000,000 rads (5×10^8 ergs of energy absorbed per gram).

Cathode rays. These high-speed electrons (beta rays) may be generated with various types of electron accelerators. This type of radiation has relatively low penetrative capacity, depending upon the energy level of the emitted electron beam. Cathode rays sterilize in a manner identical to gamma rays and without significantly raising the temperature of the material being irradiated. They have received some application in the sterilization of surgical supplies and drugs and in the experimental sterilization of foods (**Fig. 3**).

Filtration sterilization. This is the physical removal of microorganisms from liquids by filtering through materials having relatively small pores. Sterilization by filtration is employed with liquid that may be destroyed by heat, such as blood serum, enzyme solutions, antibiotics, and some bacteriological media and medium constituents. Examples of such filters are the Berkefeld filter (diatomaceous earth), Pasteur-Chamberland filter (porcelain), Seitz filter (asbestos pad), and the sintered glass filter. Most of these filters are available in different pore sizes.

The mean pore size of bacteriological filters is not the only determinant in their effectiveness. The electric charge of the pore surfaces tends to adsorb the bacteria and thus prevent their passage. Most bacteria have a net negative electrical charge on their surfaces. Usually, bacteriological filters will permit



Fig. 1. Diagrammatic sketch of an autoclave. (American Sterilizer Co.)

the passage of viruses, which are then called filterable.

A Millipore filter is a specially prepared membrane molecular filter designed to remove bacteria from water, air, and other materials, for the purpose of estimating quantitatively the bacterial population. A sterile filter disk is assembled in a filtration unit, and a specified volume of water or solution is drawn through the disk which then retains the bacteria. The filter disk is removed and placed in a sterile petri dish containing an absorbent pad, previously saturated with an appropriate bacteriological medium. Upon incubation, colonies will develop on the filter disk wherever bacteria were entrapped at the time of filtering. Special differential, or selective, media can be employed for the purpose of detecting quantitatively specific types of bacteria from the original material.

Chemical sterilization. Chemicals are used to sterilize solutions, air, or the surfaces of solids. Such chemicals are called bactericidal substances. In lower concentrations they become bacteriostatic rather than bactericidal; that is, they prevent the growth of bacteria but may not kill them. Other terms having similar meanings are employed. A disinfectant is a chemical that kills the vegetative cells of pathogenic microorganisms but not necessarily the endospores of spore-forming pathogens. An antiseptic is a chemical applied to living tissue that prevents or retards the growth of microorganisms, especially pathogenic bacteria, but which does not necessarily kill them.

The death of microorganisms subjected to bactericidal substances can be expressed exponentially, in that a straight-line graph is produced when the logarithm of survivors is plotted against time. The more concentrated the chemical employed, the greater is the rate of death.

There are hundreds of chemicals that may be considered to have sterilizing or bactericidal properties, depending upon the particular use for which they are intended. A chemical may be particularly useful for one purpose but not for another. Many are widely used for sterilization or disnfection of air, water, tabletops, surgical instruments, and so on.

The desirable features sought in a chemical sterilizer are toxicity to microorganisms but nontoxicity to humans and animals, stability, solubility, inability to react with extraneous organic materials, penetra-



Fig. 2. Cobalt-60 furnace which is designed for experimental food sterilization.



Fig. 3. Van de Graaff generator principle; packaged products sterilized by means of high-intensity electron beam which penetrates packages. (*After W. M. Urbain, Food Eng., vol. 25, February* 1953)

tive capacity, detergent capacity, noncorrosiveness, and minimal undesirable staining effects. Rarely does one chemical combine all these desirable features.

Among chemicals that have been found useful as sterilizing agents are the phenols, alcohols, chlorine compounds, iodine, heavy metals and metal complexes, dyes, and synthetic detergents, including the quaternary ammonium compounds.

Chlorine. Chlorine and chlorine-containing compounds represent the most widely used group of disinfectants. Chlorine gas is often used to purify municipal water supplies.

Various compounds of chlorine, such as the hypochlorites and chloramine, have many industrial and domestic uses as disinfectants or antiseptics.

Ozone. This is a highly oxidizing gas (O_3) used as a deodorizer, and also for disinfection of air and water. It has found some use in the food fields, but effective bactericidal concentrations may be irritating and toxic to humans.

Hydrogen peroxide. This chemical (H_2O_2) has high oxidizing and bleaching qualities and is usually employed in a 3% solution for topical application and disinfection of cuts, scratches, and minor wounds. It decomposes into water and oxygen and therefore is used where no taste, odor, or toxic residues are permitted.

Volatile organic compounds. Such compounds as formaldehyde and ethylene oxide have been used for the disinfection of sickrooms occupied by patients suffering from contagious disease (terminal disinfection), and of solids that do not permit heat treatment. These volatile substances have the advantages of effective penetrative capacity and ease of removal after treatment.

Volatile organic substances are sometimes employed as bacteriostatic agents to preserve bacteriological medium constituents until they are heat-sterilized. Charles F. Niven, Jr.

Bibliography. S. S. Block (ed.), *Disinfection, Sterilization, and Preservation*, 5th ed., 2001; P. Murray et al., *Medical Microbiology*, 2000.

Steroid

Any of a subset of biological molecules known as lipids that share a specific, fused four-ring structure comprising three rings of six carbon atoms and one ring of five carbon atoms. Different types of steroids have different chemical modifications to this common ring structure (see **illustration**). Steroids are found in animals, plants, fungi, and protozoa. In mammals, steroids are components of cell membranes, aid in digestion, and act as hormones to control a wide range of physiological responses. Because of their role in governing physiological processes, naturally occurring and synthetic steroids are used extensively in medicine to treat a wide range of conditions. *See* LIPID; STEROL.

Cholesterol. Cholesterol is the most abundant steroid in animals. It is found primarily in the plasma membrane of cells, where it is a major component of the lipid bilayer. It is also found, usually to a lesser extent, in the membranes of subcellular organelles such as mitochondria. Because cholesterol possesses a fused ring structure, it has greater rigidity than the other lipid components of these membranes and therefore tends to reduce the fluidity of membranes by restricting the movements of the more flexible lipid components. Thus, cholesterol plays an important role in determining the physical characteristics of biological membranes. In addition, cholesterol is one of the major components of lipid rafts, localized regions within the plasma membrane that are more

ordered and rigid than their surroundings. Lipid rafts associate with specific proteins and are believed to play an important role in signal transduction. While plants have very little cholesterol, a similar steroid, stigmasterol, plays an analogous role in the membranes of their cells. *See* CELL (BIOLOGY); CELL MEM-BRANES.

Because of its nonpolar nature, cholesterol has low solubility is water. Consequently, cholesterol must be transported through the bloodstream attached to carrier proteins known as lipoproteins. Low-density lipoproteins (LDL) transport cholesterol from the liver to various tissues in the body. High-density lipoproteins (HDL) transport cholesterol from various body tissues to the liver. HDL can be thought of as a "cholesterol scavenger" because it can take excess cholesterol directly from cell membranes.

High levels of cholesterol-containing LDL in the bloodstream are strongly correlated with arteriosclerosis (hardening of the arteries) and elevated risk of arterial blockage and heart attack. This is due to the deposition of excess cholesterol in the smooth muscle cells of the inner arterial wall. Over time, these deposits become calcified and can ultimately lead to artery blockage. High HDL levels have been correlated with reduced risks of heart disease. *See* CHOLESTEROL; LIPOPROTEIN.

Steroid hormones. In addition to being an important component of cell membranes, cholesterol is a precursor of a variety of critical steroid hormones. Among these are glucocorticoids, androgens, estrogens, progestins, and mineralocorticoids. Each family of steroid hormones elicits a different physiological response.

Glucocorticoids. Cortisol (also known as hydrocortisone) is an example of a glucocorticoid. This hormone is produced by the adrenal gland and participates in the regulation of carbohydrate



Structures of some important steroids.

metabolism. Cortisol causes an increase in the release of glucose from the liver and a decrease in the uptake of glucose by other tissues of the body. Consequently, cortisol effectively raises blood sugar (glucose) levels and acts in opposition to insulin. Cortisol also has anti-inflammatory and antiallergic effects, and is widely used medicinally to treat asthma, allergies, and various skin conditions.

Androgens. Testosterone is the primary androgen hormone. It is produced in the gonads and is a male sex hormone. Testosterone is required for the fetal development of male genitalia. During puberty, testosterone is required for the development of secondary sex characteristics in males, such as facial and pubic hair and a lowered voice. Testosterone also promotes protein synthesis that in turn leads to muscle production. Synthetic androgenic steroids that mimic testosterone are used medicinally to treat the severe weight loss that accompanies certain wasting diseases. To increase muscle mass and strength, many athletes have abused synthetic androgenic steroids. *See* ADRENAL GLAND.

Estrogens. Estradiol is the primary estrogen hormone. It is produced by the ovaries and placenta and is a female sex hormone. During puberty, estradiol is required for the development of female secondary sex characteristics, such as pubic hair and enlarged breasts. Estradiol causes the thickening of the uterine lining that occurs during the early phase of the menstral cycle. It also suppresses the secretion of follicle-stimulating hormone by the pituitary gland, thereby preventing ovulation. Because of its ability to suppress ovulation, estradiol is a major component of oral contraceptives. *See* ESTROGEN.

Progestins. Progesterone is the primary progestin. This hormone is produced by the adrenal gland and the corpus luteum (an ovarian follicle that has ruptured after releasing an egg). Progesterone is a female sex hormone. It causes changes in the uterine lining that enable implantation of a fertilized egg. If pregnancy results, the placenta starts to produce progesterone to prevent spontaneous abortion and prepare the mammary glands for milk production. Synthetic progestins are also major components of oral contraceptives. *See* PROGESTERONE.

Mineralocorticoids. Aldosterone is an example of a mineralocorticoid hormone. It is produced by the adrenal gland and regulates electrolyte balance. It acts on the kidneys to stimulate potassium ion excretion and to repress sodium ion excretion. *See* AL-DOSTERONE.

Mechanism of action. Steroid hormones elicit the physiological responses described above by altering the expression of various genes. Because steroids are nonpolar, they readily pass through the cell membrane. Once inside the cell, the steroid molecules bind to specific protein receptors in the cytoplasm. The hormone-protein receptor complexes then enter the nucleus and bind to specific regions of the DNA known as response elements. In many cases, binding of the hormone-protein receptor complex to response elements in the DNA induces the expression of genes adjacent to the response element. In other cases, this binding represses the expression of genes adjacent to the response element.

Bile acids. In addition to being a precursor of steroid hormones, cholesterol is a precursor of bile acids (also known as bile salts), such as cholic acid and glycocholic acid. Cholesterol is converted to bile acids in the liver, and these compounds are secreted into the small intestine via the gall bladder. In the small intestine, bile acids act as emulsifying agents to aid in the digestion and absorption of fats. While most of the bile acid in the small intestine is reabsorbed and recycled, some is not. This nonabsorption represents the major means of excretion of excess cholesterol by the body. *See* LIVER. M. Todd Washington

Bibliography. T. M. Devlin (ed.), *Textbook of Biochemistry with Clinical Correlation*, 5th ed., Wiley, New York, 2002; D. Voet and J. G. Voet, *Biochemistry*, 3d ed., Wiley, New York, 2004.

Sterol

Any of a group of naturally occurring or synthetic organic compounds with a steroid ring structure, having a hydroxyl (-OH) group, usually attached to carbon-3. This hydroxyl group is often esterified with a fatty acid (for example, cholesterol ester). The hydrocarbon chain of the fatty-acid substituent varies in length, usually from 16 to 20 carbon atoms, and can be saturated (containing no double bonds) or unsaturated (containing one or more double bonds). Sterols commonly contain one or more double bonds in the ring structure and also a variety of substituents attached to the rings. Sterols and their fatty-acid esters are essentially water insoluble. For transport in an aqueous milieu (for example, the bloodstream of mammals), sterols and other lipids are bound to specific proteins, forming lipoprotein particles. These particles are classified based on their composition and density. One lipoprotein class is abnormally high in the blood of humans prone to heart attacks

Sterols are widely distributed in nature. The biosynthetic pathway leading to the steroid ring structure is utilized by higher and lower forms of life. Modifications of the steroid ring structure are made by specific enzyme systems, producing the sterol characteristic for each species such as ergosterol in yeast. The common biosynthetic pathway begins with the two-carbon-atom compound, acetate, which is converted by several enzymic steps to a six-carbon-atom compound, mevalonate. Next, mevalonate is converted to a five-carbon-atom intermediate, isopentenyl pyrophosphate. There follows a series of condensations of six isopentenyl pyrophosphate molecules to yield a 30-carbon-atom intermediate, squalene. Squalene is next cyclized to a 30-carbon-atom sterol, lanosterol. The structure of lanosterol is then modified to yield the principal

sterol of a species. The major regulatory step in the sterol biosynthetic pathway occurs early in the process. Drugs that lower blood cholesterol levels in humans are designed to inhibit this regulatory enzyme. In addition to their conversion to sterols, several intermediates in the pathway are precursors of other important biological compounds, including chlorophyll in plants, vitamins A, D, E, and K, and regulators of membrane functions and metabolic pathways.

A universal role of sterols is to function as part of membrane structures. In addition, some insects require sterols in their diets. Cholesterol also serves as a precursor of steroid hormones (estrogens, androgens, glucocorticoids, and mineralocorticoids) and bile acids. *See* CHOLESTEROL; STEROID. Mary E. Dempsey

Bibliography. K. Bloch, The biological synthesis of cholesterol, *Science*, 150:19–28, 1965; G. W. Patterson and W. D. Nes (eds.), *Physiology and Biochemistry of Sterols*, 1992.

Stichosomida

An order of nematodes (roundworms) formerly constituting at least two important groups of animal parasites: the mermithids, which are parasites of invertebrates, and the trichinellids, which are parasites of vertebrates. These nematodes are characterized by a pharynx that is narrow and thin-walled anteriorly and is surrounded posteriorly by unicellular, glandular stichocytes, each with a duct into the pharyngeal lumen (Figs. 1 and 2). The pharynx extends one-fourth to nine-tenths of the body length in various taxa and is almost devoid of musculature. The region of the pharynx surrounded by stichocytes is known as the stichosome. Recent phylogenetic analysis based on a synthesis of molecular and morphological data suggests that the stichosome may be an example of parallel evolution and that the trichinellids and mermithids are more appropriately separated as two orders: Trichinellida, with at least six families, and Mermithida, with two families. However, debate on the higher classification of nematodes continues, and opinion will no doubt refocus as new data emerge and are applied. See MEDICAL PARASITOLOGY; NEMATA (NEMATODA).

Two life history examples provide some insight into the biology of these organisms. In *Trichinella spiralis*, encysted larvae are ingested in infected muscle tissue (raw or undercooked pork is the classic example in the case of trichinosis in humans). The cyst surrounding the larva is digested in the new host, and the larvae molt to adults, mate, and embed in the intestinal epithelium; then females produce eggs which hatch (1000 larvae per female in 5 days). The hatched larvae are distributed via the circulatory system and migrate into surrounding cells; the larvae die unless they are in striated muscle fibers. Secretory products of the stichocytes induce DNA endoredu-



Fig. 1. Anterior end of Trichinella spiralis showing stichostome. (Adapted from B. G. Chitwood, 1930)

plication (replication of nuclear DNA without mitotic cell division) and transformation of the muscle fiber into a multinucleate nurse cell, which becomes encapsulated by collagen and supplied with capillaries. The life cycle continues when the muscle is eaten by another animal.

In *Romanomermis culicivorax*, the preparasitic larva is equipped with a stylet which allows penetration through the cuticle of the mosquito larva host. The stylet is lost in subsequent larval stages,



Fig. 2. Portion of the stichosome of the whipworm, *Trichuris trichiura.* (*Photomicrograph by Sung-Jong Hong, The Korean Society for Parasitology*)

and the stichosome separates from the intestine, which becomes a trophosome (a modified nutritive storage organ) with no anal opening to the exterior. Nutrient uptake from the insect hemolymph is through a very thin cuticle, perhaps enhanced by small-diameter pores. Interestingly, the epidermal cells adjacent to the cuticle have their absorptive surface area increased by outwardly directed microvilli. The continued role of the stichosome is unclear, although it appears to be involved in protein synthesis. The larva undergoes one molt in the insect host and then ruptures the body wall as it emerges. After a final molt to the adult stage, the female deposits eggs.

The nature and function of stichocyte secretions, particularly of vertebrate parasites, constitute an area of continued interest in research on host-parasite relationships. Howard Ferris

Bibliography. B. G. Chitwood, The structure of the esophagus in the Trichuroidea, J. Parasitol., 17:35-42, 1930; P. De Ley and M. Blaxter, Systematic position and phylogeny, pp. 1-30 in D. L. Lee (ed.), The Biology of Nematodes, Taylor and Francis, London and New York, 2002; D. D. Despommier, Trichinella and Toxocara, pp. 597-607 in F. E. G. Cox, J. P. Kreier and D. Wakelin (eds.), vol. 5: Parasitology, in Topley and Wilson's Microbiology and Microbial Infections, Arnold, London, 1998; D. L. Lee, Life cycles, pp. 61-72 in D. L. Lee (ed.), The Biology of Nematodes, Taylor and Francis, London and New York, 2002; E. A. Munn and P. D. Munn, Feeding and digestion, pp. 211-232 in D. L. Lee (ed.), The Biology of Nematodes, Taylor and Francis, London and New York, 2002; G. O. Poinar, The Natural History of Nematodes, Prentice Hall, Englewood Cliffs, NJ, 1983; G. O. Poinar and R. Hess, Romanomermis culicivorax: Morphological evidence of transcuticular uptake, Exp. Parasitol., 42:27-33, 1977.

Stickleback

Any member of the fish family Gasterosteidae, order Gasterosteiformes. The body may be elongate, naked, or with bony scutes along the sides. Three to 16 well-developed isolated dorsal spines precede a normal dorsal fin having 6 to 14 rays. Additional identifying features are as follows: 12 or 13 caudal fin rays; a single spine and 1 or 2 soft rays in the pelvic fin; 3 branchiostegal rays; circumorbital ring incomplete posteriorly; epineurals present; and 28 to 42 vertebrae. A maximum length of about 18 cm (7 in.) is reached in *Spinachia spinachia. See* GAS-TEROSTEIFORMES; OSTEICHTHYES; TELEOSTEI.

Sticklebacks are small freshwater and marine fishes of the cold and temperate waters of the Northern Hemisphere. All species are of some economic importance since their diet includes mosquito larvae, and the freshwater species are especially popular aquarium fishes. Also, sticklebacks are famous as subjects of numerous studies regarding evolution, genetics, behavior, and physiology. *See* INSTINCTIVE BEHAVIOR; SEXUAL DIMORPHISM.

Notable species. The fifteen-spined stickleback (*Spinachia spinachia*) and the four-spined stickleback (*Apeltes quadracus*) are marine shorefishes. *Spinachia spinachia* is found from Norway to the Bay of Biscay (France-Spain). The adult is about 18 cm (7 in.) in length and has a life span of 1 year. In low tides waters, the male builds a nest of algae stuck together with his mucus secretions. The female enters the nest and deposits 150 to 200 eggs, which are then fertilized by the male. The female dies shortly after egg deposition, and the male guards the nest and eggs. *Apeltes quadracus* occurs in freshwaters along the Atlantic coast from Nova Scotia to Virginia.

Pungitius pungitius (Fig. 1) measures about 7.5 cm (3 in.) in length and occurs in fresh and brackish waters of North America and Europe. This species is known as the nine-spine stickleback, but the number of spines varies. Nest building is relegated to the male, which also protects the eggs and fry until they hatch in about 1 week and protects the fry.



Fig. 1. Nine-spine stickleback (Pungitius pungitius). (Photo by Konrad P. Schmidt)



Fig. 2. Three-spine stickleback (Gasterosteus aculeatus). (Bob Jenkins and Noel Burkhead; EFISH <http://www.cnr.vt.edu/efish/>)

The best-known species is the three-spine stickleback (Gasterosteus aculeatus) [Fig. 2], the behavior of which has been extensively studied. The male makes a depression in the sandy bottom; here he builds a nest of vegetation cemented together by mucous strands produced by the kidneys, which are enlarged at breeding time. During the breeding season, the abdomen of the males becomes red and, when exhibited in the territory of the nesting area, evokes an innate behavioral response and attracts the females. The male also performs an elaborate courtship dance to entice the female to enter the nest. When the female does so, the male nuzzles her near the base of the tail and thus stimulates her to oviposit. The male then fertilizes the eggs and remains to guard and aerate them. See REPRODUCTIVE BEHAVIOR. Herbert Boschung; Charles B. Curtin

Bibliography. M. A. Bell and S. A. Foster (eds.), *The Evolutionary Biology of the Threespine Stickleback*, Oxford University Press, 1994; R. Froese and D. Pauly (eds.), FishBase, World Wide Web electronic publication, version (05/2005); J. S. Nelson, *Fishes of the World*, 4th ed., Wiley, New York, 2006; R. J. Wootton, *The Biology of the Sticklebacks*, Academic Press, 1976; R. J. Wootton, *A Functional Biology of Sticklebacks*, University of California Press, 1984.

Stilbite

A mineral belonging to the zeolite family of silicates. It crystallizes in the monoclinic system in crystals that are tabular parallel to the side pinacoid. Most characteristic are sheaflike aggregates of thin tabular crystals. There is perfect cleavage parallel to the side pinacoid, and here the mineral has a pearly luster; elsewhere the luster is vitreous. The color is usually white but may be brown, red, or yellow. Hardness is $3^{1}/_{2}$ -4 on Mohs scale; specific gravity is 2.1–2.2. Stilbite is a calcium-sodium aluminum silicate, Ca(Al₂Si₇O₁₈) · 7H₂O, but usually contains some sodium replacing calcium.

Stilbite is a secondary mineral usually found in cavities in basalts and related rocks. Much less commonly it is found in granites, gneisses, and metalbearing veins. It is associated with other zeolites, datolite, prehnite, and calcite. Some notable localities are in Iceland, India, Scotland, Nova Scotia, and in the United States at Bergen Hill, New Jersey, and the Lake Superior copper district in Michigan. *See* SILICATE MINERALS; ZEOLITE. Cornelius S. Hurlbut, Jr.

Stirling engine

An engine in which work is performed by the expansion of a gas at high temperature to which heat is supplied through a wall. Like the internal combustion engine, a Stirling engine provides work by



Fig. 1. Principle of Stirling engine, displacer type.

means of a cycle in which a piston compresses gas at a low temperature and allows it to expand at a high temperature. In the former case the heat is provided by the internal combustion of fuel in the cylinder, but in the Stirling engine the heat (obtained from externally burning fuel) is supplied to the gas through the wall of the cylinder (**Fig. 1**). *See* INTERNAL COM-BUSTION ENGINE.

The rapid changes desired in the gas temperature are achieved by means of a second piston in the cylinder, called a displacer, which in moving up and down transfers the gas back and forth between two spaces, one at a fixed high temperature and the other at a fixed low temperature. When the displacer in **Fig. 2** is raised, the gas will flow from the hot space via the heater and cooler tubes into the cold space. When it is moved downward, the gas will return to the hot



Fig. 2. Stylized Stirling process.



Fig. 3. Piston and displacer movement for a Stirling engine. (a) The continuous movement of the piston and displacer shown as a function of crank angle α . There is no clear-cut division between the various phases of the cycle. The variations in volume of the hot space V_E and the cold space V_C are plotted separately in the lower part of the diagram. (b) The *P*-*V* diagram of the cycle.

space along the same path. During the first transfer stroke the gas has to yield up a large amount of heat to the cooler; an equal quantity of heat has to be taken up from the heater during the second stroke.

The regenerator shown in Figs. 1 and 2 has been inserted between the heater tube and cooler tube in order to prevent unnecessary waste of this heat. It is a space filled with porous material to which the hot gas gives up its heat before entering the cooler; the cooled gas recovers the stored heat on its way back to the heater. The displacer system serves to heat and cool the gas periodically; associated with it is a piston which compresses the gas while it is in the cold space of the cylinder and allows it to expand while in the hot space. Since expansion takes place at a higher temperature than compression, it produces a surplus of work over that required for the compression.

Stirling cycle. Any practical version of the engine will embody some kind of crank and connecting rod mechanism, in consequence of which there will be no sharp transitions between the successive phases indicated in Fig. 2; but this will not alter the principle of the cycle (nor detract from its efficiency). If, for the sake of simplicity, the piston and displacer are assumed to move discontinuously, the cycle can be divided into the following four phases. (1) The piston is in its lowest position, the displacer in its highest. All the gas is in the cold space. (2) The displacer is still in its highest position; the piston has compressed the gas at low temperature. (3) The piston is still in its highest position; the displacer has moved down and transferred the compressed gas from the cold to the hot space. (4) The hot gas has expanded, pushing the pistons, followed by the displacer, to their lowest positions. The displacer is about to rise and return the gas to the cold space, the piston remaining where it is, to give phase I again.

The actual piston and displacer movements might be as indicated in Fig. 3, which shows that the only essential condition for obtaining a surplus of work is that the maximum volume of the hot space occur before that of the cold space. This condition shows that more configurations with pistons and cylinders are possible than just the type with displacer in order to get a Stirling cycle. One of the most compact systems is shown in Fig. 4, which is the system known as the double-acting engine. In the double-acting engine there is a hot space (expansion space) at the top and a cold one (compression space) at the bottom of each of the four cylinders shown. The hot space of a cylinder is connected to the cold one through a heater, a regenerator, and a cooler. The pistons P_n of the cylinders move with a suitable phase shift between them. In the case of four cylinders, as shown in Fig. 4, this shift is 90° .

The theory of an actual engine is very complicated. In order to provide an understanding of the quantities that play a role, the formulas of the power and efficiency of an engine of highly idealized form will be given. On the assumption that the volumes of the hot (expansion) space V_E and that of the cold (compression) space V_C (Fig. 3) vary with the crank angle in a purely sinusoidal way, that both expansion and compression are isothermal at respectively T_E and T_C , and that all kinds of losses caused by the flow resistances in the tubes and regenerator, heat losses in the regenerator, and so on are ignored, the power P can be expressed as in Eq. (1). Here $\tau = T_C/T_E$,

$$P = (1 - \tau) \frac{\pi n}{60} V_{E_{\text{max}}} \cdot p_m \frac{\delta}{1 + \sqrt{1 - \delta^2}} \sin \Theta \quad (1)$$

n is the speed in rpm, p_m is the mean pressure of the pressure variation, and δ and Θ are functions of the temperature ratio, swept volume ratio, dead spaces, and phase angle between the two swept volumes. Under these conditions the thermal efficiency η is, of course, that of the Carnot cycle, and is expressed in Eq. (2).

$$\eta = \frac{T_E - T_C}{T_F} = 1 - \tau \tag{2}$$

See CARNOT CYCLE; THERMODYNAMIC CYCLE.

Engine design. Actual engines have been built in the Philips Laboratories at Eindhoven, Netherlands, as prototypes in the range of 10-500 hp (7.5-375 kW) per cylinder. After years of research on the Stirling engine, actual thermal efficiencies of 30-45% (depending on the specific output and temperature ratio) and a specific power of 1.88 hp/in.³ (85 kW/liter) swept volume of the piston were obtained with the displacer type of engine equipped with rhombic drive, as shown in Fig. 5. In the figure the piston and the displacer drive concentric rods, which are coupled to the rhombic drive turning twin timing gears. The cooler, regenerator, and heater are arranged as annular systems around the cylindrical working space. The heater tubes surround the combustion chamber. In the preheater the gas at 1470° F (800°C) from the heater is cooled to 300-390°F (150-200°C)



Fig. 4. Principle of the double-acting engine.

while heating the combustion air to about $1200^{\circ}F$ (650°C).

The rhombic drive mechanism allows complete balancing even of a single-cylinder engine and of a separate buffer space, thus avoiding heavy forces acting on the drive. The results of measurements on the first engine of this type, with hydrogen as the working fluid, are shown in **Fig. 6**. An approximate heat balance is output, 40%; exhaust and radiation, 10%; and heat rejection by cooling water, 50%. Control of engine output is by regulation of the pressure of the working fluid in the engine, while the temperature of the heater is kept constant by a thermostat;



Fig. 5. Philips Stirling engine in cross section. (Philips Gloeilampenfabrieken)



Fig. 6. Test measurements for the Stirling engine. (a) Measured shaft power P_e and efficiency of a 40-hp (30-kW) singlecylinder test engine with rhombic drive, plotted as a function of the engine speed *n*, at different values of the maximum pressure of the working fluid in the engine (p_{max}). 1 hp = 0.75 kW; 1 kgf/cm² = 98 kPa = 14.2 lbf/in.² (b) Power and efficiency of the 40-hp (30-kW) test engine, given as a function of the heater temperature and as a function of the inlet temperature of the cooling water. The curves apply to n = 1500 rpm and $p_{max} = 140$ kgf/cm² = 13.7 MPa = 1990 lbf/in.²; °F = (°C × 1.8) + 32.

hence the efficiency shows little dependence on the load.

The closed system of the Stirling engine endows this engine with many advantages and also some shortcomings. The continuous external heating of the closed system makes it possible to burn various kinds of liquid fuels and gases, without any modification whatsoever. This multifuel facility can be demonstrated with a 10-hp or 7.5-kW (at 3000 rpm) generator set (Fig. 7). The engine can operate on alcohol, various lead-containing gasolines, diesel fuel, lubricating oil, olive oil, salad oil, crude oil, propane, butane, and natural gas. Furthermore, it allows combustion to take place in such a way that air pollution is some orders of magnitude less than that due to internal combustion engines. Through the intermediary of a suitable heat transport system (for example, heat pipes) any heat source at a sufficiently high temperature can be used for this engine-radioisotopes, a nuclear reactor, heat storage, solar heat, or even the burning of coal or wood.

The almost sinusoidal cylinder pressure variation

and continuous heating make the Stirling engine very quiet in operation. An engine having four or more cylinders gives a virtually constant torque per revolution, as well as a constant dynamometer torque over a wide speed range, which is particularly valuable for traction purposes. The present configuration makes complete balancing possible, thus eliminating vibrations. There is no oil consumption and virtually no contamination because a new type of seal for the reciprocating rods shuts off the cycle hermetically from the drive mechanism. **Figure 8** shows an engine of this configuration.

Where direct or indirect air cooling is required, the closed cycle has the drawback that more heat has to be removed from the cooler than in comparable engines with open systems, where a greater quantity of heat inevitably escapes through the exhaust.

If it is envisaged as someday taking the place of existing engines, the Stirling might be ideal as a propulsion engine in yachts and passenger ships, and in road vehicles, such as city buses, where a large radiator is acceptable. The system of continuous



Fig. 7. Phillips Stirling engine with generator to demonstrate its multifuel capacity. (Philips Gloeilampenfabrieken)



Fig. 8. A 3-kW Stirling engine generator set. (General Motors Corp.)

external heating is also able to open fields of application inaccessible to internal combustion engines. Roelof J. Meijer

Bibliography. R. Darlington, *Stirling and Hot Air Engines*, 2005; T. Finkelstein and A. J. Organ, *Air Engines: The History, Science, and Reality of the Perfect Engine*, 2001; W. R. Martini, *Stirling Engine Design Manual*, 2004.

Stishovite

Naturally occurring stishovite, SiO₂, is a mineral formed under very high pressure with the silicon atom in sixfold, or octahedral, coordination instead of the usual fourfold, or tetrahedral, coordination. It was discovered in the same type of Coconino sandstone from the Meteor Crater of Arizona. Stishovite has since been found in rocks from the Ries Crater of Germany. The presence of stishovite indicates formation pressures in excess of 10⁶ lb/in.² (7.5 gigapascals). The possibility of the existence of stishovite at great depths strongly influences the interpretations of geophysicists and solid-state physicists regarding the phase transitions of mineral matter, as well as the interpretation of seismic data in the study of such regions of the interior of the Earth. See SILICA MINER-ALS.



Comparison of the mean index of refraction and specific gravity of stishovite and other polymorphic modifications of silica.

Natural occurrence. Stishovite occurs in submicrometer size in very small amounts (less than 1% of the rock) in samples of Coconino sandstones from the Meteor Crater of Arizona, which contains up to 10% of coesite, the other high-pressure polymorph of silica. Besides stishovite and coesite, these rocks consist mainly of quartz, with small amounts of silica glass (lechatelierite) and traces of zircon and, in some cases, rutile. Because of its extremely fine grain size and because of the sparsity of this mineral in the rock, positive identification of the mineral is possible only by the x-ray diffraction method after chemical concentration. *See* COESITE; METEORITE.

Properties. Stishovite is much more resistant to attack by concentrated hydrofluoric acid than coesite. It can therefore be extracted or concentrated by prolonged treatment of the stishovite-bearing sandstone in hydrofluoric acid. Concentrates of stishovite, all of submicrometer size, are pale gray and have a roughly prismatic habit. Because of its fine-grained habit, only the mean index of refraction of the natural material can be measured. It is 1.80.

Crystallographically, stishovite is SiO₂ in the rutile structure, tetragonal with a = 0.41790 nanometer and c = 0.26649 nm (with a standard error of ± 0.00004 nm). The axial ratio is c/a is 0.6377, and the volume of the tetragonal unit cell is 0.046541 nm³. The specific gravity of stishovite, calculated from the x-ray data, is 4.28, as reported by Chao and others, compared with the value of 4.35 for the synthetic material reported by Stishov and Popova. It is 46% denser than coesite and much denser than other modifications of silica (see **illus.**). Edward C. T. Chao

Bibliography. G. Beggerow, *Higb-Pressure Properties of Matter*, 1980; E. C. T. Chao et al., Stishovite, SiO₂, a very high pressure new mineral from Meteor Crater, Arizona, *J. Geophys. Res.*, 67(1):419-421, 1962; P. J. Heaney, C. T. Prewitt, and G. V. Gibbs (eds.), *Silica: Physical Behavior, Geochemistry, and Materials Applications (Reviews in Mineralogy, Vol. 29)*, 1994.

Stochastic control theory

A branch of control theory which aims at predicting and minimizing the magnitudes and limits of the random deviations of a control system output through optimizing the design of the controller. Such deviations occur when random noise and disturbance processes are present in a control system, so that the system does not follow its prescribed course but deviates from the latter by a randomly varying amount.

In contrast to deterministic signals, random signals cannot be described as given functions of time such as a step, a ramp, or a sine wave. The exact function is unknown to the system designer; only some of its statistical properties are known.

A random signal may be generated by one of nature's processes, for instance, radar noise and windor wave-induced forces and moments on a radar antenna or a ship. Alternatively, a random signal may be generated by human intelligence, for instance, the bearing of a zigzagging aircraft, or the contour to be followed by a duplicating machine.

One outstanding experimental fact about nature's random processes is that these signals are approximately gaussian. The word gaussian refers to a mathematical concept which describes one or more signals i_1, i_2, \ldots, i_n having the following properties:

1. The amplitude of each signal is normally distributed.

2. The joint distribution function of any number of signals at the same or different times taken from the set is a multivariate normal distribution. This experimental fact is not surprising in view of the fact that a random process of nature is usually the sum total of the effects of a large number of independent contributing factors. For instance, thermal noise is due to the thermal motions of billions of electrons and atoms. An ocean-wave height at any particular time and place is the sum of wind-generated waves at previous times over a large area. Gaussian signals have the property that when passed through linear systems they remain gaussian. *See* DISTRIBU-TION (PROBABILITY); ELECTRICAL NOISE; STOCHASTIC PROCESS.

In the following, the mathematical characterization of the random processes, their transmission properties, measurements and simulation, effects on the system output, and control system optimization will be introduced.

Mathematical characterization. The random processes encountered in control work can be classified as stationary and nonstationary processes. The underlying mechanism that generates a random process can usually be described in physical or mathematical terms. For instance, the underlying mechanism that generates shot-effect noise is thermionic emission; the underlying mechanism that generates ocean waves is essentially wind force in conjunction with the Earth's gravity; and the outcome of dice throwing is determined by a probability of 1/6 for each face of each die independent of all others.

If the generating mechanism does not change with time, any measured average property of the random process such as mean or variance is independent of the time of measurement aside from statistical fluctuations, and the random process is called stationary. For instance, ocean-wave height in a given sea state, thermal noise, and a telegraphic signal of a certain language are examples of stationary random processes. If the generating mechanism does change, the random process is called nonstationary, in which case the generating mechanism may change in a predetermined fashion or at random. An example of the former is found in missile launching, where the unwanted fluctuations in vertical and lateral thrusts are generated by one or a group of rocket motors of which the operating condition is a predetermined function of time. Alternatively, the generating mechanism may change at random, as in the course of flight of an aircraft, where the wind disturbance depends on the local atmospheric condition which may be calm or turbulent.

Statistical ensemble. The random processes or events generated by the same underlying mechanism form a statistical ensemble. The ensemble average is the averaged value over all possible events weighted according to the probabilities of such events. For a stationary process there is little reason to distinguish between the ensemble average at any particular time and the time average of a single event. The life history of a single event can be chopped up into an infinite number of events. Since these events are generated by the same underlying mechanism which does not change with time, they can be thought of as arbitrarily selected from the entire ensemble, and therefore they represent the latter. The assumption of equivalence between time and ensemble averages is called the ergodic hypothesis. See STATISTICAL MECHANICS; TOPOLOGICAL DYNAMICS.

The ensemble average of a random variable or function is referred to as its mean value. For stationary systems, the mean value can be obtained by taking either time or ensemble average. Let \overline{x} denote the mean value of a random variable x, and \tilde{x} its random variation from the mean: $x = \overline{x} + \tilde{x}$. As \overline{x} can be absorbed as part of the deterministic input, the random signals may be defined as variations about the mean values, and thus they have zero mean themselves.

Correlation functions. The statistical properties of a set of random signals (which may include the random inputs and system variables in response to these inputs) can be described by their correlation functions. The correlation function $\phi_{ab}(t_1,t_2)$ is defined as the ensemble average of signal a(t) at time t_1 and b(t) at time t_2 , Eq. (1). When the two signals a(t) and b(t)

$$\phi_{ab}(t_1, t_2) \equiv \overline{a(t_1)b(t_2)} \tag{1}$$

are different, $\phi_{ab}(t_1,t_2)$ is called the cross-correlation function. When a(t) and b(t) are one and the same, $\phi_{ab}(t_1,t_2)$ is called the autocorrelation function.

For stationary systems, the only significant independent variable is $t_2 - t_1$, denoted as τ , Eq. (2). The

$$\phi_{ab}(\tau) = \overline{a(t)b(t+\tau)} \tag{2}$$

time interval τ is called the correlation time. The useful properties of the correlation function given in Eqs. (3)-(7) follow from Eq. (2). Equation (6)

$$\phi_{aa}(0) = \overline{a(t)^2} \tag{3}$$

$$\phi_{ab}(-\tau) = \phi_{ba}(\tau) \tag{4}$$

$$\phi_{aa}(-\tau) = \phi_{aa}(\tau) \tag{5}$$

$$|\phi_{ab}(\tau)| \le \sqrt{\phi_{aa}(0)\phi_{bb}(0)} \tag{6}$$

$$|\phi_{aa}(\tau)| \le \phi_{aa}(0) \tag{7}$$

is obtained by taking the ensemble average of Eq. (8), yielding Eq. (9), and combining this with

Eq. (10). Equation (7) is a special case of Eq. (6) with b(t) = a(t).

$$0 \leq \left\{ \frac{|a(t)|}{\sqrt{\phi_{aa}(0)}} - \frac{|b(t+\tau)|}{\sqrt{\phi_{bb}(0)}} \right\}^{2} \\ = \frac{a(t)^{2}}{\phi_{aa}(0)} - \frac{2|a(t)b(t+\tau)|}{\sqrt{\phi_{aa}(0)\phi_{bb}(0)}} + \frac{b(t+\tau)^{2}}{\phi_{bb}(0)}$$
(8)

$$0 \le 2 - \frac{\overline{2 |a(t) \ b(t+\tau)|}}{\sqrt{\phi_{aa}(0) \ \phi_{bb}(0)}} \tag{9}$$

$$|\phi_{ab}(\tau)| = \overline{|a(t)b(t+\tau)|} \tag{10}$$

Equations (3)-(7) show that the autocorrelation function is even in τ , and its value at $\tau = 0$ is the mean-square value of the random signal itself. The magnitude of the cross-correlation function cannot exceed the product of the root-mean-square values of the two signals.

Spectral-density functions. If the chopped signals $a_T(t)$ and $b_T(t)$ are defined as the same random signals a(t) and b(t), respectively, in the interval $0 \le t \le T$ but zero outside this interval, and $A_T(s)$ and $B_T(s)$ represent their Laplace transforms, then a function $\Phi_{ab}(s)$ can be defined by Eq. (11). Equation (11) implies that Eqs. (12) and (13) are valid. Equations (12)

$$\Phi_{ab}(s) \equiv \lim_{T \to \infty} \frac{1}{T} \overline{A_T(-s)B_T(s)}$$
(11)

$$\Phi_{ab}(j\omega) = \Phi_{ba}(-j\omega) \tag{12}$$

$$\Phi_{aa}(j\omega) = \Phi_{aa}(-j\omega) \rangle 0 \tag{13}$$

and (13) show that $\Phi_{ab}(j\omega)$ is the complex conjugate of $\Phi_{ba}(j\omega)$ and that $\Phi_{aa}(j\omega)$ is a real positive even function of ω . The function $\Phi_{ab}(s)$ is called the spectral-density function when *a* and *b* are the same, or the cross-spectral-density function when *a*(*t*) and *b*(*t*) represent different signals.

It is readily shown that $\phi_{ab}(\tau)$ and $\Phi_{ab}(s)$ are related by Laplace transforms, Eqs. (14) and (15). The

$$\phi_{ab}(\tau) = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} \Phi_{ab}(s) e^{s\tau} \, ds \qquad (14)$$

$$\Phi_{ab}(s) = \int_{-\infty}^{\infty} \phi_{ab}(\tau) e^{-s\tau} d\tau \qquad (15)$$

transform relationship between correlation functions and spectral densities as represented by Eqs. (14) and (15) is called Wiener's theorem.

Transmission properties. When stochastic processes are applied to a physical system, the random output variations are related to the inputs and to each other. These relationships are expressed in terms of the transmission properties.

The **illustration** shows a multiple-input and multiple-output system where the vector $\mathbf{x}(t)$ denotes inputs $x_1(t), x_2(t), \ldots, x_n(t)$ and $\mathbf{y}(t)$ denotes outputs $y_1(t), y_2(t), \ldots, y_m(t)$, through filters with transfer functions $H_{ik}(s)$, Eq. (16). The filters are

$$Y_i(s) = \sum_k H_{ik}(s) X_k(s) \tag{16}$$



Transmission of stochastic processes through the linear filter H.

assumed to be stable (with all their poles in the left half of the *s*-plane). The spectral densities of components of **y** are expressed in terms of that of **x** by Eq. (17). Equation (17) is readily derived from

$$\Phi_{y_l y_j}(j\omega) = \sum_k \sum_l H_{ik}(-j\omega) \ \Phi_{x_k x_l}(j\omega) H_{jl}(j\omega)$$
(17)

Eq. (11) and can be written in matrix form, Eq. (18),

$$\Phi_{y}(j\omega) = \mathbf{H}(-j\omega)\Phi_{x}(j\omega)\mathbf{H}^{T}(j\omega)$$
(18)

where Φ_{y_i} , Φ_{x_i} , and **H** are $m \times m$, $n \times n$, and $m \times n$ matrices with $\Phi_{y_i}y_j$, $\Phi_{x_i}x_j$, H_{ij} as their i-jth components respectively. Note that $\mathbf{H}(-j\omega)$ and $\mathbf{H}^T(j\omega)$ are hermitian adjoints of each other. Equations (17) and (18) are perfectly general and can be used to obtain cross-spectral density between input and output of a filter as well. For instance, by regarding **x** as the m + 1- to m + n-th components of an enlarged output vector, Eq. (18) readily leads to Eq. (19), where

$$\Phi_{xy}(j\omega) = \Phi_x(j\omega)\mathbf{H}^T(j\omega)$$
(19)

 Pbi_{xy} is an $n \times m$ matrix with $Pbi_{xi}y_j$ as its *ij*th component. See LINEAR SYSTEM ANALYSIS.

If the process is nonstationary, frequency domain analysis cannot be used. Equation (18) is replaced by Eq. (20), where $h_{kl}(t,\tau)$ denotes the response of filter

$$W_{y}(t,t') = \int_{-\infty}^{t} \int_{-\infty}^{t'} \mathbf{h}(t,\tau) \phi_{x}(\tau,\tau') \\ \times \mathbf{h}^{T}(t',\tau') \, d\tau \, d\tau' \quad (20)$$

 η_{kl} at time *t* to unit impulse input at time τ .

White stochastic process. A stochastic process x with n components x_1, x_2, \ldots, x_n is said to be white if there is no correlation between any of its components at different times, Eq. (21). This includes auto-

$$\phi_{ik}(t, t') = 0 \text{ if } t \neq t'$$
 (21)

correlations with i = k. Equation (21) has an interesting implication: at t = t', the autocorrelation function for a white process must have infinite instantaneous amplitude if it is to produce any physically significant effect. This can be proven as follows: A physical system can be represented by (or approximated as) a stable filter with finite impulse response $b(t,\tau)$. The mean-square value of its response a(t) to a white noise input x(t) can be obtained from Eq. (20), yielding Eq. (22). If it is assumed to the contrary that x(t) is

$$[a(t)]^{2} = \phi_{aa}(t, t)$$

= $\int_{-\infty}^{t} \int_{-\infty}^{t} b(t, \tau) b(t, \tau') \phi_{xx}(\tau, \tau') d\tau d\tau'$ (22)

finite, then $\phi_{xx}(\tau, \tau') \leq M^2$. This condition, together with Eq. (21), gives the integral of Eq. (22) to be zero.

Equation (21) and the requirement of finite physical effect lead to the correlation function of Eq. (23)

$$\phi_{ik}(t, t') = W_{ik}(t)\delta(t - t')$$
 (23)

for the white noise process, where the matrix $\mathbf{W}(t)$ with $W_{ik}(t)$ as its elements is positive semidefinite for all *t* (that is, $\mathbf{v}^T \mathbf{W} \mathbf{v} \ge 0$ for any vector **v** with real components); and $\delta(t)$ is the Dirac function defined by Eqs. (24) and (25). For a stationary white process,

$$\delta(t) = 0 \text{ if } t \neq 0$$
(24)
$$\int_{a}^{a} \delta(t) \, dt = 1 \text{ for all } a \rangle 0$$
(25)

Wiener's theorem and Eq. (23) give the spectral density as Eq. (26).Equation (26) illustrates the origin of

$$\Phi_{ik}(j\omega) = W_{ik} \tag{26}$$

the description "white" in the sense that the spectral density is independent of frequency.

A white process is purely a mathematical concept as no physical process can have infinite amplitude or infinite bandwidth. It is nevertheless a very important concept in simulation of stochastic processes and in system optimization. A random process can be regarded as white in an approximate sense if Eq. (21) is satisfied for $|t-t'| > \Delta$, where Δ is a time interval much smaller than the response time of the control system under consideration. An equivalent condition is for Eq. (26) to hold for all $\omega < \omega_k$, where ω_k is at least an order of magnitude larger than the system bandwidth.

Simulation of stochastic processes. A white process can be simulated either digitally or by an analog method. Digitally it is simulated by a sequence of random numbers at a rate at least a decade higher than the system bandwidth. The numbers are then used directly in a simulation study or converted into an electrical signal in actual tests. In analog white noise generators, shot effect noise or thermal noise is amplified by a wide-band amplifier. Correlated white processes with matrix **W** are generated by using independent white noise sources and a resistive network. The required number of independent sources is equal to the rank of **W**.

Nonwhite stochastic processes can be simulated as outputs of filters with white noise sources as inputs. Assuming $Phi_x(j\omega)$ to be the identity matrix, Eqs. (18) and (20) give respectively Eqs. (27) and (28). The problem is then to find $H(j\omega)$ so that $\phi_{\nu}(j\omega)$

$$\phi_{\gamma}(j\omega) = \mathbf{H}(-j\omega)\mathbf{H}^{T}(j\omega)$$
(27)

$$\phi_{y}(t,t') = \int_{-\infty}^{t} \mathbf{h}(t,\tau) \mathbf{h}^{T}(t',\tau) d\tau \quad t' \ge t \quad (28)$$

is equal to the spectral density matrix of the stochastic processes to be simulated or to find $b(t,\tau)$ for the nonstationary case. *See* SIMULATION.

Regulator problem. The deterministic regulator problem is encountered in the optimal control of linear systems. Its state equation may be modified to include a white stochastic process $\mathbf{w}(t)$, as shown in Eq. (29). Here $\mathbf{x}(t)$ is the state vector, $\mathbf{u}(t)$ is the

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} + \mathbf{B}(t)\mathbf{u} + \mathbf{w}(t)$$
(29)
$$\mathbf{x}(t_0) \text{ specified}$$

input vector, and $\mathbf{w}(t)$ is a white stochastic process with correlation function given by Eq. (23). The performance index *J*, as defined by Eq. (30), where \mathbf{S}_1

$$J = \mathbf{x}^{T}(t_{1})\mathbf{S}_{1}\mathbf{x} (t_{1})$$
$$+ \int_{t_{0}}^{t_{1}} \mathbf{x}^{T}\mathbf{Q}(t)\mathbf{x} + \mathbf{u}^{T}\mathbf{R}(t)\mathbf{u} dt \quad (30)$$

and $\mathbf{Q}(t)$ are symmetric, nonnegative definite matrices which penalize the system for large excursions, depends now on $\mathbf{w}(t)$. Therefore its expected value or ensemble average over all possible $\mathbf{w}(t)$ must be used, Eq. (31).

Performance index =
$$E(J)$$
 (31)

See OPTIMAL CONTROL (LINEAR SYSTEMS).

As $\mathbf{w}(t)$ is white, its future value has no correlation with its past, and consequently has no correlation with $\mathbf{x}(t)$. Furthermore, $\mathbf{w}(t)$, $t > t_0$, is completely unpredictable and nothing can be done about it. The best that can be done is to minimize the contribution to E(J) by $\mathbf{x}(t)$. Thus, the optimal control law of Eq. (32) with **S** defined by the Riccati differential equation (33) and the boundary condition $\mathbf{S}(t_1) = \mathbf{S}_1$

$$\mathbf{u}(t) = -\mathbf{R}^{-1}(t)\mathbf{B}^{T}(t)\mathbf{S}(t)\mathbf{x}(t)$$
(32)

 $\dot{\mathbf{S}} + \mathbf{S}\mathbf{A}(t) + \mathbf{A}^{T}(t)\mathbf{S}$

$$-\mathbf{SB}(t)\mathbf{R}^{-1}(t)\mathbf{B}^{T}(t)\mathbf{S} + \mathbf{Q}(t) = 0 \quad (33)$$

is still valid. This heuristic reasoning has been substantiated by rigorous mathematical proof.

However, the minimum value of E(J) is different from that given by Eq. (34) because of the added

$$J_{\min}(\mathbf{x}(t_0), t_0) = \mathbf{x}^T(t_0)\mathbf{S}(t_0)\mathbf{x}(t_0)$$
(34)

disturbance $\mathbf{w}(t)$. Straightforward calculations give Eq. (35), where Tr stands for the trace of the matrix inside the braces.

$$\mathbf{E}(J)_{\min} = \mathbf{x}^{T}(t_{0})\mathbf{S}(t_{0})\mathbf{x}(t_{0}) + \int_{t_{0}}^{t_{1}} \operatorname{Tr} \{\mathbf{S}(t)\mathbf{W}(t)\} dt \quad (35)$$

In practical control systems, the state vector $\mathbf{x}(t)$ is not exactly known because of instrument noise and such, and the disturbance $\mathbf{w}(t)$ may not have white spectrum. However, with any given $Phi_W(j\omega)$ or $\phi_W(t,t')$, $\mathbf{w}(t)$ can be simulated as the output of linear filters with white noise processes $\mathbf{v}(t)$ as its inputs, as in Eqs. (36) and (37). A generalized vec-

$$\dot{\mathbf{w}}(t) = \mathbf{C}(t) \,\mathbf{w}(t) + \mathbf{v} \tag{36}$$

$$\phi_{\nu}(t,t') = \delta(t-t') \tag{37}$$

tor $\mathbf{z}(t)$ is defined by Eq. (38). If matrices **D**, **E**, and **F** are defined by Eqs. (39), where **1** and **0** denote

$$\mathbf{z}(t) = (\mathbf{x}(t)\mathbf{w}(t)) \tag{38}$$

$$\mathbf{D}(t) = \begin{pmatrix} \mathbf{A}(t) & \mathbf{1} \\ \mathbf{0} & \mathbf{C}(t) \end{pmatrix}$$

$$\mathbf{E}(t) = \begin{pmatrix} \mathbf{B}(t) \\ \mathbf{0} \end{pmatrix} \qquad \mathbf{F} = \begin{pmatrix} \mathbf{0} \\ \mathbf{1} \end{pmatrix}$$
(39)

the identity matrix and null matrix of appropriate dimensions, then Eqs. (29) and (36) can be combined into one, giving Eq. (40). Since **Fv** is a white distur-

$$\dot{\mathbf{z}}(t) = \mathbf{D}(t)\mathbf{z}(t) + \mathbf{E}(t)\mathbf{u} + \mathbf{F}\mathbf{v}$$
(40)

bance, the problem is reduced to an optimal regulator problem with white disturbance. However, $\mathbf{z}(t)$ is not completely known.

One of the most important results in estimator theory is that the residue error $\mathbf{z}(t) - \hat{\mathbf{z}}(t)$, where $\hat{\mathbf{z}}(t)$ is the optimum estimate of $\mathbf{z}(t)$, cannot be inferred in any way from past measurements. Consequently the best that can be done is to apply regulatory action to the known part of $\mathbf{z}(t)$, using Eq. (41), where

$$\mathbf{u}(t) = -\mathbf{R}^{-1}(t)\mathbf{E}^{T}(t)\mathbf{S}(t)\hat{\mathbf{z}}(t)$$
(41)

S(t) is the solution to Riccati equation for the new problem defined by Eq. (40), and $\hat{z}(t)$ is the best or optimum estimate of z(t). Equation (41) states that the optimal overall controller is obtained by applying the optimal control law to the separately optimally estimated state $\hat{z}(t)$. It is known as the separation theorem.

For single-input, single-output, time-invariant systems with infinite time interval, both the estimation problem and the control problem may be reformulated, leading to an integral equation known as the Wiener-Hopf equation. The solution to this equation has been obtained and is a classical mathematical result in the theory of stochastic processes. Further, the method discussed above as well as the Wiener-Hopf approach may also be applied to discrete-time systems. *See* ESTIMATION THEORY.

Structural uncertainty. In Eq. (29), the random variable $\mathbf{w}(t)$ is an additive term. Quite often the system structure is not completely known, and $\mathbf{A}(t)$ or $\mathbf{B}(t)$ has stochastic, uncertain, or unknown, components. These are known as structural uncertainty. If the structural uncertainty is white gaussian, the optimal control law remains the same, with one

or more terms added to the Riccati equation. If the structural uncertainty is nonwhite, or constant but unknown, it can be predicted to some extent from past measurements. Adaptive controllers are generally used. If the structural uncertainty varies in some unknown manner, minimax, least-sensitive, or guaranteed-cost controllers are used. Minimax means minimizing the largest possible value of the performance index. Least-sensitive means designing the controller in such a way that the performance index is not sensitive to the structural uncertainty. Guaranteed-cost aims at simple robust design which guarantees that the performance index does not exceed some given limit. *See* ADAPTIVE CONTROL; CON-TROL SYSTEMS; OPTIMAL CONTROL THEORY.

S. S. L. Chang; Gerald Cook

Bibliography. A. Bagchi, *Optimal Control of Stochastic Systems*, 1993; A. Benoussian, *Stochastic Control of Partially Observable Systems*, 1992; J. A. Borrie, *Stochastic Systems for Engineers*, 1992; G. R. Chen, G. Chen, and S.-H. Hsu, *Linear Stochastic Control Systems*, 1995; M. J. Grimble and M. A. Johnson, *Optimal Control and Stochastic Estimation: Theory and Applications*, 2 vols., 1988; F. L. Lewis and V. L. Syrmos, *Optimal Estimation: With an Introduction to Stochastic Control Theory*, 2d ed., 1995; R. F. Stengel, *Stochastic Optimal Control: Theory and Application*, 1986.

Stochastic process

A physical stochastic process is any process governed by probabilistic laws. Examples are (1) development of a population as controlled by mendelian genetics; (2) brownian motion of microscopic particles subjected to molecular impacts or, on a different scale, the motion of stars in space; (3) succession of plays in a gambling house; and (4) passage of cars by a specified highway point.

In each case, a probabilistic system is evolving; that is, its state is changing with time. Thus the state at time t depends on chance: It is a random variable x(t). The parameter set of values of t involved is usually (and will always be in this article) either an interval (continuous parameter stochastic process) or a set of integers (discrete parameter stochastic process). Some authors, however, apply the term stochastic process only to the continuous parameter case.

If the state of the system is described by a single number, x(t) is numerical-valued. In other cases, x(t)may be vector-valued or even more complicated. The discussion in this article will usually be restricted to the numerical case. As the state changes, its values determine a function of time, the sample function; and the probability laws governing the process determine the probabilities assigned to the various possible properties of sample functions.

A mathematical stochastic process is a mathematical structure inspired by the concept of a physical stochastic process, and studied because it is a mathematical model of a physical stochastic process or

because of its intrinsic mathematical interest and its applications both in and outside the field of probability. The mathematical stochastic process is defined simply as a family of random variables. That is, a parameter set is specified, and to each parameter point t a random variable x(t) is specified. If one recalls that a random variable is itself a function, if one denotes a point of the domain of the random variable x(t) by ω , and if one denotes the value of this random variable at ω by $x(t,\omega)$, it results that the stochastic process is completely specified by the function of the pair (t,ω) just defined, together with the assignment of probabilities. If t is fixed, this function of two variables defines a function of ω , namely, the random variable denoted by x(t). If ω is fixed, this function of two variables defines a function of t, a sample function of the process.

Probabilities are ordinarily assigned to a stochastic process by assigning joint probability distributions to its random variables. These joint distributions, together with the probabilities derived from them, can be interpreted as probabilities of properties of sample functions. For example, if t_0 is a parameter value, the probability that a sample function is positive at time t_0 is the probability that the random variable $x(t_0)$ has a positive value. The fundamental theorem at this level is that, to any self-consistent assignment of joint probability distributions, there corresponds a stochastic process.

The concept of a stochastic process is so general that the study of stochastic processes includes all of probability theory. Although it would be impossible to make a useful but more restrictive definition, what the probabilist usually has in mind in using the term stochastic process (unless he is interested in the mathematical foundations of the general theory) is a stochastic process whose random variables have some sort of interesting mutual relations. For example, one such relation is that of independence, and one type of stochastic process which was studied long before the term was invented is a sequence of independent random variables. For historical reasons, such a sequence is not commonly thought of as a stochastic process (although its continuous parameter analog, a process with independent increments, to be discussed below, is). The rest of this article is devoted to a general discussion of specific types of stochastic processes which have received the most attention, because they are important in mathematical and nonmathematical applications.

Stationary processes. These are the stochastic processes for which the joint distribution of any finite number of the random variables is unaffected by translations of the parameter; that is, the distribution of $x(t_1 + b), \ldots, x(t_n + b)$ does not depend on *b*. For a more complete discussion *see* PROBABILITY

Markov processes. A Markov process is a process for which, if the present is given, the future and past are independent of each other. More precisely, if $t_1 < \cdots < t_n$ are parameter values, and if 1 < j < n, then the sets of random variables $[x(t_1), \ldots, x(t_{j-1})]$ and $[x(t_{j+1}), \ldots, x(t_n)]$ are mutually independent for given $x(t_j)$. Equivalently, the conditional probability distribution of $x(t_n)$ for given $x(t_1), \ldots, x(t_{n-1})$ depends only on the specified value of $x(t_{n-1})$ and is in fact the conditional probability distribution of $x(t_n)$, given $x(t_{n-1})$. An important and simple example is the Markov chain, in which the number of states is finite or denumerably infinite. (The terminology varies somewhat here.) One simple type of discreteparameter Markov chain is the following. Let (p_{ij}) be a set of numbers, where *i* and *j* range over a finite or infinite set of integers. (In physical language, the number p_{ij} will be the probability that some system has a transition from state *i* to state *j* in one step.) The numbers p_{ij} are to satisfy relation (1). The ran-

$$p_{ij} \ge 0 \qquad \sum_{j} p_{ij} = 1 \tag{1}$$

dom variables of the associated Markov process are integral-valued, denoted by $x(0), x(1), \ldots$ If i_0 is prescribed as the initial state, that is, if x(0) is assigned the value i_0 identically, the probability that x(k) has the value i_k , for $k = 1, \ldots, N$, is product (2). For

$$p_{i_0i_1}\cdots p_{i_{N-1}i_N} \tag{2}$$

this process, p_{ij} is the probability that x(n + 1) has the value *j* if x(n) has the value *i*. The number p_{ij} is also (and this is the characteristic property of Markov processes) the probability that x(n + 1) has the value *j* if x(n) has the value *i* and if also x(n - 1) has any prescribed value $a_1, x(n - 2)$ the value a_2 , and so on. What makes this a special Markov chain, aside from the fact that the states are denoted by integers, is that the conditional probability just described does not depend on *n*. The chain is therefore described as having stationary transition probabilities. If the initial state is given a distribution, say by prescribing that x(0) have the value *i* with probability p_i , evaluation (2) becomes sum (3). Here the p_i 's are any nonnega-

$$\sum_{i} p_{i} p_{ii_1} \cdots p_{i_{N-1}i_N} \tag{3}$$

tive numbers with sum 1. If the initial distribution is chosen, as is not always possible, in such a way that the probability that x(n) has the value *j* is p_j , not only for n = 0 but for all values of *n*, the resulting process is stationary.

In constructing the corresponding continuousparameter Markov chain, it is supposed that, for each pair (i_j) , there is a function $p_{ij}(.)$, defined for strictly positive *t*, satisfying relations (4). The equations of

$$b_{ij} \ge (t)0 \qquad \sum_{j} p_{ij}(t) = 1 \\ p_{ij}(s+t) = \sum_{k} p_{ik}(s) p_{kj}(t)$$
(4)

the system in the last line are known as the Chapman-Kolmogorov equations. A Markov stochastic process with continuous parameter ranging from 0 to ∞ can be constructed for which, if again p_j is the probability that x(0) has the value j, and if $0 < t_1 < \cdots < t_n$, the probability that $x(t_k)$ has the value i_k for

 $k = 1, \dots, N$ is given by sum (5). For this process,

$$\sum_{i} p_{i} p_{ii_{1}}(t_{1}) p_{i_{1}i_{2}}(t_{2}-t_{1}) \cdots p_{i_{N-1}i_{N}}(t_{N}-t_{N-1})$$
 (5)

if s > 0, the probability that x(u + s) has the value *j*, if x(u) has the value *i*, is $p_{ij}(s)$. The number $p_{ij}(s)$ is also the probability (and this is the characteristic property of Markov processes) that x(u + s) has the value *j* if x(u) has the value *i*; and if also $x(u_1)$ has any specified value $a_1, x(u_2)$, the value a_2 , and so on, where u_1, u_2, \ldots , are any positive numbers less than *u*. This example is not the general continuousparameter Markov chain because the transition probability just described does not depend on *u*, that is, because the chain has stationary transition probabilities. The process is stationary if the probability that x(u) has the value *j* does not depend on *u*, for all *j*. The second line of relation (4) has a simple interpretation: The probability, if x(u) has the value *i*, that x(u + s + t) has the value *j*, is the sum over *k* of the probability that x(u + s) has the value k multiplied by the probability that, if x(u) has the value *i*, and if x(u + s) has the value k, then x(u + s + t) has the value j. Without the Markov property, the second factor might depend on *i*. If the number of states is not finite or denumerably infinite, the preceding discussion is modified by replacing sums in notations (1), (3), (4), and (5) by integrals.

Typical applications. Typical questions that have been raised, and solved to a varying degree, about Markov processes with stationary transition probabilities are the following. They are phrased in the continuous-parameter case, for the Markov chain just described, and it is assumed that relation (6) holds. For con-

$$\lim_{t \to 0} P_{ij}(t) = 1 \quad \text{for all } i \tag{6}$$

venience one defines $p_{ij}(0)$ as 1 if i = j and as 0 otherwise.

1. Does $p_{ij}(t)$ have a limit when $t \to \infty$? In other words, for each *j* is there a limiting probability that the system is in state *j* as time passes? The answer is yes, and the limiting probability depends only on the end state *j*, not on the initial state *i*, if transitions between all pairs of states are possible.

2. What are the asymptotic properties of $p_{ij}(t)$ as $t \rightarrow 0$? The answer is that $p'_{ij}(0)$ always exists and is finite except possibly when i = j. Under further hypotheses on the transition probability functions, always satisfied if there are only finitely many states, $p'_{ij}(0)$ is finite and relations (7) hold. These equations

$$p'_{ij}(t) = \sum_{k} p_{ik}(t) p'_{kj}(0)$$

$$p'_{ij}(t) = \sum_{k} p'_{ik}(0) p_{kj}(t)$$
(7)

can be used to determine the transition probability functions in terms of assigned derivatives when t = 0. For example, if *c* is a strictly positive constant, and if $p'_{ij}(0)$ is specified as *c* if j = i + 1, as -c if j = i, and as 0 otherwise, it is shown that $p_{ij}(t) = 0$ if j < i, and that otherwise relation (8) is true. The process

$$p_{ij}(t) = \frac{(ct)^{j-i}e^{-ct}}{(j-i)!}$$
(8)

with these transition probabilities is known as the Poisson process.

3. What are the properties of the sample functions? Under further restrictions on the process, the sample functions are constant on intervals, changing in jumps from one state to the next, and $-p'_{ii}(0)/p'_{ii}(0)$ is the probability that, if the system is in state i, its next jump will be into state *j*. If it is in the *i*th state, the time the system remains in this state thereafter is a random variable with density qe^{-tq} , where $q = -p'_{ii}(0)$. For example, in the case of the Poisson process described above, it is shown that, under proper normalization, and if x(0) = 0, the sample functions are integral-valued and monotone, increasing in unit jumps. This process is a mathematical model for the physical process of radioactive decay. That is, x(t) can be interpreted as the number of radioactive disintegrations of a substance by time t. In other interpretations, x(t) is taken as the number of telephone calls initiated by time t, or the number of cars that have passed a given highway point by time t. The constant c is the rate at which these various events occur. In fact, the expected value of x(u + b) - x(u), that is, the expected number of events in a time interval of length b (b here is of course positive) is cb, and the probability that an event will occur in an interval of length b, regardless of the past history of the process, is ch up to higher powers of b.

There are many special types of Markov chains, for which more detailed questions become important. For example, consider the branching processes. In a system of particles, all of the same type, a particle will occasionally split, independently of its past history and of the other particles, into $j \ge 0$ particles with probability q_i . If a particle is observed at time t, the probability that it will split by time t + b is cb, up to higher powers of *b*, where *c* is a strictly positive constant. Then the number of particles at time *t* is a random variable x(t). The x(t) process is a Markov process, and $p'_{ij}(0)$ is easily determined in terms of q_i and c. A more general branching process would permit particles of several types, and each particle would be allowed to split into particles of the various types. The rate c would depend on the particle type. In this case, x(t) is defined as a vector whose *i*th component is the number of particles of type *i* at time *t*. The x(t) stochastic process is a vector-valued Markov process. In studying branching processes, the most natural questions to ask are: What is the probability that the population will die out? If it does not die out, what is the asymptotic distribution of population as time passes? The answers are too technical to be given here.

Transition probabilities. If the states of a Markov process comprise all real numbers, the character of the process may be similar to that of a chain but may also be quite different. For example, the

sample functions of the process may be continuous. The most important examples of this type are the diffusion processes. Simplifying somewhat, but not assuming stationary transition probabilities, consider a Markov process for which the probability distribution of the state at time t, given state ξ at time s < t, has density $p(s, \xi, t, \eta)$. Then the basic conditions satisfied by the transition density, corresponding to relation (1), are relations (9). Now suppose that limits (10) and (11) exist and have the indicated

$$p(s,\xi,t,\eta) \ge 0 \qquad \int_{-\infty}^{\infty} p(s,\xi,t,\eta) \, d\eta = 1$$

$$p(s,\xi,t,\eta) = \int_{-\infty}^{\infty} p(s,\xi,u,\varsigma) p(u,\varsigma,t,\eta) \, d\xi$$
if $s < u < t$
(9)

$$\lim_{b \downarrow 0} \int_{-\infty}^{\infty} p(s,\xi,s+b,\eta) \left(\eta-\xi\right) \frac{d\eta}{b} = m(s,\xi) \quad (10)$$

$$\lim_{b \downarrow 0} \int_{-\infty}^{\infty} p(s,\xi,s+b,\eta) (\eta-\xi)^2 \frac{d\eta}{b} = \sigma^2(s,\xi) \quad (11)$$

values. These limit relations make m and σ^2 the instantaneous rates of change of the displacement and its variance, given a specified time and state. If mand σ are sufficiently regular, and if a further condition is imposed which, roughly, makes improbable significant sample function changes in short times, the corresponding Markov process, when properly normalized, will have continuous sample functions. Moreover, the transition density will then satisfy backward-diffusion equation (12) and forward diffusion equation (13), also known as the Fokker-Planck

$$\frac{\partial p(s,\xi,t,\eta)}{\partial s} + m(s,\xi) \frac{\partial p}{\partial \xi} + \frac{1}{2}\sigma^2(s,\xi)\frac{\partial^2 p}{\partial \xi^2} = 0 \quad (12)$$

$$\frac{\partial p(s,\xi,t,\eta)}{\partial s} + \frac{\partial}{\partial \eta} [m(t,\eta)p] - \frac{1}{2} \frac{\partial^2}{\partial \eta^2} [\sigma^2(t,\eta)p] = 0 \quad (13)$$

equation. Conversely, given a pair of coefficient functions m and 0 and σ these second-order parabolic equations can be used to derive the corresponding transition densities.

The simplest nontrivial example of a diffusion process corresponds to the specification m = 0 and σ a constant function. In this case, the diffusion process is the brownian motion process, or Wiener process: The increment x(t) - x(s) has a gaussian distribution with mean value 0 and variance $\sigma^2 | t - s |$. This is a mathematical model for the physical brownian motion. That is, if x(t) - x(s) represents the displacement in a given direction of a brownian particle between times *s* and *t*, this process is a good model for the actual motion.

Martingales. A martingale is a stochastic process with the property that, if $t_1 < \cdots < t_n$ are parameter values, the expected value of $x(t_n)$, for given $x(t_1), \cdots, x(t_{n-1})$ is equal to $x(t_{n-1})$. That is, the expected future value, given present and past values,

is equal to the present value. The interpretation that a martingale can be thought of as the fortune of a player after the successive plays of a fair gambling game is obvious.

Typical results on martingales are the following. If a sequence of random variables is a martingale, it converges under weak conditions which impose certain "bounds" on the random variables, for example, if the expectation of the absolute value of the *n*th random variable is bounded independently of *n*. The sample functions of a properly normalized continuous-parameter martingale do not have oscillatory discontinuities.

The applications of martingale theory are too technical to be given here, but one suggestive example will be given, which indicates at least how the theory can be usefully applied to information theory. Let $y, x_1, x_2, ...$ be random variables, and let y_n be the expected value of y knowing $x_1, ..., x_n$. Then y_n is a random variable which is a function of x_1 , $..., x_n$ and the sequence $y_1, y_2, ...$ is a martingale. That is, the expected values of a random variable, if one knows more and more, defines a sequence of random variables which is a martingale.

Processes with independent increments. Such a process is a continuous-parameter process with the property that, if $t_1 < \cdots < t_n$ are parameter values, the successive increments in notation (14) are mutu-

$$x(t_2) - x(t_1), \ldots, x(t_n) - x(t_{n-1})$$
 (14)

ally notation (14) independent. If $y(t) = x(t) - x(t_0)$, where t_0 is fixed, the y(t) process is then a Markov process. Both the Poisson and the brownian motion processes described above have independent increments.

Typical results on these processes include the following. If such a process is properly normalized, its sample functions do not have oscillatory discontinuities. Moreover, the distribution of any increment x(t) - x(s) is infinitely divisible, and there is a standard form for the characteristic function of any such distribution. *See* GAME THEORY; INFORMATION THE-ORY; LINEAR PROGRAMMING; OPERATIONS RESEARCH. Joseph L. Doob

Bibliography. R. N. Bhattacharya and E. C. Waymire, Stochastic Processes with Applications, 1990; J. L. Doob, Stochastic Processes, 1953, paper 1990; G. F. Lawler, Introduction to Stochastic Processes, 1995; J. Y. Medhi, Stochastic Processes, 2d ed., 1994; M. K. Ochi, Applied Probability and Stochastic Processes in Engineering and the Physical Sciences, 1990; A. Papoulis, Probability, Random Variables, and Stochastic Processes, 3d ed., 1991; P. Todorovic and J. O. Berger, An Introduction to Stochastic Processes and Their Applications, 1992.

Stoichiometry

The field of chemistry that includes all chemical measurements, such as the measurements of atomic and molecular weights and sizes, gas volumes, vapor densities, deviation from the gas laws, and the structure of molecules. In the long struggle to determine the relative weights of the atoms, scientists relied upon combining ratios, specific heats, and measurements of gas volumes. All such measurements, and the calculations that relate them to each other, constitute the field of stoichiometry. Since measurements are expressed in mathematical terms, stoichiometry can be considered to be the mathematics of general chemistry. Thus, stoichiometry is not part of inorganic, organic, physical, or analytical chemistry, but is an essential part of all of them. Chemistry is an exact science, and it depends upon exact measurements of weights, lengths, and volumes, and on the amounts of energy which are absorbed or evolved in chemical reactions.

In a more usual usage, the term stoichiometry refers to the relationships between the measured quantitites of substances or of energy involved in a chemical reaction; the calculations of these quantities include the assumption of the validity of the laws of definite proportions and of conservation of matter and energy. *See* CONSERVATION OF ENERGY; CONSERVATION OF MASS.

Laboratory measurements are made in terms of a unit of weight (grams, ounces, and such), volume (milliliters and such) or energy (ergs, calories, and such), but since atoms and molecules of different substances have different weights, it is necessary to convert numbers of atoms or molecules into the measured quantities. These quantities can be expressed in any convenient units. In laboratory work, weights are usually expressed in grams, but in chemical engineering, in pounds or tons. *See* ATOMIC MASS; GRAM-MOLECULAR WEIGHT; MOLECULAR WEIGHT.

A typical stoichiometric problem involves predicting the weight of reactant needed to produce a desired amount of a product in a chemical reaction. For example, phosphorus can be extracted from calcium phosphate, $Ca_3(PO_4)_2$, by a certain process with a 90% yield (some calcium phosphate fails to react or some phosphorus is lost). In a specific problem, it might be necessary to determine the mass of calcium phosphate required to prepare 16.12 lb of phosphorus by this process. The balanced equation for the preparation is shown in reaction (1). In this reac-

$$\begin{array}{c} 2\text{Ca}_3(\text{PO}_4)_2 + 10\text{C} + 6\text{SiO}_2 \rightarrow \\ 6\text{Ca}\text{SiO}_3 + 10\text{CO} + \text{P}_4 \quad (1) \end{array}$$

tion, 2 moles of calcium phosphate are required to produce 1 mole of phosphorus. Two moles of calcium phosphate have a mass of 620 lb, and 1 mole of phosphorus as P_4 has a mass of 124 lb. Using these relationships, calculation (2) is made. Since the yield

$$\frac{2 \text{ moles } Ca_3(PO_4)}{1 \text{ mole } P_4} = \frac{620 \text{ lb } Ca_3(PO_4)_2}{124 \text{ lb } P_4}$$
$$= \frac{X \text{ lb } Ca_3(PO_4)_2}{16.12 \text{ lb } P_4} \qquad (2)$$
$$X = 80.6 \text{ lb}$$

of phosphorus is only 90%, extra $Ca_3(PO_4)_2$ must be used: 88.1 lb is the mass of calcium phosphate required to yield 16.12 lb of phosphorus by this process.

Calculations of this sort are important in chemical engineering processes, in which amounts and yields of products must be known. The same reasoning is used in calculations of energy generated or required. In this case, the energy involved in the reaction of a known weight of the material in question must be known or determined. There are many variations on this theme. For example, if it is necessary to determine the composition of a mixture of copper and zinc powders, a given weight of the material can be dissolved in hydrochloric acid, which reacts with the zinc, but not with the copper [reaction (3)]. From

$$Zn + 2HCI \rightarrow ZnCI_2 + H_2$$
 (3)

the volume of hydrogen gas liberated, the amount of zinc dissolved can be calculated, and hence the composition of the mixture. Again, if 5.0 g of sodium is placed in a vessel containing 5.0 g of chlorine, a violent reaction takes place [reaction (4)]. However,

$$2Na + Cl_2 \rightarrow 2NaCl \tag{4}$$

some of the sodium remains unattacked. It can be calculated that 5.0 g of chlorine can react with only 3.2 g of sodium; 1.8 g of sodium will remain. The chlorine is said to be the limiting reagent—when it is used up, the reaction must stop. *See* AVOGADRO'S LAW; COMBINING VOLUMES, LAW OF; DALTON'S LAW; DEFINITE COMPOSITION, LAW OF; MULTIPLE PROPORTIONS, LAW OF; SPECIFIC HEAT.

The calculations discussed in this article involve compounds in which the ratio of atoms is generally simple. For a discussion of compounds in which the relative number of atoms cannot be expressed as ratios of small whole numbers. *See* NONSTOICHIOMET-RIC COMPOUNDS.. John C. Bailar, Jr.

Bibliography. J. A. Olmsted and G. M. Williams, *Chemistry*, 4th ed., 2004; K. C. Timberlake, *Chemistry*, 9th ed., 2005.

Stoker

A mechanical means for feeding coal into, and for burning coal in, a furnace. There are three basic types of stokers. Chain or traveling-grate stokers have a moving grate on which the coal burns; they carry the coal from a hopper into the furnace and move the ash out. Spreader stokers mechanically or pneumatically distribute the coal from a hopper at the furnace front wall and move it onto the grate which usually moves continuously to dispose of the ash after the coal is burned. Underfeed stokers are arranged to force fresh coal from the hopper to the bottom of the burning coal bed, usually by means of a screw conveyor. The ash is forced off the edges of the retort peripherally to the ashpit or is removed by hand. See STEAM-GENERATING FURNACE. George W. Kessler

Pages from



Stokes' theorem The assertion that under certain light restrictions the surface integral of $(\Delta_{\widetilde{\mathcal{C}}} \times \widetilde{\mathcal{C}}) \cdot \mathbf{v}$ over a surface patch *S* is equal to the line integral of $\widetilde{\mathcal{C}} \cdot \tau$ taken around *C*, the boundary curve of *S*, provided the sense of transcription of *C* is right-handed relative to **v**. This can be expressed as

ISBN 0071429573

5th Edition (2005)

$$\iint (\nabla \times \mathfrak{F}) \cdot \mathbf{v} \, dS = \int \mathfrak{F} \cdot \tau \, ds$$

Here \mathfrak{F} is a vector function, \mathbf{v} is one of the two unit normals to the two-sided surface S, s is arc length measured positively in the sense which is right-handed relative to \mathbf{v} , and τ is the unit tangent vector to C in the sense of increasing s. See CALCULUS OF VECTORS. [H.V.C.]

Stolonifera An order of the Alcyonaria which lacks coenenchyme. They form either simple or rather complex colonies. The polyp has a cylindrical body with a retractile oral portion which can withdraw into a solid anthostele or calyx protected by many calcareous spicules. The base of the mature polyp is attached to a creeping stolon which is a ribbonlike network or thin flat mat from which daughter polyps arise. Daughter polyps never bud from the wall of the primary polyp. Each polyp is connected by solenial tubes of the stolons, or by transverse platforms. *See* ALCYONARIA. [K.At.]

Stomach The tubular or saccular abdominal organ of the digestive system adapted for temporary food storage and preliminary stages of food breakdown.

In some primitive vertebrates the stomach may be little more than a simple tube quite similar to other portions of the gastrointestinal tract. In other forms the stomach is a distinct, and frequently large, saclike structure of variable shape. Carnivorous forms typically have a better-developed stomach than herbivores, probably reflecting the larger but less numerous feedings characteristic of the former, but exceptions are numerous.

In birds the stomach consists of a proventriculus and a gizzard. The former is well supplied with glands which secrete softening and digestive enzymes; the latter is a strong, muscular grinding organ whose action is often enhanced by the ingestion of small stones.

Mammals have stomachs which vary considerably in structure. Although a single chamber is most common, some mammals, such as cows and their relatives (ruminants), have as many as four. These chambers may have developed either from modifications of the posterior portions of the esophagus or from alterations of the stomach itself.

The human stomach is located beneath the diaphragm, through which the posterior, terminal end of the esophagus passes. The stomach appears as a dilated tube continuous with the distal end of the esophagus. The upper curvature of the stomach is usually above and to the left of the esophageal orifice. This expanded anterior portion is the fundus and is commonly filled with air or gas. The body (corpus) of the stomach is directed toward the attenuated right extremity or pyloric region and is subject to variations in size and shape, depending upon functional activities, habits, disease, and volume of the contents. The pyloric walls are marked by the heavy sphincter muscle which controls the passage of chyme (a semiliquid fluid produced by the mechanical and chemical changes of preliminary digestion) into the duodenum.

The stomach of vertebrates is lined by a mucous membrane that is usually thrown into longitudinal folds called rugae. Most of the surface is covered with mucus-secreting epithelial cells, but scattered throughout the lining are many small glandular pits which are lined with one or more types of secretory cells. See DIGESTIVE SYSTEM. [T.S.P.] **Stomatopoda** The only order of the superorder Hoplocarida belonging to the subclass Malacostraca of the class Crustacea. This order of the mantis shrimps contains four families: Lysioquillidae, Bathysquillidae, Squillidae, and Gonodactylidae. Stomatopoda are among the larger Crustacea; the size of an adult ranges from about 0.5 in. (12 mm) to over 12 in. (300 mm).

The bodies of stomatopods are narrow and elongate, almost eruciform. Only part of the cephalic and thoracic somites are fused and covered by the dorsal shield, the carapace. Those cephalic somites which bear the antennulae and the eyes are free and visible anterior to the carapace, and the last four thoracic somites are similarly exposed. The tail fan consists of a well-developed, sometimes peculiarly sculptured or deformed, median plate, the telson, and the two uropods. Anteriorly, the carapace bears a flattened movable plate, the rostrum. The eyes are large, stalked, and movable. Of eight pairs of thoracic appendages the first is narrow, slender, and hairy, probably being used for cleaning purposes. The second thoracic leg is very strong and heavy. It has become a large raptorial claw that shows a great resemblance to that of the praying mantis, and for this reason the Stomatopoda are often given the name mantis shrimps.

The Stomatopoda are marine animals rarely found in brackish water. Most species are confined to tropical and subtropical areas, though some occur in the boreal and antiboreal regions. The majority of stomatopods live in the littoral and sublittoral zones, but a few species have been found in greater depths, down to 2500 ft (760 m). See MALACOSTRACA. [L.B.Ho.]

Stone and stone products The term stone is applied to rock that is cut, shaped, broken, crushed, and otherwise physically modified for commercial use. The two main divisions are dimension stone and crushed stone. Other descriptive terms may be used, for example, building stone, roofing stone, or precious stone. *See* GEM; ROCK.

The term dimension stone is applied to blocks that are cut and milled to specified sizes, shapes, and surface finishes. The principal uses are for building and ornamental applications. Granites, limestones, sandstones, and marbles are widely used; basalts, diabases, and other dark igneous rocks are used less extensively. Soapstone is used to some extent. Rock suitable for use as dimension stone must be obtainable in large, sound blocks, free from incipient cracks, seams, and blemishes, and must be without mineral grains that might cause stains as a result of weathering. It must have an attractive color, and generally a uniform texture.

Slate differs from other dimension stone because it can be split into thin sheets of any desired thickness. Commercial slate must be uniform in quality and texture and reasonably free from knots, streaks, or other imperfections, and have good splitting properties. Roofing slates are important products of most slate quarries. However, the roofing-slate industry has declined considerably because of competition from other types of roofing. Slate is also used for milled products such as blackboards, electrical panels, window and door sills and caps, baseboards, stair treads, and floor tile. See SLATE.

Nearly all the principal types of stone—granite, diabase, basalt, limestone, dolomite, sandstone, and marble—may be used as sources of commercial crushed stone; limestone is by far the most important. Crushed stone is made from sound, hard stone, free from surface alteration by weathering. Stone that breaks in chunky, more or less cubical fragments is preferred. Commercial stone should be free from certain deleterious impurities, such as opalescent quartz, and free from clay or silt. Crushed stone is used principally as concrete aggregate, as road stone, or as railway ballast. Other uses for limestone are as a fluxing material to remove impurities from ores smelted in metallurgical furnaces, in the manufacture of alkali chemicals, calcium carbide, glass, paper, paint, and sugar, and for filter beds and for making mineral wool. [R.L.B.]

2120 Storage tank

Storage tank A container for storing liquids or gases. A tank may be constructed of ferrous or nonferrous metals or alloys, reinforced concrete, wood, or filament-wound plastics, depending upon its use. Tanks resting on the ground have flat bottoms: those supported on towers have either flat or curved bottoms. Standpipes, which are usually cylindrical shells of steel or reinforced concrete resting on the ground, are frequently of great height and comparatively small diameter. They are built to contain water for a distribution system, and height is required to maintain pressure in the system. Tanks for other liquids and for gases, where storage is more important than pressure, are generally lower and of greater diameter. [C.M.A.]

Storm An atmospheric disturbance involving perturbations of the prevailing pressure and wind fields on scales ranging from tornadoes (0.6 mi or 1 km across) to extratropical cyclones (1.2–1900 mi or 2–3000 km across); also, the associated weather (rain storm, blizzard, and the like). Storms influence human activity in such matters as agriculture, transportation, building construction, water impoundment and flood control, and the generation, transmission, and consumption of electric energy. See WIND.

The form assumed by a storm depends on the nature of its environment, especially the large-scale flow patterns and the horizontal and vertical variation of temperature; thus the storms most characteristic of a given region vary according to latitude, physiographic features, and season. Extratropical cyclones and anticyclones are the chief disturbances over roughly half the Earth's surface. Their circulations control the embedded smaller-scale storms. Large-scale disturbances of the tropics differ fundamentally from those of extratropical latitudes. *See* HURRICANE; SQUALL; THUNDERSTORM; TORNADO; TROPICAL METEO-ROLOGY.

Cyclones form mainly in close proximity to the jet stream, that is, in strongly baroclinic regions where there is a large increase of wind with height. Weather patterns in cyclones are highly variable, depending on moisture content and thermodynamic stability of air masses drawn into their circulations. Warm and occluded fronts, east of and extending into the cyclone center, are regions of gradual upgliding motions, with widespread cloud and precipitation but usually no pronounced concentration of stormy conditions. Extensive cloudiness also is often present in the warm sector. Passage of the cold front is marked by a sudden wind shift, often with the onset of gusty conditions, with a pronounced tendency for clearing because of general subsidence behind the front. Showers may be present in the cold air if it is moist and unstable because of heating from the surface. Thunderstorms, with accompanying squalls and heavy rain, are often set off by sudden lifting of warm, moist air at or near the cold front, and these frequently move eastward into the warm sector. See Cyclone; JET STREAM; WEATHER.

Extratropical cyclones alternate with high-pressure systems or anticyclones, whose circulation is generally opposite to that of the cyclone. The circulations of highs are not so intense as in well-developed cyclones, and winds are weak near their centers. In low levels the air spirals outward from a high; descent in upper levels results in warming and drying aloft. Anticyclones fall into two main categories, the warm "subtropical" and the cold "polar" highs.

Between the scales of ordinary air turbulence and of cyclones, there exist a variety of circulations over a middle-scale or mesoscale range, loosely defined as from about one-half up to a few hundred miles. Alternatively, these are sometimes referred to as subsynoptic-scale disturbances because their dimensions are so small that they elude adequate description by the ordinary synoptic network of surface weather stations. Thus their detection often depends upon observation by indirect sensing systems. *See* METEOROLOGICAL SATELLITES; RADAR METEOROLOGY; STORM DETECTION. [C.W.N.] **Storm detection** Identifying storm formation, monitoring subsequent storm evolution, and assessing the potential for destruction of life and property through application of various methods and techniques. Doppler radars, satellite-borne instruments, lightning detection networks, and surface observing networks are used to detect the genesis of storms, to diagnose their nature, and to issue warnings when a threat to life and property exists. *See* STORM.

Radar surveillance. Radars emit pulses of electromagnetic radiation that are broadcast in a beam, whose angular resolution is about 1° with a range resolution of about 0.5 km (0.3 mi). The radar beam may intercept precipitation particles in a storm that reflect a fraction of the transmitted energy to the transmitter site (generally called reflectivity or the scatter cross section per unit volume). As the transmitter sweeps out a volume by rotating and tilting the transmitting antenna, the reflectivity pattern of the precipitation particles embodied in the storm is defined. Doppler radars also can measure the velocities of precipitation particles along the beam (radial velocity). Reflectivity and velocity patterns of the storm hydrometeors then make it possible to diagnose horizontal and vertical circulations that may arise within the storm, and to estimate the type and severity of weather elements attending the storm, such as rainfall, hail, damaging winds, and tornadoes. See DOPPLER RADAR; PRECIPITATION (METEOROLOGY); RADAR; RADAR METEOROLOGY; WIND.

Satellite surveillance. Since the early 1960s, meteorological data from satellites have had an increasing impact on storm detection and monitoring. In December 1966 the first geostationary *Applications Technology Satellite (ATS 1)* allowed forecasters in the United States to observe storms in animation. A Geostationary Operational Environmental Satellite (GOES) program was initiated within the National Oceanic and Atmospheric Administration (NOAA) with the launch of GOES 1 in October 1975. The visible and infrared spin scan radiometer (VISSR) provided imagery, which significantly advanced the ability of meteorologists to detect and observe storms by providing frequent-interval visible and infrared imagery of the Earth surface, cloud cover, and atmospheric moisture patterns.

The first of NOAA's next generation of geostationary satellites, GOES 8 was launched in the spring of 1994. GOES 8 introduced improved capabilities to detect and observe storms. The GOES 8 system includes no conflict between imaging and sounding operation, multispectral imaging with improved resolution and better signal-to-noise in the infrared bands, and more accurate temperature and moisture soundings of the storm environment. The Earth's atmosphere is observed nearly continuously.

Derived-product images showing fog and stratus areas from *GOES 8* are created by combining direct satellite measurements, such as by subtracting brightness temperatures at two different wavelengths. *GOES 8* shows the fog and stratus much more clearly because of its improved resolution. This capability enables forecasters to detect boundaries between rain-cooled areas having fog or low clouds, and clear areas. Such boundaries are frequently associated with future thunderstorm development. The sounder on *GOES 8* is capable of fully supporting routine forecasting operations. This advanced sounding capability consists of better vertical resolution in both temperature and moisture, and improved coverage of soundings in and around cloudy weather systems. *See* CLOUD; FOG; METEOROLOGICAL SATELLITES; SATELLITE METEOROLOGY.

Surface observing systems. Larger convective storm systems such as squall lines and mesoscale convective systems can be detected (but not fully described) by the temperature, moisture, wind, and pressure patterns observed by appropriate surface instrumentation. Automatic observing systems provide frequent data on pressure, temperature, humidity, wind, cloud base, and most precipitation types, intensity, and accumulation. Analyses of these data, combined with improved conceptual models of convective storm systems, enable forecasters

to detect and monitor the intense mesoscale fluctuations in pressure and winds that often accompany the passage of convective weather systems such as bow echoes, derechos (strong, straightline winds), and squall lines. A bow echo is a specific radar reflectivity pattern associated with a line of thunderstorms. The middle portion of the thunderstorm line is observed to move faster than the adjacent portions, causing the line of storms to assume a bowed-out configuration. Other analyses of these mesoscale data fields aid the forecaster in detecting favorable areas for thunderstorm cell regeneration, which may produce slowly moving mesoscale convective storms attended by heavy rains and flash floods. See SQUALL LINE; WEATHER OBSERVATIONS.

Cloud-to-ground lightning detectors. Lightning location stations provide forecasters with the location, polarity, peak current, and number of strokes in a flash to ground within seconds of the flash occurrence. Useful applications have emerged with regard to the detection and tracking of thunderstorms, squall lines, other mesoscale convective systems, and the weather activity that accompany these phenomena, such as tornadoes and hail. *See* LIGHTNING; MESOMETEOROLOGY; SFERICS; THUNDER-STORM; WEATHER FORECASTING AND PREDICTION. [C.F.Ch.]

Storm electricity Processes responsible for the separation of positive and negative electric charges in the atmosphere during storms, including the spectacular manifestation of this charge separation: lightning discharges. Cloud electrification is almost invariably associated with convective activity and with the formation of precipitation in the form of liquid water (rain) and ice particles (graupel and hail). The most vigorous convection and active lightning occurs in the summertime, when the energy source for convection, water vapor, is most prevalent. Winter snowstorms can also be strongly electrified, but they produce far less lightning than summer storms. Electrified storm clouds occasionally occur in complete isolation; more commonly they are found in convective clusters or in lines that may extend horizontally for hundreds of kilometers. *See* PRECIPITATION (METEO-ROLOGY).

Measurements of electric field at the ground and from instrumented balloons within thunderclouds have disclosed an electrostatic structure that appears to be fairly systematic throughout the world. The measurements show that the principal variations in charge occur in the vertical and are affected by the temperature of the cloud. The charge structure within a thundercloud is tripolar, with a region of dominant negative charge sandwiched between an upper region of positive charge and a subsidiary lower region of positive charge. In addition to the charge accumulations described within the cloud, electrical measurements disclose the existence of charge-screening layers at the upper cloud boundary and a layer of positive charge near the Earth's surface beneath the cloud. These secondary charge accumulations arising from charge migration outside the cloud are caused by electrostatic forces of attraction set up by the charges within the cloud.

Large differences of electric potential are associated with the distribution of charge maintained by active thunderclouds. These large differences in potential are maintained by charging currents that result from the motions of air and particles. The charging currents range from milliamperes in small clouds that are not producing lightning to several amperes for large storms with high rates of lightning. *See* CLOUD PHYSICS.

In response to charge separation within a thundercloud, the electric field increases to a value of approximately 10^6 V/m (300,000 V/ft) at which point dielectric breakdown occurs and lightning is initiated. Most lightning extends through the cloud at speeds of 10,000–100,000 m/s (22,000–220,000 mi/h). The peak temperature of lightning, which is a highly ionized plasma, may exceed 30,000 K (54,000°F). The acoustic disturbance caused by the sudden heating of the atmosphere by lightning is thunder. *See* LIGHTNING; THUNDER.

Meteorologists have shown a growing interest in the largescale display of real-time lightning activity, since lightning is one of the most sensitive indicators of convective activity. Research has expanded into relationships between lightning characteristics and the meteorological evolution of different types of storms. The discovery of the sensitive dependence of local lightning activity on the temperature of surface air has led to research focused on the use of the global electrical circuit as a diagnostic for global temperature change. *See* ATMOSPHERIC ELECTRICITY; STORM DE-TECTION; THUNDERSTORM. [E.Wi.]

Storm surge An anomalous rise in water elevations caused by severe storms approaching the coast. A storm surge can be succinctly described as a large wave that moves with the storm that caused it. The surge is intensified in the nearshore, shallower regions where the surface stress caused by the strong onshore winds pile up water against the coast, generating an opposing pressure head in the offshore direction. However, there are so many other forces at play in the dynamics of the storm surge phenomenon, such as bottom friction, Earth's rotation, inertia, and interaction with the coastal geometry, that a simple static model cannot explain all the complexities involved. Scientists and engineers have dedicated many years in the development and application of sophisticated computer models to accurately predict the effects of storm surges.

The intensity and dimension of the storm causing a surge, and thus the severity of the ensuing surge elevations, depend on the origin and atmospheric characteristics of the storm itself. Hurricanes and severe extratropical storms are the cause of most significant surges. In general, hurricanes are more frequent in low to middle latitudes, and extratropical storms are more frequent in middle to high latitudes. *See* HURRICANE; STORM. [S.R.S.]

Straight-line mechanism A mechanism that produces a straight-line (or nearly so) output motion from an input element that rotates, oscillates, or moves in a straight line. Common machine elements, such as linkages, gears, and cams, are often used in ingenious ways to produce the required controlled motion. The more elegant designs use the properties of special points on one of the links of a four-bar linkage. *See* MECHANISM.

Four-bar linkages that generate approximate straight lines are not new. In 1784 James Watt applied the concept to the verticalcylinder beam engine. By selecting the appropriate link lengths, the designer can easily develop a mechanism with a high-quality approximate straight line. Contemporary kinematicians have contributed to more comprehensive studies of the properties of the mechanisms that generate approximate straight lines. The work not only describes the various classical mechanisms, but also provides design information on the quality (the amount of deviation from a straight line) and the length of the straight-line output. See FOUR-BAR LINKAGE; LINKAGE (MECHANISM).

Gears can also be used to generate straight-line motions. The most common combination would be a rack-and-pinion gear. *See* GEAR.

Cam mechanisms are generally not classified as straight-line motion generators, but translating followers easily fall into the classical definition. See CAM MECHANISM. [J.A.Sm.]

Strain gage A device which measures mechanical deformation (strain). Normally it is attached to a structural element, and uses the change of electrical resistance of a wire or semiconductor under tension. Capacity, inductance, and reluctance are also used.

The strain gage converts a small mechanical motion to an electrical signal by virtue of the fact that when a metal (wire or foil) or semiconductor is stretched, its resistance is increased. The change in resistance is a measure of the mechanical motion. In addition to their use in strain measurement, these gages are used in sensors for measuring the load on a mechanical member,

2122 Strake

forces due to acceleration on a mass, or stress on a diaphragm or bellows. See STRAIN. [J.H.Z.]

Strake The slender forward extension of the inboard region of the wing of a combat aircraft used to provide increased lift in the high angle-of-attack maneuvering condition. In contrast to the normal attached-flow design principles, these strakes are built to allow the flow to separate along the leading edge in the high angle-of-attack range and to roll up into strong leading-edge vortices. The illustration shows a typical strake installation and the vortices generated. These are highly stable, and



Strake and resulting vortex flow.

their strong swirling motion creates a lower pressure area on the strake upper surface, resulting in a large incremental increase in lift force known as vortex lift. This vortex lift increases rapidly in the high angle-of-attack range and is less susceptible to the normal stalling characteristics encountered with conventional lifting surfaces. *See* AIRFOIL; SUBSONIC FLIGHT; VORTEX; WING. [E.C.Po.]

Strand line The line that marks the separation of land and water along the margin of a pond, lake, sea, or ocean; also called the shoreline. The strand line is very dynamic. It changes with the tides, storms, and seasons, and as long-term sea-level changes take place. The sediments on the beach respond to these changes, as do the organisms that live in this dynamic environment. On a beach organisms move with the tides, and on a rocky coast they tend to be organized relative to the strand line because of special limitations or adaptations to exposure.

Geologists who study ancient coastal environments commonly try to establish where the strand line might be in the rock strata. This can sometimes be determined by a combination of the nature and geometry of individual laminations in the rock, by identifying sedimentary structures that occur at or near the strand line. The most indicative of these structures are swash marks, which are very thin accumulations of sand grains that mark the landward uprush of a wave on the beach. *See* FACIES (GEOLOGY); PALEOGEOGRAPHY; STRATIGRAPHY. [R.A.D.]

Strange particles Bound states of guarks, in which at least one of these constituents is of the strange (s) type. Strange quarks are heavier than the up (u) and down (d) quarks, which form the neutrons and protons in the atomic nucleus. Neutrons (udd) and protons (uud) are the lightest examples of a family of particles composed of three guarks, known as baryons. These and other composite particles which interact dominantly through the strong (nuclear) force are known as hadrons. The first strange hadron discovered (in cosmic rays in 1947) was named the lambda baryon, Λ ; it is made of the three-guark combination uds. A baryon containing a strange quark is also called a hyperon. Although strange particles interact through the strong (nuclear) force, the strange quark itself can decay only by conversion to a quark of different type (such as *u* or *d*) through the weak interaction. For this reason, strange particles have very long lifetimes, of the order of 10^{-10} s, compared to the lifetimes

of the order of 10^{-23} s for particles which decay directly through the strong interaction. This long lifetime was the origin of the term strange particles. *See* BARYON; HADRON; NEUTRON; PROTON; STRONG NUCLEAR INTERACTIONS.

In addition to strange baryons, strange mesons occur. The lightest of these are the kaons ($K^+ = u\bar{s}$ and $K^0 = d\bar{s}$) and the antikaons ($\bar{K}^0 = s\bar{d}$ and $\bar{K}^- = s\bar{u}$). Kaons and their antiparticles have been very important in the study of the weak interaction and in the detection of the very weak *CP* violation, which causes a slow transition between neutral kaons and neutral antikaons. See ELEMENTARY PARTICLE; MESON; QUARKS. [T.Bar.]

Strangles A highly contagious disease of the upper respiratory tract of horses and other members of the family Equidae, characterized by inflammation of the pharynx and abscess formation in lymph nodes. This disease occurs in horses of all ages throughout the world. The causative agent is *Streptococcus equi*, a clonal pathogen apparently derived from an ancestral strain of *S. zooepidemicus*. It is an obligate parasite of horses, donkeys, and mules. *See* STREPTOCOCCUS.

Strangles is most common and most severe in young horses, and is very prevalent on breeding farms. The causative agent has been reported to survive for 7 weeks in pus but dies in a week or two on pasture. Transmission is either direct by nose or mouth contact or aerosol, or indirect by flies, drinking buckets, pasture, and feed. The disease is highly contagious under conditions of crowding, exposure to severe climatic conditions such as rain and cold, and prolonged transportation. Carrier animals, although of rare occurrence, are critical in maintenance of the streptococcus and in initiation of outbreaks.

The mean incubation period is about 10 days, with a range of 3-14 days. The animal becomes quieter, has fever of $39-40.5^{\circ}$ C ($102-105^{\circ}$ F), nasal discharge, loss of appetite, and swelling of one or more lymph nodes of the mouth. Pressure of a lymph node on the airway may cause respiratory difficulty. Abscesses in affected lymph nodes rupture in 7-14 days, and rapid clinical improvement and recovery then ensues. Recovery is associated with formation of protective antibodies in the nasopharynx and in the serum. See ANTIBODY.

Streptococcus equi is easily demonstrated in smears of pus from abscesses and in culture of pus or nasal swabs on colistin– nalidixic acid blood agar. Acutely affected animals also show elevated white blood cell counts and elevated fibrinogen.

Commercially available vaccines are injected in a schedule of two or three primary inoculations followed by annual boosters. However, the clinical attack rate may be reduced by only 50%, a level of protection much lower than that following the naturally occurring disease. See IMMUNITY.

Procaine penicillin G is the antibiotic of choice and quickly brings about reduction of fever and lymph node enlargement. See BIOLOGICALS. [J.F.T.]

Stratigraphy A discipline involving the description and interpretation of layered sediments and rocks, and especially their correlation and dating. Correlation is a procedure for determining the relative age of one deposit with respect to another. The term "dating" refers to any technique employed to obtain a numerical age, for example, by making use of the decay of radioactive isotopes found in some minerals in sedimentary rocks or, more commonly, in associated igneous rocks. To a large extent, layered rocks are ones that accumulated through sedimentary processes beneath the sea, within lakes, or by the action of rivers, the wind, or glaciers; but in places such deposits contain significant amounts of volcanic material emplaced as lava flows or as ash ejected from volcanoes during explosive eruptions. *See* DATING METHODS; IGNEOUS ROCKS; ROCK AGE DETERMINATION; SEDIMENTARY ROCKS.

Sedimentary successions are locally many thousands of meters thick owing to subsidence of the Earth's crust over millions of years. Sedimentary basins therefore provide the best available record of Earth history over nearly 4 billion years. That record includes information about surficial processes and the varying environment at the Earth's surface, and about climate, changing sea level, the history of life, variations in ocean chemistry, and reversals of the Earth's magnetic field. Sediments also provide a record of crustal deformation (folding and faulting) and of largescale horizontal motions of the Earth's lithospheric plates (continental drift). Stratigraphy applies not only to strata that have remained flat-lying and little altered since their time of deposition, but also to rocks that may have been strongly deformed or recrystallized (metamorphosed) at great depths within the Earth's crust, and subsequently exposed at the Earth's surface as a result of uplift and erosion. As long as original depositional layers can be identified, some form of stratigraphy can be undertaken. See BASIN; CONTINENTAL DRIFT; FAULT AND FAULT STRUCTURES; SEDIMEN-TOLOGY.

An important idea first articulated by the Danish naturalist Nicolaus Steno in 1669 is that in any succession of strata the oldest layer must have accumulated at the bottom, and successively younger layers above. It is not necessary to rely on the present orientation of layers to determine their relative ages because most sediments and sedimentary rocks contain numerous features, such as current-deposited ripples, minor erosion surfaces, or fossils of organisms in growth position, that have a well-defined polarity with respect to the up direction at the time of deposition (so-called geopetal indicators). This principle of superposition therefore applies equally well to tilted and even overturned strata. Only where a succession is cut by a fault is a simple interpretation of stratigraphic relations not necessarily possible, and in some cases older rocks may overlie younger rocks structurally. See DEPOSITIONAL SYSTEMS AND ENVIRONMENTS.

The very existence of layers with well-defined boundaries implies that the sedimentary record is fundamentally discontinuous. Discontinuities are present in the stratigraphic record at a broad range of scales, from that of a single layer or bed to physical surfaces that can be traced laterally for many hundreds of kilometers. Large-scale surfaces of erosion or nondeposition are known as unconformities, and they can be identified on the basis of both physical and paleontological criteria. *See* PALEONTOLOGY; UNCONFORMITY.

Most stratal discontinuities possess time-stratigraphic significance because strata below a discontinuity tend to be everywhere older than strata above. To the extent that unconformities can be recognized and traced widely within a sedimentary basin, it is possible to analyze sedimentary rocks in a genetic framework, that is, with reference to the way they accumulated. This is the basis for the modern discipline of sequence stratigraphy, so named because intervals bounded by unconformities have come to be called sequences.

Traditional stratigraphic analysis has focused on variations in the intrinsic character or properties of sediments and rocksproperties such as composition, texture, and included fossils (lithostratigraphy and biostratigraphy)—and on the lateral tracing of distinctive marker beds such as those composed of ash from a single volcanic eruption (tephrostratigraphy). The techniques of magnetostratigraphy and chemostratigraphy are also based on intrinsic characteristics, although these techniques require sophisticated laboratory analysis. Sequence stratigraphy attempts to integrate these approaches in the context of stratal geometry, thereby providing a unifying framework in which to investigate the time relations between sediment and rock bodies as well as to measure their numerical ages (chronostratigraphy and geochronology). Seismic stratigraphy is a variant of the technique of sequence stratigraphy in which unconformities are identified and traced in seismic reflection profiles on the basis of reflection geometry. See GEOCHRONOMETRY; SEISMIC STRATIG-RAPHY; SEISMOLOGY.

Conventional stratigraphy currently recognizes two kinds of stratigraphic unit: material units, distinguished on the basis of some specified property or properties or physical limits; and temporal or time-related units. A common example of a material unit is the formation, a lithostratigraphic unit defined on the basis of lithic characteristics and position within a stratigraphic succession. Each formation is referred to a section or locality where it is well developed (a type section), and assigned an appropriate geographic name combined with the word formation or a descriptive lithic term such as limestone, sandstone, or shale (for example, Tapeats Sandstone). Some formations are divisible into two or more smaller-scale units called members and beds. In other cases, formations of similar lithic character or related genesis are combined into composite units called groups and supergroups.

Sequence stratigraphy differs from conventional stratigraphy in two important respects. The first is that basic units (sequences) are defined on the basis of bounding unconformities and correlative conformities rather than material characteristics or age. The second is that sequence stratigraphy is fundamentally not a system for stratigraphic classification, but a procedure for determining how sediments accumulate. See SEQUENCE STRATIGRAPHY. [N.C.B.]

Stratosphere The atmospheric layer that is immediately above the troposphere and contains most of the Earth's ozone. Here temperature increases upward because of absorption of solar ultraviolet light by ozone. Since ozone is created in sunlight from oxygen, a by-product of photosynthesis, the stratosphere exists because of life on Earth. In turn, the ozone layer allows life to thrive by absorbing harmful solar ultraviolet radiation. The mixing ratio of ozone is largest (10 parts per million by volume) near an altitude of 30 km (18 mi) over the Equator. The distribution of ozone is controlled by solar radiation, temperature, wind, reactive trace chemicals, and volcanic aerosols. *See* AT-MOSPHERE; TROPOSPHERE.

The heating that results from absorption of ultraviolet radiation by ozone causes temperatures generally to increase from the bottom of the stratosphere (tropopause) to the top (stratopause) near 50 km (30 mi), reaching 280 K ($45^{\circ}F$) over the summer pole. This temperature inversion limits vertical mixing, so that air typically spends months to years in the stratosphere. See TEMPERATURE INVERSION; TROPOPAUSE.

The lower stratosphere contains a layer of small liquid droplets. Typically less than 1 micrometer in diameter, they are made primarily of sulfuric acid and water. Occasional large volcanic eruptions maintain this aerosol layer by injecting sulfur dioxide into the stratosphere, which is converted to sulfuric acid and incorporated into droplets. Enhanced aerosol amounts from an eruption can last several years. By reflecting sunlight, the aerosol layer can alter the climate at the Earth's surface. By absorbing upwelling infrared radiation from the Earth's surface, the aerosol layer can warm the stratosphere. The aerosols also provide surfaces for a special set of chemical reactions that affect the ozone layer. Liquid droplets and frozen particles generally convert chlorinebearing compounds to forms that can destroy ozone. They also tend to take up nitric acid and water and to fall slowly, thereby removing nitrogen and water from the stratosphere. The eruption of Mount Pinatubo (Philippines) in June 1991 is believed to have disturbed the Earth system for several years, raising stratospheric temperatures by more than 1 K $(1.8^{\circ}F)$ and reducing global surface temperatures by about 0.5 K (0.9°F). See AEROSOL.

Ozone production is balanced by losses due to reactions with chemicals in the nitrogen, chlorine, hydrogen, and bromine families. Reaction rates are governed by temperature, which depends on amounts of radiatively important species such as carbon dioxide. Human activities are increasing the amounts of these molecules and are thereby affecting the ozone layer. Evidence for anthropogenic ozone loss has been found in the Antarctic lower stratosphere. Near polar stratospheric clouds, chlorine and bromine compounds are converted to species that, when the Sun comes up in the southern spring, are broken apart by ultraviolet

2124 Stratospheric ozone

radiation and rapidly destroy ozone. This sudden loss of ozone is known as the anthropogenic Antarctic ozone hole. See STRATO-SPHERIC OZONE. [M.H.H.]

Stratospheric ozone While ozone is found in trace quantities throughout the atmosphere, the largest concentrations are located in the lower stratosphere in a layer between 9 and 18 mi (15 and 30 km). Atmospheric ozone plays a critical role for the biosphere by absorbing the ultraviolet radiation with wavelength (λ) 240–320 nanometers. This radiation is lethal to simple unicellular organisms (algae, bacteria, protozoa) and to the surface cells of higher plants and animals. It also damages the genetic material of cells and is responsible for sunburn in human skin. The incidence of skin cancer has been statistically correlated with the observed surface intensity of ultraviolet wavelength 290–320 nm, which is not totally absorbed by the ozone layer. See OZONE; STRATOSPHERE; ULTRAVIOLET RADIATION (BIOLOGY).

Ozone also plays an important role in photochemical smog and in the purging of trace species from the lower atmosphere. Furthermore, it heats the upper atmosphere by absorbing solar ultraviolet and visible radiation ($\lambda < 710$ nm) and thermal infrared radiation ($\lambda \simeq 9.6$ micrometers). As a consequence, the temperature increases steadily from about -60° F (220 K) at the tropopause (5–10 mi or 8–16 km altitude) to about 45°F (280 K) at the stratopause (30 mi or 50 km altitude). This ozone heating provides the major energy source for driving the circulation of the upper stratosphere and mesosphere. See ATMOSPHERIC GENERAL CIRCULATION; TROPOPAUSE.

Above about 19 mi (30 km), molecular oxygen (O_2) is dissociated to free oxygen atoms (O) during the daytime by ultraviolet photons, ($h\nu$), as shown in reaction (1). The oxygen atoms produced then form ozone (O_3) by reaction (2), where M is an

$$O_2 + h\nu \to 0 + 0 \qquad \lambda < 242 \text{ nm} \tag{1}$$

$$0 + O_2 + M \rightarrow O_3 + M \tag{2}$$

arbitrary molecule required to conserve energy and momentum in the reaction. Ozone has a short lifetime during the day because of photodissociation, as shown in reaction (3). However,

$$0_3 + h\nu \to 0_2 + 0 \qquad \lambda < 710 \text{ nm} \tag{3}$$

except above 54 mi (90 km), where O_2 begins to become a minor component of the atmosphere, reaction (3) does not lead to a net destruction of ozone. Instead the O is almost exclusively converted back to O_3 by reaction (2). If the odd oxygen concentration is defined as the sum of the O_3 and O concentrations, then odd oxygen is produced by reaction (1). It can be seen that reactions (2) and (3) do not affect the odd oxygen concentrations but merely define the ratio of O to O_3 . Because the rate of reaction (2) decreases with altitude while that for reaction (3) increases, most of the odd oxygen below 36 mi (60 km) is in the form of O_3 while above 36 mi (60 km) it is in the form of O. The reaction that is responsible for a small fraction of the odd oxygen removal rate is shown as reaction (4). A significant fraction

$$0 + 0_3 \rightarrow 0_2 + 0_2$$
 (4)

of the removal is caused by the presence of chemical radicals [such as nitric oxide (NO), chlorine (Cl), bromine (Br), hydrogen (H), or hydroxyl (OH)], which serve to catalyze reaction (4) (see illustration).

The discovery in the mid-1980s of an ozone hole over Antarctica, which could not be explained by the classic theory of ozone and had not been predicted by earlier chemical models, led to many speculations concerning the causes of this event, which can be observed each year in September and October. As suggested by experimental and observational evidence, heterogeneous reactions on the surface of liquid or solid particles that



Principal chemical cycles in the stratosphere. The destruction of ozone is affected by the presence of radicals which are produced by photolysis or oxidation of source gases. Chemical reservoirs are relatively stable but are removed from the stratosphere by transport toward the troposphere and rainout.

produce Cl_2 , HOCl, and $ClNO_2$ gas, and the subsequent rapid photolysis of these molecules, produces chlorine radicals (Cl, ClO) which in turn lead to the destruction of ozone in the lower stratosphere by a catalytic cycle [reactions (5)–(7)]. Solar radia-

$$CI + O_3 \rightarrow CIO + O_2 \tag{5}$$

$$CIO + CIO \rightarrow CI_2O_2 \tag{6}$$

 $Cl_2O_2 + h\nu \to 2Cl + O_2 \tag{7}$

tion is needed for these processes to occur.

Sites on which the reactions producing Cl₂, HOCl, and ClNO₂ can occur are provided by the surface of ice crystals in polar stratospheric clouds (PSCs). These clouds are formed between 8 and 14 mi (12 and 22 km) when the temperature drops below approximately -123° F (187 K). Other types of particles are observed at temperatures above the frost point of -123° F (187 K). These particles provide additional surface area for these reactions to occur. Clouds are observed at high latitudes in winter. Because the winter temperatures are typically $20-30^{\circ}F(10-15 \text{ K})$ colder in the Antarctic than in the Arctic, their frequency of occurrence is highest in the Southern Hemisphere. Thus, the formation of the springtime ozone hole over Antarctica is explained by the activation of chlorine and the catalytic destruction of O_3 which takes place during September, when the polar regions are sunlit but the air is still cold and isolated from midlatitude air by a strong polar vortex. Satellite observations made since the 1970s suggest that total ozone in the Arctic has been abnormally low during the 1990s, probably in relation to the exceptionally cold winter tempratures in the Arctic lower stratosphere recorded during that decade. [G.P.B.; R.G.Pr.]

Strawberry Low-growing perennials, spreading by stolons, with fruit consisting of a fleshy receptacle, and "seeds" in pits or nearly superficial on the receptacle. The strawberry in the United States is derived from two species: *Fragaria chihensis*, which grows along the Pacific Coast of North and South America, and *F. uirginiana*, the eastern meadow strawberry, both members of the order Resales. *See* ROSALES.

The strawberry is the most universally grown of the small fruits, both in the home garden and in commercial plantings. Home garden production is possible in nearly all of the states, provided water can be supplied where rainfall is insufficient. Commercial production is important in probably three-fourths of the states. The following states are large producers: Oregon, California, Tennessee, Michigan, Louisiana, Washington, Arkansas, Kentucky, and New York. See FRUIT. [J.H.C.] **Stream function** In fluid mechanics, a mathematical idea which satisfies identically, and therefore eliminates completely, the equation of mass conservation. If the flow field consists of only two space coordinates, for example, *x* and *y*, a single and very useful stream function $\psi(x, y)$ will arise. If there are three space coordinates, such as (x, y, z), multiple stream functions are needed, and the idea becomes much less useful and is much less widely employed.

The stream function not only is mathematically useful but also has a vivid physical meaning. Lines of constant ψ are streamlines of the flow; that is, they are everywhere parallel to the local velocity vector. No flow can exist normal to a streamline; thus, selected ψ lines can be interpreted as solid boundaries of the flow.

Further, ψ is also quantitatively useful. In plane flow, for any two points in the flow field, the difference in their stream function values represents the volume flow between the points. See CREEPING FLOW; FLUID FLOW. [F.M.W.]

Stream gaging The measurement of water discharge in streams. Discharge is the rate of movement of the stream's water volume. It is the product of water velocity times cross-sectional area of the stream channel. Several techniques have been developed for measuring stream discharge; selection of the gaging method usually depends on the size of the stream. The most accurate methods for measuring stream discharge make use of in-stream structures through which the water can be routed, such as flumes and weirs.

A flume is a constructed channel that constricts the flow through a control section, the exact dimensions of which are known. Through careful hydraulic design and calibration by laboratory experiments, stream discharge through a flume can be determined by simply measuring the water depth (stage) in the inlet or constricted sections. Appropriate formulas relate stage to discharge for the type of flume used.

A weir is used in conjunction with a dam in the streambed. The weir itself is usually a steel plate attached to the dam that has a triangular, rectangular, or trapezoidal notch over which the water flows. Hydraulic design and experimentation has led to calibration curves and appropriate formulas for many different weir designs. To calculate stream discharge through a weir, only the water stage in the reservoir created by the dam needs to be measured. Stream discharge can be calculated by using the appropriate formula that relates stage to discharge for the type of weir used. *See* HYDROLOGY; SURFACE WATER. [T.C.Wi.]

Stream pollution Biological, or bacteriological, pollution in a stream indicated by the presence of the coliform group of organisms. While nonpathogenic itself, this group is a measure of the potential presence of contaminating organisms. Because of temperature, food supply, and predators, the environment provided by natural bodies of water is not favorable to the growth of pathogenic and coliform organisms. Physical factors, such as flocculation and sedimentation, also help remove bacteria. Any combination of these factors provides the basis for the biological self-purification capacity of natural water bodies.

Nonpolluted natural waters are usually saturated with dissolved oxygen. They may even be supersaturated because of the oxygen released by green water plants under the influence of sunlight. When an organic waste is discharged into a stream, the dissolved oxygen is utilized by the bacteria in their metabolic processes to oxidize the organic matter. The oxygen is replaced by reaeration through the water surface exposed to the atmosphere. This replenishment permits the bacteria to continue the oxidative process in an aerobic environment. In this state, reasonably clean appearance, freedom from odors, and normal animal and plant life are maintained.

An increase in the concentration of organic matter stimulates the growth of bacteria and increases the rates of oxidation and oxygen utilization. If the concentration of the organic pollutant is so great that the bacteria use oxygen more rapidly than it can be replaced, only anaerobic bacteria can survive and the stabilization of organic matter is accomplished in the absence of oxygen. Under these conditions, the water becomes unsightly and malodorous, and the normal flora and fauna are destroyed. Furthermore, anaerobic decomposition proceeds at a slower rate than aerobic. For maintenance of satisfactory conditions, minimal dissolved oxygen concentrations in receiving streams are of primary importance.

Municipal sewage and industrial wastes affect the oxygen content of a stream. Cooling water, used in some industrial processes, is characterized by high temperatures, which reduce the capacity of water to hold oxygen in solution. Municipal sewage requires oxygen for its stabilization by bacteria. Oxygen is utilized more rapidly than it is replaced by reaeration, resulting in the death of the normal aquatic life. Further downstream, as the oxygen demands are satisfied, reaeration replenishes the oxygen supply.

Polluted waters are deprived of oxygen by the exertion of the biochemical oxygen demand, which is defined as the quantity of oxygen required by the bacteria to oxidize the organic matter. Factors such as the turbulence of the stream flow, biological growths on the stream bed, insufficient nutrients, and inadequate bacteria in the river water influence the rate of oxidation in the stream as well as the removal of organic matter.

When a significant portion of the waste is in the suspended state, settling of the solids in a slow-moving stream is probable. The organic fraction of the sludge deposits decomposes anaerobically, except for the thin surface layer which is subjected to aerobic decomposition due to the dissolved oxygen in the overlying waters. In warm weather, when the anaerobic decomposition proceeds at a more rapid rate, gaseous end products, usually carbon dioxide and methane, rise through the supernatant waters. The evolution of the gas bubbles may raise sludge particles to the water surface. Although this phenomenon may occur while the water contains some dissolved oxygen, the more intense action during the summer usually results in depletion of dissolved oxygen.

Water may absorb oxygen from the atmosphere when the oxygen in solution falls below saturation. Dissolved oxygen for receiving waters is also derived from two other sources: that in the receiving water and the waste flow at the point of discharge, and that given off by green plants.

Unpolluted water maintains in solution the maximum quantity of dissolved oxygen. The saturation value is a function of temperature and the concentration of dissolved substances, such as chlorides. When oxygen is removed from solution, the deficiency is made up by the atmospheric oxygen, which is absorbed at the water surface and passes into solution. The oxygen balance in a stream is determined by the concentration of organic matter and its rate of oxidation, and by the dissolved oxygen concentration and the rate of reaeration. *See* ESTUARINE OCEANOGRAPHY; FRESH-WATER ECOSYSTEM; SEWAGE DISPOSAL; WATER POLLUTION. [D.J.O.]

Stream transport and deposition The sediment debris load of streams is a natural corollary to the degradation of the landscape by weathering and erosion. Eroded material reaches stream channels through rills and minor tributaries, being carried by the transporting power of running water and by mass movement, that is, by slippage, slides, or creep. The size represented may vary from clay to boulders. At any place in the stream system the material furnished from places upstream either is carried away or, if there is insufficient transporting ability, is accumulated as a depositional feature. The accumulation of deposited debris tends toward increased ease of movement, and this tends eventually to bring into balance the transporting ability of the stream and the debris load to be transported. [LB.L]

Streaming potential The potential which is produced when a liquid is forced to flow through a capillary or a porous

2126 Streamlining

solid. The streaming potential is one of four related electrokinetic phenomena which depend upon the presence of an electrical double layer at a solid-liquid interface. This electrical double layer is made up of ions of one charge type which are fixed to the surface of the solid and an equal number of mobile ions of the opposite charge which are distributed through the neighboring region of the liquid phase. In such a system the movement of liquid over the surface of the solid produces an electric current, because the flow of liquid causes a displacement of the mobile counterions with respect to the fixed charges on the solid surface. The applied potential necessary to reduce the net flow of electricity to zero is the streaming potential. [Q.V.W.]

Streamlining The contouring of a body to reduce its resistance (drag) to motion through a fluid.

For fluids with relatively low viscosity such as water and air, effects of viscous friction are confined to a thin layer of fluid on the surface termed the boundary layer. Under the influence of an increasing pressure, the flow within the boundary layer tends to reverse and flow in an upstream direction. Viscosity tends to cause the flow to separate from the body surface with consequent formation of a region of swirling or eddy flow (termed the body wake; illus. *a*). This eddy formation leads to a reduction in the downstream pressure on the body and hence gives rise to a force opposite to the body motion, known as pressure drag. *See* WAKE FLOW.

In general, streamlining in subsonic flow involves the contouring of the body in such a manner that the wake is reduced and hence the pressure drag is reduced. The contouring should provide for gradual deceleration to avoid flow separation, that is, reduced adverse pressure gradients. These considerations lead to the following general rules for subsonic streamlining: The forward portion of the body should be well rounded, and the body should curve back gradually from the forward section to a tapering aftersection with the avoidance of sharp corners along the body surface. These conditions are well illustrated by teardrop shapes (illus. *b*).

At supersonic speeds the airflow can accommodate sudden changes in direction by being compressed or expanded. Where this change in direction occurs at the nose of the body, a compression wave is created, the strength of which depends upon the magnitude of the change in flow direction. Lowering the bodyinduced flow angle weakens this compression shock wave. When the flow changes direction again at the midpoint of the body, the air will expand to follow the shape of the body. This change in



Flow about bodies in uniform subsonic flow. (a) Blunt body. (b) Streamlined body.

direction creates expansion waves. At the tail of the body the direction changes again, creating another compression or shock wave. At each of these shock waves, changes in pressure, density, and velocity occur, and in this process energy is lost. This loss results in a retarding force known as wave drag. *See* SHOCK WAVE.

Bodies which are streamlined for supersonic speeds are characterized by a sharp nose and small flow deflection angles. Because the intensity of the shock wave and the drag level is dependent upon the magnitude of the change in flow direction, the width or thickness of the body should be minimal. *See* BOUNDARY-LAYER FLOW. [A.G.H.; D.M.Bu.]

Strength of materials A branch of applied mechanics concerned with the behavior of materials under load, relationships between externally applied loads and internal resisting forces, and associated deformations. Knowledge of the properties of materials and analysis of the forces involved are fundamental to the investigation and design of structures and machine elements. *See* MACHINE DESIGN; STRUCTURAL MATERIALS.

Investigation of the resistance of a member, dealing with internal forces, is called free-body analysis. Determination of the distribution and intensity of the internal forces and the associated deformations is called stress analysis. See STRESS AND STRAIN.

A material offers resistance to external load only insofar as the component elements can furnish cohesive strength, resistance to compaction, and resistance to sliding. The relations developed in strength of materials analysis evaluate the tensile, compressive, and shear stresses that a material is called upon to resist. The most important factors in determining the suitability of a structural or machine element for a particular application are strength and stiffness. *See* SHEAR. [J.B.S.]

Strepsiptera An order of twisted-wing insects that spend most of their life cycle as internal parasitoids of other insects. The adult male is only a few millimeters in wing span, is free living, and has only one pair of full flight wings (the posterior, or metathoracic, pair). The front (mesothoracic) wings are reduced to narrow, clublike organs that may function as halteres, or flight balancers. The male eyes are coarsely faceted and berrylike, and the antennae have four to seven segments, with some segments having finger- or bladelike extensions. Most adult females are immobile, blind, and larviform, and live inside the insect host. Rarely, the female is free living, with legs and eyes but no wings. More than 600 species of Strepsiptera are known, many of them not yet formally described and named. The order's relationship to other insect orders remains uncertain.

Both male and female begin larval life as mobile first-stage organisms with eyes and three pairs of functional legs. These triunguloids eventually attack the immature or adult forms of the host and enter their bodies, where they molt to an apodous instar. Two subsequent molts result in the divergence of the sexes and the differentiation of a hardened forebody that eventually protrudes through the host integument; the molted larval integuments are not shed but are incorporated in the general puparial wall arid the neotenous adult female capsule. Males (and some adult females) emerge from the puparium and begin the search for a mate, probably through the mediation of airborne pheromones.

Strepsiptera are found worldwide in temperate and tropical habitats. They parasitize insects, mostly of orders Hemiptera and Hymenoptera, but also some cockroaches, mantids, orthopterans, flies, and silverfish. Their effect on the host is variable, ranging from reproductive failure to death. *See* ENDOPTERYGOTA; PHEROMONE. [W.L.Bro.]

Streptococcus A large genus of spherical or ovoid bacteria that are characteristically arranged in pairs or in chains resembling strings of beads. Many of the streptococci that constitute part of the normal flora of the mouth, throat, intestine,

and skin are harmless commensal forms; other streptococci are highly pathogenic. The cells are gram-positive and can grow either anaerobically or aerobically, although they cannot utilize oxygen for metabolic reactions. Glucose and other carbohydrates serve as sources of carbon and energy for growth. All members of the genus lack the enzyme catalase. Streptococci can be isolated from humans and other animals.

Streptococcus pyogenes is well known for its participation in many serious infections. It is a common cause of throat infection, which may be followed by more serious complications such as rheumatic fever, glomerulonephritis, and scarlet fever. Other beta-hemolytic streptococci participate in similar types of infection, but they are usually not associated with rheumatic fever and glomerulonephritis. Group B streptococci, which are usually beta-hemolytic, cause serious infections in newborns (such as meningitis) as well as in adults. Among the alpha-hemolytic and nonhemolytic streptococci, S. pneumoniae is an important cause of pneumonia and other respiratory infections. Vaccines that protect against infection by the most prevalent capsular serotypes are available. The viridans streptococci comprise a number of species commonly isolated from the mouth and throat. Although normally of low virulence, these streptococci are capable of causing serious infections (endocarditis, abcesses). See PNEUMONIA. [K.Ru.]

Streptothricosis An acute or chronic infection of the epidermis, caused by the bacterium *Dermatophilus congolensis*. which results in an oozing dermatitis with scab formation. Streptothricosis includes dermatophilosis, mycotic dermatitis, lumpy wool, strawberry foot-rot, and cutaneous streptothricosis—diseases having a worldwide distribution and affecting a wide variety of species, including humans.

The infectious form of the bacterium is a coccoid, motile zoospore that is released when the skin becomes wet. Thus, the disease is closely associated with rainy seasons and wet summers. Zoospores lodge on the skin of susceptible animals and germinate by producing filaments which penetrate to the living epidermis, where the organism proliferates by branching mycelial growth.

Early cutaneous lesion of dermatophilosis in cattle reveals small vesicles, papules, and pus formation under hair plaques. An oozing dermatitis then appears as the disease progresses and exudates coalesce to form scabs, which change to hard crusts firmly adherent to the skin. The crusts enlarge and harden, and are often devoid of hair. Lesions occur on most areas of sheep, but the characteristic lesions in the wooled areas occur as numerous hard masses of crust or scab scattered irregularly over the back, flanks, and upper surface of the neck. Lesion resolution has been found to correlate with the presence of immunoglobulin A–containing plasma cells in the dermis and with the antibody levels to *D. congolensis* at the skin surface of infected sheep and cattle.

No single treatment is considered specific for dermatophilosis. Some observers claim that topical agents are successful for sheep and cattle. A number of systemic antibiotics are effective in treating the disease. A combination of streptomycin and penicillin has given good therapeutic results in both bovine and equine infections. [S.S.S.]

Stress (psychology) Generally, environmental events of a challenging sort as well as the body's response to such events. Of particular interest has been the relationship between stress and the body's adaptation to it on the one hand and the body's susceptibility to disease on the other. Both outcomes involve behavioral and brain changes as well as psychosomatic events, that is, changes in body function arising from the ability of the brain to control such function through neural output as well as hormones. One problem is that both environmental events and bodily responses have been referred to interchangeably as stress. It is preferable to refer to the former as the stressor

and the latter as the stress response. The stress response consists of a cascade of neural and hormonal events that have shortand long-lasting consequences for brain and body alike. A more serious issue is how an event is determined to be a stressor. One view is to define a stressor as an environmental event causing a negative outcome, such as a disease. Another approach is to view stressors as virtually any challenge to homeostasis and to regard disease processes as a failure of the normal operation of adapative mechanisms, which are part of the stress response. With either view, it is necessary to include psychological stressors, such as fear, that contain implied threats to homeostasis and that evoke psychosomatic reactions. These are reactions that involve changes in neural and hormonal output caused by psychological stress. Psychosomatic reactions may lead to adaptive responses, or they may exacerbate disease processes. Whether the emphasis is on adaptation or disease, it is essential to understand the processes in the brain that are activated by stressors and that influence functions in the body. See HOMEOSTASIS; PSYCHOSOMATIC DISORDERS.

Among the many neurotransmitter systems activated by stress is noradrenaline, produced by neurons with cell bodies in the brainstem that have vast projections up to the forebrain and down the spinal cord. Stressful experiences activate the noradrenergic system and promote release of noradrenaline; severe stress leads to depletion of noradrenaline in brain areas such as the hypothalamus. This release and depletion of noradrenaline stores results in changes at two levels of neuronal function: phosphorylation is triggered by the second-messenger cyclic AMP and occurs in the presynaptic and postsynaptic sites where noradrenaline is released and where it also acts; synthesis of new protein is induced via actions on the genome. Both processes enhance the ability of the brain to form noradrenaline when the organism is once again confronted with a stressful situation. Other neurotransmitter systems may also show similar adaptive changes in response to stressors. See NORADRENERGIC SYSTEM.

Stress also activates the neurally mediated discharge of adrenaline from the adrenal medulla and of hypothalamic hormones that initiate the neuroendocrine cascade, culminating in glucocorticoid release from the adrenal cortex. Thus, the activity of neurons triggered by stressful experiences, physical trauma, fear, or anger leads to hormone secretion that has effects throughout the body. Virtually every organ of the body is affected by stress hormones. The hypothalamic hormone (corticotrophinreleasing hormone) that triggers the neuroendocrine cascade directly stimulates the pituitary to secrete ACTH. In response to certain stressors, the hypothalamus also secretes vasopressin and oxytocin, which act synergistically with corticotrophin-releasing hormone on the pituitary to potentiate the secretion of ACTH. Various stressors differ in their ability to promote output of vasopressin and oxytocin, but all stressors stimulate release of corticotrophin-releasing hormone. Other hormones involved in the stress response include prolactin and thyroid hormone; the metabolic hormones insulin, epinephrine, and glucagon; and the endogenous opiates endorphin and enkephalin. See ENDOR-PHINS

Of all the hormones in the endocrine cascade initiated by stress, the glucocorticoids are the most important because of their widespread effects throughout the body and in the brain. The brain contains target cells for adrenal glucocorticoids secreted in stress, and receptors in these cells are proteins that interact with the genome to affect expression of genetic information. Thus, the impact of stress-induced activation of the endocrine cascade that culminates in glucocorticoid release is the feedback of glucocorticoids on target brain cells. The effect is to alter the structure and function of the brain cells over a period of time ranging from hours to days.

In the case of noradrenaline, glucocorticoids have several types of feedback effects that modify how the noradrenergic system responds to stress. Glucocorticoids inhibit noradrenaline release, and they reduce the second-messenger response of brain
2128 Stress and strain

structures such as the cerebral cortex to noradrenaline. Glucocorticoid feedback also affects the serotonin system, facilitating serotonin formation during stress but at the same time altering the levels of several types of serotonin receptors in different brain regions, which has the net effect of shifting the balance within the serotonergic system. Taken together, evidence points to a role of glucocorticoid secretion in leading to restoration of homeostatic balance by counteracting the acute neural events such as increased activity of noradrenaline and serotonin, which are turned on by stressful experiences. Other neurotransmitter systems may also respond to glucocorticoid action. Moreover, the other hormones activated by stress have effects on the brain and body that must be considered. See SEROTONIN.

In general, stress hormones are protective and adaptive in the immediate aftermath of stress, and the organism is more vulnerable to many conditions without them. However, the same hormones can promote damage and accelerate pathophysiological changes, such as bone mineral loss, obesity, and cognitive impairment, when they are overproduced or not turned off. This wear-and-tear on the body has been called allostatic load. It is based upon the notion that allostasis is the active process of maintaining stability, or homeostasis, through change, and allostatic load is the almost inevitable cost to the body of doing so.

Stress hormone actions have important effects outside the brain on such systems as the immune response. Glucocorticoids and catecholamines from sympathetic nerves and the adrenal medulla participate in the mobilization and enhancement of immune function in the aftermath of acute stress. These effects improve the body's defense against pathogens but can exacerbate autoimmune reactions. When they are secreted chronically, the stress-related hormones are generally immunosuppressive; such effects can be beneficial in the case of an autoimmune disease but may compromise defense against a virus or bacterial infections. At the same time, glucocorticoids are important agents for containing the acute-phase response to an infection or autoimmune disturbance. In the absence of such containment, the organism may die because of the excessive inflammatory response. See IMMUNOLOGY.

Besides affecting the immune response, stressors are believed to exacerbate endogenous depressive illness in susceptible individuals. Major depressive illness frequently results in elevated levels of cortisol in the blood. It is not clear whether the elevated cortisol is a cause or strictly a result of the brain changes involved in depressive illness. *See* AFFECTIVE DISORDERS. [B.McE.]

Stress and strain Related terms defining the intensity of internal reactive forces in a deformed body and associated unit changes of dimension, shape, or volume caused by externally applied forces. Stress is a measure of the internal reaction between elementary particles of a material in resisting separation, compaction, or sliding that tend to be induced by external forces. Total internal resisting forces are resultants of continuously distributed normal and parallel forces that are of varying magnitude and direction and are acting on elementary areas throughout the material. These forces may be distributed uniformly or nonuniformly. Stresses are identified as tensile, compressive, or shearing, according to the straining action.

Strain is a measure of deformation such as (1) linear strain, the change of length per unit of linear dimensions; (2) shear strain, the angular rotation in radians of an element undergoing change of shape by shearing forces; or (3) volumetric strain, the change of volume per unit of volume. The strains associated with stress are characteristic of the material. Strains completely recoverable on removal of stress are called elastic strains. Above a critical stress, both elastic and plastic strains exist, and that part remaining after unloading represents plastic deformation called inelastic strain. Inelastic strain reflects internal changes in the crystalline structure of the metal. Increase of resistance to continued plastic



Stress-strain diagram for a low-carbon steel. ΔS = change in stress; $\Delta \epsilon$ = change in strain; P = force; A_0 = area of cross section.

deformation due to more favorable rearrangement of the atomic structure is strain hardening.

A stress-strain diagram is a graphical representation of simultaneous values of stress and strain observed in tests and indicates material properties associated with both elastic and inelastic behavior (see illustration). It indicates significant values of stressaccompanying changes produced in the internal structure. See ELASTICITY; STRENGTH OF MATERIALS. [J.B.S.]

Strigiformes The owls, an order of nocturnal, worldwide predacious birds that are probably most closely related to the goatsuckers. The strigiforms are arranged in four families: Ogygoptyngidae (fossil), Protostrigidae (fossil), Tytonidae (barn owls; 11 species), and Strigidae (typical owls; 135 species). The bay owls (*Phodilus*), occurring in southern Asia and Africa, are somewhat intermediate between the Tytonidae and the Strigidae and are sometimes placed in a separate family; here they are treated as a subfamily, Phodilinae, of the Tytonidae.

Owls are small to medium in size, with soft plumage of somber colors. The large head has a facial disk that covers the feathered parabolic reflectors of the bird's acute directional hearing system. The eyes are large and capable of sight in very dim light, and the bill is strong and hooked. The wings are long and rounded, and the flight feathers are fringed for silent flight. The ulna has a unique bony arch on the shaft. Their strong legs are short to medium in length and terminate in strong feet. Owls are excellent fliers but walk poorly. Of the four toes on each foot, two point forward and two backward and bear strong claws. Prey is detected by acute night vision or by directional hearing; owls can locate and catch their prey in total darkness. Owls are generally nonmigratory and solitary, but some species live in small flocks. Courtship takes place at night with a male hooting to a female, which answers. A strong pair bond exists between the monogamous male and female, which usually build their nest in a tree cavity. Some species nest on the ground, on cliff ledges, or in abandoned crow or hawk nests. The clutch of up to seven eggs is incubated by both sexes, and after hatching, the young stay in the nest and are cared for by both parents. See AVES. [WJB]

Stripping The removal of volatile component from a liquid by vaporization. The stripping operation is an important step in many industrial processes which employ absorption to purify gases and to recover valuable components from the vapor phase. In such processes, the rich solution from the absorption step must be stripped in order to permit recovery of the absorbed solute and recycle of the solvent. *See* GAS ABSORPTION OPERATIONS.

Stripping may be accomplished by pressure reduction, the application of heat, or the use of an inert gas (stripping vapor). Many processes employ a combination of all three; that is, after absorption at elevated pressure, the solvent is flashed to In addition to its use in conjunction with gas absorption, the term stripping is also used quite generally in technical fields to denote the removal of one or more components from a mixed system. Such usage covers (1) the distillation operation which takes place in a distilling column in the zone below the feed point, (2) the extraction of one or more components from a liquid by contact with a solvent liquid, (3) the removal of organic or metal coatings from solid surfaces, and (4) the removal of color from dyed fabrics. *See* DISTILLATION; ELECTROPLATING OF METALS; SOLVENT EXTRACTION. [A.L.K.]

Stroboscope An instrument for observing moving bodies by making them visible intermittently and thereby giving them the optical illusion of being stationary. A stroboscope may operate by illuminating the object with brilliant flashes of light or by imposing an intermittent shutter between the viewer and the object.

Stroboscopes are used to measure the speed of rotation or frequency of vibration of a mechanical part or system. They have the advantage over other instruments of not loading or disturbing the equipment under test. Mechanical equipment may be observed under actual operating conditions with the aid of stroboscopes. Parasitic oscillations, flaws, and unwanted distortion at high speeds are readily detected.

The flashing-light stroboscopes employ gas discharge tubes to provide a brilliant light source of very short duration. Tubes may vary from neon glow lamps, when very little light output is required, to special stroboscope tubes capable of producing flashes of several hundred thousand candlepower with a duration of only a few millionths of a second. See NEON GLOW LAMP; VAPOR LAMP. [A.R.E.]

Stroboscopic photography Stroboscopic or "strobe" photography generally refers to pictures of both single and multiple exposure taken by flashes of light from electrical discharges. Originally the term referred to multiple-exposed photographs made with a Stroboscopic disk as a shutter. One essential feature of modern Stroboscopic photography is a short exposure time, usually much shorter than can be obtained by a mechanical shutter.

High-speed photography with Stroboscopic light has proved to be one of the most powerful research tools for observing fast motions in engineering and in science. Likewise, the electrical system of producing flashes of light in xenon-filled flash lamps is of great utility for studio, candid, and press photography. *See* PHOTOGRAPHY; STROBOSCOPE. [H.E.E.]

Stromatolite A laminated, microbial structure in carbonate rocks (limestone and dolomite). Stromatolites are the oldest macroscopic evidence of life on Earth, at least 2.5 billion years old, and they are still forming in the seas. During the 1.5 billion years of Earth history before marine invertebrates appeared, stromatolites were the most obvious evidence of life, and they occur sporadically throughout the remainder of the geologic record. In Missouri and Africa, stromatolite reefs have major accumulations of lead, zinc, or copper; and in Montana, New Mexico, and Oman, stromatolites occur within oil and gas reservoirs. For geologists, the shapes of stromatolites are useful indications of their environmental conditions, and variations in form and microstructure of the laminations may be age-diagnostic in those most ancient sedimentary rocks that lack invertebrate fossils. See DOLOMITE; LIMESTONE; REEF. Stromatolites are readily recognizable in outcrops by their characteristic convex-upward laminated structure. Individual, crescent-shaped laminations, which are generally about a millimeter thick, are grouped together to produce an enormous range of shapes and sizes.

The tiny, filamentous cyanobacteria (blue-green algae) that make present-day stromatolites, and similar filaments associated with the oldest stromatolites known, are considered one of the most successful organisms on Earth. Living stromatolites in the Bahamas and Western Australia possess laminations that record the episodic trapping and binding of sediment particles by the microbial mat. In the modern oceans, stromatolites develop almost exclusively in extreme marine conditions that exclude or deter browsing invertebrates and fish from destroying the microbial mats and inhibit colonization by competing algae. *See* ALGAE; CYANOBACTERIA. [R.Gi.]

Stromatoporoidea An extinct order thought to be sponges because of its close similarity to the class Sclerospongiae. Generally accepted stromatoporoids first appeared at the beginning of the Middle Ordovician and died out at the end of the Devonian. The Stromatoporoidea are worldwide in distribution, most commonly associated with bedded limestones and fossil reefs.

The stromatoporoid skeleton is called a coenosteum, which begins development as a thin, encrusting layer of irregular rods and plates called the peritheca. Most coenostea are hemispherical, or thick laminar or lens-shaped structures. However, some stromatoporoids are thin and irregularly undulatory and others are cylindrical, cone-shaped, branching, or anastomosing. *See* SCLEROSPONGIAE. [J.St.J.]

Strong nuclear interactions One of the fundamental physical interactions, which acts between a pair of hadrons. Hadrons include the nucleons, that is, neutrons and protons; the strange baryons, such as lambda (Λ) and sigma (Σ); the mesons, such as pion (π) and rho (ρ); and the strange meson, kaon (K). The nature of the interaction is determined principally through observations of the collision of a hadron pair. From this it is found that the interaction has a short range of about 10^{-15} m (10^{-13} in.) and is by far the dominant force within this range, being much larger than the electromagnetic interaction, which is next in magnitude. The strong interaction conserves parity and is time-reversal-invariant. See BARYON; HADRON; MESON; NUCLEON; PARITY (QUANTUM MECHANICS); STRANGE PARTICLES; SYMMETRY LAWS (PHYSICS); TIME REVERSAL INVARIANCE.

The interaction between baryons for distances greater than 10^{-15} m arises from the exchange of mesons. At relatively large distances, single-pion exchange dominates (illus. *a*). At shorter separation distances, the two-pion systems such as the ρ become important (illus. *b*). The interaction between the strange baryons, and between the strange baryons and the nucleons, is moderated by kaon exchange. To summarize this description, the interaction between baryons in the SU(3) multiplet is the consequence of the exchange of SU(3) spin 0 and spin 1 bosons. In a second



Interaction between nucleons (a) from exchange of single pion, (b) from exchange of ρ -meson, a two-pion system, (c) from exchange of two separate pions with formation of excited state of nucleon, and (d) without formation of excited state.

2130 Strongylida

approximation, the exchange of two pions (illus. *c*, *d*), or more generally, the exchange of two members of the SU(3) spin 0 and spin 1 multiplet, is responsible for a component of the strong nuclear interaction. See UNITARY SYMMETRY.

The range of the interaction generated by these exchanges can be calculated by using the formula below, where m is the mass

Range
$$=\frac{\hbar}{m}$$

of the exchanged particles, \hbar is Planck's constant divided by 2π , and c is the speed of light. According to the above equation, the range of the interaction developed when a single pion is exchanged (illus. 1a) is equal to 1.4×10^{-15} m (5.5 $\times10^{-14}$ in.), while that due to two-pion exchange (illus. 1d) is 0.7 $\times10^{-15}$ m (2.8 $\times10^{-14}$ in.).

At short separation distances the quark-gluon structure of the baryons must be taken into account. The interaction must be considered as a property of the six-quark-plus-gluon system. The decisive elements are the Pauli principle obeyed by the quarks, and the mismatch between the six-quark wave function and the two-baryon wave function. Thus, at short distances the interaction is effectively repulsive or more generally independent of the kinetic energy of the baryons at infinite separations. *See* ELEMEN-TARY PARTICLE; EXCLUSION PRINCIPLE; FUNDAMENTAL INTERACTIONS; GLUONS; QUARKS.

Strongylida An order of nematodes in which the cephalic region may be adorned with three or six labia or the labia may be replaced by a corona radiata. All strongylid nematodes are parasitic. The order embraces eight superfamilies: Strongyloidea, Diaphanocephaloidea, Ancylostomatoidea, Trichostrongyloidea, Metastrongyloidea, Cosmocercoidea, Oxyuroidea, and Heterakoidea.

Strongyloidea. The Strongyloidea contain important parasites of reptiles, birds, and mammals. The early larval stages may be free-living microbivores, but the adults are always parasitic. Three species are important parasites of horses, *Strongylus vulgaris, S. equinus*, and *S. edentatus*. All three undergo direct life cycles; that is, infestations are acquired by ingestion of contaminated food.

Trichostrongyloidea. The Trichostrongyloidea comprise obligate parasites of all vertebrates but fishes. Normally they are intestinal parasites, but some are found in the lungs. The species are important parasites of sheep, cattle, and goats. The adult females lay eggs in the intestinal tract, which are passed out with the feces. In the presence of oxygen the eggs hatch in a few days. When the larvae are ingested by an appropriate host, their protective sheath is lost, and they proceed through the fourth larval stage to adulthood in the intestinal tract, where they may enter the mucosa. No migration takes place outside the gastrointestinal tract.

Metastrongyloidea. The Metastrongyloidea comprise obligate parasites of terrestrial and marine mammals, found commonly in the respiratory tract. In their life cycle they utilize both paratenic and intermediate hosts, among them a variety of invertebrates, including earthworms and mollusks. Two important species are *Metastrongylus apri* (swine lungworm) and *Angiostrongylus cantonensis* (rodent lungworm).

Heterakoidea. The Heterakoidea are capable of parasitizing almost any warm-blooded vertebrate as well as reptiles and amphibians. The species *Ascaridia galli* is the largest known nematode parasite of poultry; males are 2–3 in. (50–76 mm) long, and females 3–4.5 in. (75–116 mm). [A.R.M.]

Oxyuroidea. The Oxyuroidea constitute a large group of the phylum Nemata. Hosts include terrestrial mammals, birds, reptiles, amphibians, fishes, insects, and other arthropods.

The species are small to medium sized and thin bodied. With one exception, known life cycles are direct. Typically the eggs pass out of the host's alimentary tract onto the ground, where they become fully embryonated and infective. Normally the infective egg does not hatch until a susceptible animal ingests it. The cecum and colon of the host are the typical locations of these parasites. Larvae in all stages of development and adults occur in the gut.

The human pinworm, *Enterobius vermicularis*, is probably the most contagious of all helminthic diseases. It is estimated that 10% of the world's population suffer from this parasite, the majority being children. Indeed, incidence among schoolchildren in the cool regions of the world often approaches 100%. Infection occurs when eggs are inhaled or ingested. The most common method of transmission is from anus to mouth. Because of the aerial transmission, this disease is highly contagious. Though the infection is seldom serious, the behavioral symptoms are disturbing: nail biting, teeth grinding, anal scratching, insomnia, nightmares, and even convulsions. Several medical treatments are available, but there is often the danger of reinfestation from contaminated objects within the household or institution. *See* NEMATA.

Strontianite The mineral form of strontium carbonate, usually with some calcium replacing strontium. It characteristically occurs in veins with barite or celestite or as masses in certain sedimentary rocks. Strontianite is normally prismatic, but it may also be massive. It may be colorless or gray with yellow, green, or brownish tints. The hardness is $3^{1/2}$ on Mohs scale, and the specific gravity of 3.76. It occurs at Strontian, Scotland, and in Germany, Austria, Mexico, and India and, in the United States, in the Strontium Hills of Calfornia. See CARBONATE MINERALS; STRONTIUM. [R.I.Ha.]

Strontium A chemical element, Sr, atomic number 38, and atomic weight 87.62. Strontium is the least abundant of the alkaline-earth metals. The crust of the Earth is 0.042% strontium, making this element as abundant as chlorine and sulfur. The main ores are celestite, SrSO₄, and strontianite, SrCO₃. See ALKALINE-EARTH METALS; PERIODIC TABLE; STRONTIANITE.

Strontium nitrate is used in pyrotechnics, railroad flares, and tracer bullet formulations. Strontium hydroxide forms soaps and greases with a number of organic acids which are structurally stable, resistant to oxidation and breakdown over a wide temperature range.

Properties of strontium				
Property	Value			
Atomic number Atomic weight Isotopes (stable) Boiling point, °C Melting point, °C Density, g/cm3 at 20°C	38 87.62 84, 86, 87, 88, 90 1638(?) 704(?) 2.6			

Strontium is divalent in all its compounds which are, aside from the hydroxide, fluoride, and sulfate, quite soluble. Strontium is a weaker complex former than calcium, giving a few weak oxy complexes with tartrates, citrates, and so on. Some physical properties of the element are given in the table. [R.F.]

Strophanthus A genus of woody climbers of the dogbane family (Apocynaceae), natives of tropical Asia and Africa. They are the source of arrow poisons. The dried, greenish, ripe seeds of *Strophanthus hispidus* and *S. kombe* contain the glucoside strophanthin, which is much used in treating heart ailments. Strophanthin acts directly on heart muscle, increasing muscular force. It causes the heart to beat more regularly and decreases the pulse rate. Strophanthin is a precursor of cortisone, which is used in the treatment of arthritis. *See* GENTIANALES. [PD.St.; E.L.C.]

Strophomenida An extinct order of brachiopods, in the subphylum Rhynchonelliformea, that inhabited shallow shelf seas of the Paleozoic Era. It contains the greatest taxonomic diversity of all rhynchonelliform brachiopod orders, and its members exhibit a diverse range of morphologies and modes of life.

Strophomenida contains four suborders: Strophomenidina, Chonetidina, Productidina, and Strophalosiidina. A fifth suborder, Orthotetidina, shares characteristics with both strophomenids and primitive orthids. Strophomenids are distinguished from all other articulate brachiopods by their concavo-convex or plano-convex lateral profiles and the lack of a functional pedicle in adults. They also possess strophic (straight) hinge lines and variable ornamentation from fine-ribbed early Paleozoic forms to more elaborate spine-bearing late Paleozoic forms. Adult strophomenids were mostly free-living, sessile, epifaunal suspension feeders. The concavo-convex lateral profile and elaborate spines were adaptations for living on soft substrates. *See* BRACHIOPODA; RHYNCHONELLIFORMEA. [M.E.P.]

Structural analysis A detailed evaluation intended to assure that, for any structure, the deformations will be sufficiently below allowable values that structural failure will not occur. The deformations may be elastic (fully recoverable) or inelastic (permanent). They may be small, with an associated structural failure that is cosmetic; for example, the deflection of a beam supporting a ceiling may cause cracking of the plaster. They may be large, with an associated structural failure that is catastrophic; for example, the buckling of a column or the fracture of a tension member causes complete collapse of the structure.

Structural analysis may be performed by tests on the actual structure, on a physical model of the structure to some scale, or through the use of a mathematical model. Tests on an actual structure are performed in those cases where many similar structures will be produced, for example, automobile frames, or where the cost of a test is justified by the importance and difficulty of the project, for example, a lunar lander. Physical models are sometimes used where subassemblages of major structures are to be investigated. The vast majority of analyses, however, are on mathematical models, particularly in the field of structural engineering which is concerned with large structures such as bridges, buildings, and dams. See BRIDGE; BUILDINGS; DAM; STRUCTURE (ENGINEERING).

The advent of the digital computer made it possible to create mathematical models of great sophistication, and almost all complex structures are now so analyzed. Programs of such generality have been written as to permit the analysis of any structure. These programs permit the model of the structure to be two- or three-dimensional, elastic or inelastic, and determine the response to forces that are static or dynamic. Most of the programs utilize the stiffness method, in which the stiffnesses of the individual elements are assembled into a stiffness matrix for the entire structure, and analysis is performed in which all behavior is assumed to be linearly elastic. *See* DIGITAL COMPUTER; ELASTICITY.

The structural engineer's function continues to require training and experience in conceptualizing the structure, choosing the appropriate model, estimating the loads that will be of importance, coding the information for the program, and interpreting the results. The analyst usually enters the process after the conceptualization. Most structures consist of assemblies of members connected at joints. While all real members transmit axial, torsional, and bending actions, the majority of buildings and bridges are analyzed as trusses, beams, and frames with either axial or bending forces predominant. *See* BEAM; ENGINEERING DESIGN; STRESS AND STRAIN; STRUCTURAL DESIGN; TRUSS.

Whether the model selected is detailed or simplified, one extremely important part of the analysis consists of the estimate of the loads to be resisted. For bridges and buildings, the primary vertical loads are gravity loads. These include the weight of the structure itself, and such appurtenances as will be permanent in nature. These are referred to as dead loads. The loads to be carried, the live loads, may consist of concentrated loads (heavy objects occupying little space, for example, a printing press), or loads distributed over relatively large areas (such as floor and deck coverings). Horizontal loads on buildings are produced by wind and by the inertia forces created during earthquakes. In seismic analysis, computers are used to simulate the dynamic characteristics of the structure. The accelerations actually measured during earthquakes are then used to determine the response of the structure. See LOADS, DYNAMIC; LOADS, TRANSVERSE; SEISMIC RISK. [G.D.B.R.]

Structural chemistry Much of chemistry is explainable in terms of the structures of chemical compounds. The understanding of these structures hinges very strongly on understanding the electronic configurations of the elements. The union of atoms, and therefore the formation of compounds from the elements, is associated with interactions among the extranuclear electrons of the individual atoms. Electronic interactions among atoms may occur in two ways: Electrons may be transferred from one atom to another, or they may be shared by two (or more) atoms. The first type of interaction is called electrovalence and results in the formation of electrically charged monatomic ions. The second, covalence, leads to the formation of molecules and complex ions. *See* CHEMICAL BONDING. [D.H.B.]

In considering structures more complex than those derived from simple monoatomic ions, the logical step is to consider single polyhedral aggregates of atoms. In its most precise sense, structure is used to denote a knowledge of the bonding distances and angles between atoms in chemical compounds and, in turn, the geometrical arrangements which they form. These atomic arrangements and the associated distances and angles serve uniquely as "fingerprints" of these atom spatial configurations, and depend very much on the electronic configurations around atoms. The chemical combination of neutral atoms to produce uncharged species results in molecule formation, whereas the similar combination of atoms or ions possessing a net charge results in the formation of complex ions. A basic understanding of the species formed involves the concept of the coordination polyhedron, which allows a simple classification of the structures of many polyatomic molecules and ions. This type of classification is particularly useful because it conveniently explains the packing together of simple chemical molecules or ions in terms of highly symmetrical polyhedra. There is an obvious connection between polyhedra and the structures found in crystalline solids formed from them. Crystal formation often involves the linking of convex polyhedra by the sharing of corners, edges, or faces, ultimately forming space-filling assemblies in which all faces of each polyhedron are in contact with faces of other polyhedra. The most important simple polyhedrons are the tetrahedron, the trigonal bipyramid, the octahedron, the pentagonal bipyramid, and the cube. The most commonly observed of these polyhedral configurations are the tetrahedron (four faces) and the octahedron (six faces).

The simplest correlative device which accurately summarizes a very large number of structures and enables the chemist to predict, with a good chance of success, the geometric array of the atoms in a compound of known composition, is based on an extreme electrostatic model. This model, or theory, represents the bonds in a purely formal way. The central atom is considered to be a positive ion having a charge equal to its oxidation state. The groups attached to the central atom (the ligands) are then treated either as negative ions or as neutral dipolar molecules. The principal justification for this approach lies in its successful correlation of a vast amount of information.

A number of significant observations can be made with regard to these formulations. There are series of ions, or ions and molecules, having the same type of composition, differing only in the nature of the central ion and the net charge on the aggregate. Examples are found in the series: NO_3^{-} , CO_3^{2-} , BO_3^{3-} , CIO_3^{-} , SO_3^{2-} , PO_3^{3-} ; CIO_4^{-} , SO_4^{2-} , PO_4^{3-} , SiO_4^{4-} ; AIF_6^{3-} , SiF_6^{2-} , PF_6^{-} . The numbers of atomic nuclei and of electrons are the same for all the members of each series; consequently, these are called isoelectronic series. Not only are the several chemical entities in such series isoelectronic, but they are usually identical in geometrical structure (isostructural).

It may also be observed that corresponding ions from a given vertical family of the periodic table commonly vary in coordination number. A useful example is found in N⁵⁺ and P⁵⁺ which form NO₃⁻ and PO₄³⁻, respectively. In addition, some neutral molecules expand their coordination numbers to form stable anionic halo complexes, whereas others do not. Thus, SiF₄ reacts with fluoride ion to form SiF₆²⁻, whereas CF₄ does not form a similar complex ion. The most satisfactory explanation of these and many related observations is conveniently formulated in terms of the electrostatic model chosen here.

The necessary condition for stability of the coordination polyhedron MA_n requires that the anions A are each in contact with the central atom M. As a consequence of this condition, the limit of stability of the structure arises in those cases where the anions are also mutually in contact. Larger ligands, or anions, would not be in contact with the central ion. This relationship is usually summarized in terms of the limiting ratio of the radius of the cation, r_M , to that of the anion, r_A , below which the anions would no longer be in contact with the cation.

According to the valence-bond theory, the principal requirements for the formation of a covalent bond are a pair of electrons and suitably oriented electron orbitals on each of the atoms being bonded. The geometry of the atoms in the resulting coordination polyhedron is correlated with the orientation of the orbitals on the central atom. The orbitals used depend on the energies of the electrons in them. In general, the order of increasing energy of the electron orbitals is (n - 1)d < ns< np < nd. It is concluded that a nontransition atom having one valence electron will form a covalent bond utilizing an s orbital. In those cases where an unshared pair of electrons may be assigned to the *ns* orbital, as many as three equivalent bonds may be formed by utilizing the three np orbitals of the central atom. Because of the orientation of these p orbitals with respect to each other, the three resulting bonds should be at 90° to each other. This expectation is nearly realized in PH₃. In order to account for four or six equivalent bonds, or for that matter in order to account for all the remaining polyhedral and polygonal structures, except the angular structure for a coordination number of 2 (with two unshared pairs of electrons on the central atom), an additional assumption is necessary. It is assumed that s and p, s and d, or s, p, and dorbitals, are replaced by new orbitals, called hybridized orbitals. These hybridized orbitals are derived from the original orbitals (mathematically) in such a way that the required number of equivalent bonds may be formed. In the simplest case, it is shown that s and p may be combined to form two equivalent sp hybridized orbitals directed at 180° to each other. Other sets of hybridized orbitals have been shown to be appropriate to describe the bonding in other structures. See LIGAND FIELD THEORY.

Among inert-gas ions of the first row of eight elements in the periodic table, there are four orbitals available for covalent bond formation, one 2s and three 2p. Consequently, a maximum of four bonds may be formed. This is in general agreement with the existence of the tetrahedron as the limiting coordination polyhedron among these elements, for example, BeF_4^{2-} , BF_4^{-} , CCl_4 , NH_4^+ . Although only Li⁺ deviates from this pattern, having a coordination number of 6 in its crystalline halides, these compounds are best treated as simple electrovalent salts. In keeping with the limitation of only four orbitals, the formation of dou-

ble or triple bonds between atoms of these elements reduces the coordination number of the central atom. Thus, the highest coordination number of a first-row element forming one double bond is 3. This is illustrated by the structures below.

In these and similar examples, the geometric array is determined by the formation of three single bonds utilizing sp^2 hybridized orbitals on the central atom and p orbitals on the ligand. In general, the bonds determining the geometry of a molecule or ion (in this way) are called σ bonds. The double bond results from the superposition of a second bond, a π bond, between two atoms. In this example the formation of a π bond reduces the number of σ bonds from four to three, thus changing the geometries of the corresponding molecules or ions from tetrahedral to trigonal planar.

The formation of a second π bond (a triple bond or two double bonds) reduces the coordination number of the atom in question still further, resulting in the linear *sp* set of hybridized orbitals being utilized in σ -bond formation. See VALENCE.

With regard to the nature of doubly bonded compounds, another problem arises when such structures are viewed from the standpoint of valence-bond theory. In the species BCl₃, COCl₂, NO₂Cl, and many similar substances, nonequivalent bonds are predicted. The doubly bonded oxygen should be closer to the central carbon atom than the singly bonded ones. This is not found to be true experimentally so long as the similar atoms are otherwise equivalent. There is only one observable C-O distance in carbonate, one N-O distance in nitrate, and so on. To account for such facts as these, the concept of resonance must be introduced. If the π bond exists, it must exist equally between the central atom and all the equivalent oxygen atoms. The resonance method of describing this situation is to say that one of the pictorial structures is inadequate to describe the substance properly, but that enough pictorial structures (resonance structures) should be considered to permute the double bond about all the equivalent bonds. The true structure is assumed to be something intermediate to all the resonance structures and more stable than any of them because it exists in preference to any one of them. The resonance structures for CO_3^{2-} are the following:



See CONJUGATION AND HYPERCONJUGATION; RESONANCE (MOLE-CULAR STRUCTURE).

The classic homologous series of compounds in organic chemistry provide useful examples involving the condensation of polyhedrons containing the same central element in the individual units. The general formula, $C_nH_{2n} + 2$, represents a large number of compounds extending from the lowest member, methane, CH_4 , to polyethylene, a plastic of economic importance in which *n* is a very large number. Two ways exist for the linking together of these tetrahedrons. This gives rise to two molecular forms, both of which are stable, well-known compounds. It is



Fig. 1. Geometric structure of propane.



Fig. 2. Cis-trans isomerism among olefins.



Fig. 3. Benzene molecule. (a) Structural formula. (b) Two forms in resonance.

an essential part of these structures that each -C link is linear (because it is merely a σ -bond); however, when two carbon atoms are linked to a third, the C-C-C angle is essentially determined by the bond angle of the central carbon atom (that is, the other carbons may be treated as ligands to the first (Fig. 1). The other familiar homologous series of organic chemistry differ from the saturated hydrocarbons in having at least one unique coordination polyhedron of a different type. The olefins contain two doubly bonded, or unsaturated, carbon atoms, whose polyhedral structures are trigonal planar, and the remainder of the carbons are tetrahedral. As in the case of the aliphatic hydrocarbons, the olefins exhibit an isomerism which is associated with the branching of the chain structure. In addition, the presence of two linked trigonal planar carbon atoms and the fact that the polyhedrons cannot rotate about the double bond give rise to a different kind of isomerism, called cis-trans isomerism (Fig. 2). See BOND ANGLE AND DISTANCE.

The existence of a predicted isomerism provides one of the most important confirmations of the theories of chemical structure. In general, the polyhedral view of molecular structure, as described here, has been thoroughly verified by the discovery of the many types of predicted isomerism. The first really convincing proof of the tetrahedral structures of saturated carbon atoms involved optical isomerism. See OPTICAL ACTIVITY.

The aromatic hydrocarbons are characterized by cyclic arrangements of trigonal planar carbon atoms (Fig. 3*a*). The highly symmetrical nature of the benzene molecule is not fully represented by such a structure. The figure indicates the presence of three double and three single bonds in the ring. It has been shown that the C—C bonds are all the same and, consequently, the true structure of the substance must be represented by two resonance structures which interchange the single and double bonds (Fig. 3*b*). See BENZENE. [J.M.Wi.]

Structural connections Methods of joining the individual members of a structure to form a complete assembly. The connections furnish supporting reactions and transfer loads from

Structural connections 2133

one member to another. Loads are transferred by fasteners (rivets, bolts) or welding supplemented by suitable arrangements of plates, angles, or other structural shapes. When the end of a member must be free to rotate, a pinned connection is used.

The suitability of a connection depends on its deformational characteristics as well as its strength. Rotational flexibility or complete rigidity must be provided according to the degree of end restraint assumed in the design. A rigid connection maintains the original angles between connected members virtually unchanged after loading. Flexible or nonrestraining connections permit rotation approximately equal to that at the ends of a simply supported beam. Intermediate degrees of restraint are called semirigid.

A commonly used form of connection for rolled-beam sections, called a web connection, consists of two angles which are attached to opposite sides of a member and which are in turn connected to the web of a supporting beam, girder, column, or framing at right angles. A shelf angle may be added to facilitate erection (Fig. 1).

A bracket or seat on which the end of the beam rests is a seat connection; it is intended to furnish the end reaction of the supported beam. Two general types are used: The unstiffened seat provides bearing for the beam by a projecting plate or angle leg which offers resistance only by its own flexural strength (Fig. 2); the stiffened seat is supported by a vertical plate or angle which transfers the reaction force to the supporting member without flexural distortion of the outstanding seat.

When the action line of a transferred force does not pass through the centroid of the connecting fastener group or welds, the connection is subjected to rotational moment which produces additional shearing stresses in the connectors. The load transmitted by diagonal bracing to a supporting column flange through a gusset plate is eccentric with reference to the connecting fastener group.



Fig. 1. Riveted or bolted web connections. (*a*) Beam-toginder, elevation view. (*b*) Plan view. (*c*) Bending of angle legs.



Fig. 2. Unstiffened seat connections. (a) Fastened angle seat. (b) Welded angle seat. R = reaction load; e = distance to reaction from column face.

2134 Structural deflections

In beam-to-column connections and stiffened seat connections or when members transfer loads to columns by a gusset plate or a bracket, the fasteners are subjected to tension forces caused by the eccentric connection. Although there are initial tensions in the fasteners, the final tension is not appreciably greater than the initial tension.

Rigidity and moment resistance are necessary at the ends of beams forming part of a continuous framework which must resist lateral and vertical loads. Wind pressures tend to distort a building frame, producing bending in the beams and columns which must be suitably connected to transfer moment and shear. The resisting moment can be furnished by various forms of angle T for fasteners or welded or bracket connections.

Where appreciably angular change between members is expected, and in special cases where a hinge support without moment resistance is desired, connections are pinned. Many bridge trusses and large girder spans have pin supports. *See* BOLT; JOINT (STRUCTURES); RIVET. [J.B.S.]

Structural deflections The deformations or movements of a structure and its components, such as beams and trusses, from their original positions. It is as important for the designer to determine deflections and strains as it is to know the stresses caused by loads. *See* STRESS AND STRAIN.

Deflections may be computed by any of several methods. Generally the computation is based on the assumption that stress is proportional to strain. As a result, deflection equations involve the modulus of elasticity E, which is a measure of the stiffness of a material.

The relation between deflections at different parts of a structure is indicated by Maxwell'slaw of reciprocal deflections. This states that if a load P is applied at any point A in any direction aand causes a shift of another point B in direction b, the same load applied at B in direction b will cause an equal shift of Ain direction a (see illustration). The law is used in a number of ways such as in simplifying deflection calculations, checking the accuracy of computations, and producing influence lines. See STRUCTURAL ANALYSIS.



Example of Maxwell's law of reciprocal deflections.

Beam and truss deflections usually are computed by similar methods, except that integration is used for equations and summation for trusses. Beam deflection equations involve bending moments and moments of inertia. Truss deflection equations are based on the stresses and cross-sectional areas of chords and web members. Deflections may also be determined graphically. See BEAM; TRUSS. [J.B.S.]

Structural design The selection of materials and member type, size, and configuration to carry loads in a safe and serviceable fashion. In general, structural design implies the engineering of stationary objects such as buildings and bridges, or objects that may be mobile but have a rigid shape such as ship hulls and aircraft frames. Devices with parts planned to move with relation to each other (linkages) are generally assigned to the area of mechanical design.

Structural design involves at least five distinct phases of work: project requirements, materials, structural scheme, analysis, and design. For unusual structures or materials a sixth phase, testing, should be included. These phases do not proceed in a rigid progression, since different materials can be most effective in different schemes, testing can result in changes to a design, and a final design is often reached by starting with a rough estimated design, then looping through several cycles of analysis and redesign. Often, several alternative designs will prove quite close in cost, strength, and serviceability. The structural engineer, owner, or end user would then make a selection based on other considerations.

Before starting design, the structural engineer must determine the criteria for acceptable performance. The loads or forces to be resisted must be provided. For specialized structures this may be given directly, as when supporting a known piece of machinery, or a crane of known capacity. For conventional buildings, building codes adopted on a municipal, county, or state level provide minimum design requirements for live loads (occupants and furnishings, snow on roofs, and so on). The engineer will calculate dead loads (structure and known, permanent intallations) during the design process. For the structure to be serviceable or useful, deflections must also be kept within limits, since it is possible for safe structures to be uncomfortably "bouncy." Very tight deflection limits are set on supports for machinery, since beam sag can cause driveshafts to bend, bearings to burn out, parts to misalign, and overhead cranes to stall. Beam stiffness also affects floor "bounciness," which can be annoying if not controlled. In addition, lateral deflection, sway, or drift of tall buildings is often held within approximately height/500 (1/500)of the building height) to minimize the likelihood of motion discomfort in occupants of upper floors on windy days. See LOADS, DYNAMIC; LOADS, TRANSVERSE.

Technological advances have created many novel materials such as carbon fiber- and boron fiber-reinforced composites, which have excellent strength, stiffness, and strenth-to-weight properties. However, because of the high cost and difficult or unusual fabrication techniques required, glass-reinforced composites such as fiberglass are more common, but are limited to lightly loaded applications. The main materials used in structural design are more prosaic and include steel, aluminum, reinforced concrete, wood, and masonry. *See* COMPOSITE MATERIAL; MA-SONRY; PRECAST CONCRETE; PRESTRESSED CONCRETE; REINFORCED CONCRETE; STRUCTURAL MATERIALS.

In an actual structure, various forces are experienced by structural members, including tension, compression, flexure (bending), shear, and torsion (twist). However, the structural scheme selected will influence which of these forces occurs most frequently, and this will influence the process of material selection. *See* SHEAR; TORSION.

Analysis of structures is required to ensure stability (static equilibrium), find the member forces to be resisted, and determine deflections. It requires that member configuration, approximate member sizes, and material properties be known or assumed. Aspects of analysis include: equilibrium; stress, strain, and elastic modulus; linearity; plasticity; and curvature and plane sections. Various methods are used to complete the analysis.

Once a structure has been analyzed (by using geometry alone if the analysis is determinate, or geometry plus assumed member sizes and materials if indeterminate), final design can proceed. Deflections and allowable stresses or ultimate strength must be checked against criteria provided either by the owner or by the governing building codes. Safety at working loads must be calculated. Several methods are available, and the choice depends on the types of materials that will be used. Once a satisfactory scheme has been analyzed and designed to be within project criteria, the information must be presented for fabrication and construction. This is commonly done through drawings, which indicate all basic dimensions, materials, member sizes, the anticipated loads used in design, and anticipated forces to be carried through connections. [R.L.T.; L.M.J.]

Structural geology The branch of geology that deals with study and interpretation of deformation of the Earth's crust. Deformation brings about changes in size (dilation), shape

(distortion), position (translation), or orientation (rotation). Evidence for the changes caused by deformation are commonly implanted into geologic bodies in the form of recognizable structures, such as faults and joints, folds and cleavage, and foliation and lineation. The geologic record of structures and structural relations is best developed and most complicated in mountain belts, the most intensely deformed parts of the Earth's crust. *See* MOUNTAIN SYSTEMS.

The discipline of structural geology harnesses three interrelated strategies of analysis: descriptive analysis, kinematic analysis, and dynamic analysis. Descriptive analysis is concerned with recognizing and describing structures and measuring their orientations. Kinematic analysis focuses on interpreting the deformational movements responsible for the structures. Dynamic analysis interprets deformational movements in terms of forces, stresses, and mechanics. The ultimate goal of these interdependent approaches is to interpret the physical evolution of crustal structures, that is, tectonic analysis. A major emphasis in modern structural geology is strain analysis, the quantitative analysis of changes in size and shape of geologic bodies, regardless of scale.

There are many significant practical applications of structural geology. An understanding of the descriptive and geometric properties of folds and faults, as well as mechanisms of folding and faulting, is of vital interest to exploration geologists in the petroleum industry. Ore deposits commonly are structurally controlled, or structurally disturbed, and for these reasons detailed structural geologic mapping is an essential component of mining exploration. Other applications of structural geology include the evaluation of proposals for the disposal of radioactive waste in the subsurface, and the targeting of safe sites for dams, hospitals, and the like in regions marked by active faulting. See FAULT AND FAULT STRUCTURES; FOLD AND FOLD SYSTEMS; ORE AND MINERAL DEPOSITS; PETROLEUM GEOLOGY. [G.H.D.]

Structural materials Construction materials which, because of their ability to withstand external forces, are considered in the design of a structural framework.

Brick is the oldest of all artificial building materials. It is classified as face brick, common brick, and glazed brick. Face brick is used on the exterior of a wall and varies in color, texture, and mechanical perfection. Common brick consists of the kiln run of brick and is used behind whatever facing material is employed providing necessary wall thickness and additional structural strength. Glazed brick is employed largely for interiors where beauty, ease of cleaning, and sanitation are primary considerations. *See* BRICK.

Structural clay tiles are burned-clay masonry units having interior hollow spaces termed cells. Such tile is widely used because of its strength, light weight, and insulating and fire-protection qualities. *See* TILE.

Architectural terra-cotta is a burned-clay material used for decorative purposes. The shapes are molded either by hand in plaster-of-paris molds or by machine, using the stiff-mud process.

Building stones generally used are limestone, sandstone, granite, and marble. Until the advent of steel and concrete, stone was the most important building material. Its principal use now is as a decorative material because of its beauty, dignity, and durability. *See* GRANITE; LIMESTONE; MARBLE; SANDSTONE; STONE AND STONE PRODUCTS.

Concrete is a mixture of cement, mineral aggregate, and water, which, if combined in proper proportions, form a plastic mixture capable of being placed in forms and of hardening through the hydration of the cement. *See* CONCRETE; PRESTRESSED CONCRETE; REINFORCED CONCRETE.

The cellular structure of wood is largely responsible for its basic characteristics, unique among the common structural materials. When cut into lumber, a tree provides a wide range of material which is classified according to use as yard lumber, factory or shop lumber, and structural lumber. Laminated lumber is used for beams, columns, arch ribs, chord members, and other structural members. Plywood is generally used as a replacement for sheathing, or as form lumber for reinforced concrete structures. *See* LUMBER; PLYWOOD; WOOD ANATOMY; WOOD PRODUCTS.

Important structural metals are the structural steels, steel castings, aluminum alloys, magnesium alloys, and cast and wrought iron. Steel castings are used for rocker bearings under the ends of large bridges. Shoes and bearing plates are usually cast in carbon steel, but rollers are often cast in stainless steel. Aluminum alloys are strong, lightweight, and resistant to corrosion. The alloys most frequently used are comparable with the structural steels in strength. Magnesium alloys are produced as extruded shapes, rolled plate, and forgings. The principal structural applications are in aircraft, truck bodies, and portable scaffolding. Gray cast iron is used as a structural material for columns and column bases, bearing plates, stair treads, and railings. Malleable cast iron has few structural applications. Wrought iron is used extensively because of its ability to resist corrosion. It is used for blast plates to protect bridges, for solid decks to support ballasted roadways, and for trash racks for dams. See ALUMINUM: CAST IRON; MAGNESIUM ALLOYS; STEEL; STRUCTURAL STEEL; WROUGHT IRON. [C.M.A.]

Composite materials are engineered materials that contain a load-bearing material housed in a relatively weak protective matrix. A composite material results when two or more materials, each having its own, usually different characteristics, are combined, producing a material with properties superior to its components. The matrix material (metallic, ceramic, or polymeric) bonds together the reinforcing materials (whiskers, laminated fibers, or woven fabric) and distributes the loading between them. See COMPOSITE MATERIAL; CRYSTAL WHISKERS.

Fiber-reinforced polymers (FRP) are a broad group of composite materials made of fibers embedded in a polymeric matrix. Compared to metals, they generally have relatively high strength-to-weight ratios and excellent corrosion resistance. They can be formed into virtually any shape and size. Glass is by far the most used fiber in FRP (glass-FRP), although carbon fiber (carbon-FRP) is finding greater application. Although complete FRP shapes and structures are possible, the most promising application of FRP in civil engineering is for repairing structures or infrastructure. FRP can be used to repair beams, walls, slabs, and columns. See POLYMERIC COMPOSITE. [M.Sc.]

Structural petrology The study of the structural aspects of rocks, as distinct from the purely chemical and mineralogical studies that are generally emphasized in other branches of petrology. The term was originally used synonymously with petrofabric analysis, but is sometimes restricted to denote the analysis of only microscopic structural and textural features. *See* PETROFABRIC ANALYSIS; PETROGRAPHY. [J.M.Ch.]

Structural plate A simple rolled steel section used as an isolated structural element, as a support of other structural elements, or as part of other structural elements. When isolated plates are extremely thick, their design is controlled by shear; plates of moderate thickness are controlled by bending (with some torsion), and very thin plates carry their loads principally by tensile membrane action. Although stresses of all types exist in all plates, it is usually sufficient to deal with only the most significant. Bending is the most common design criterion.

Plates are commonly used as cover plates on wide-flange beams, as the flanges and webs of plate girders, and as the sides of tube-shaped beams and columns. In all these cases, serious consideration must be given to the fact that the plate may buckle when compressed. Fortunately, the plates have edge supports in the direction of the stress, so they function as panels rather than as beams. Their ratios of length to width are large enough that the resistance to local buckling of the plate element depends upon its width-thickness ratio, practically independent of its length. (The length of the overall section is still significant in

2136 Structural steel

determining the member's capacity.) See BEAM; COLUMN; JOINT (STRUCTURES); LOADS, TRANSVERSE; STRUCTURAL STEEL. [G.D.Br.]

Structural steel Steel used in engineering structures, usually manufactured by either the open-hearth or the electric-furnace process. The exception is carbon-steel plates and shapes whose thickness is $7/_{16}$ in. (11 mm) or less and which are used in structures subject to static loads only. These products may be made from acid-Bessemer steel. The physical properties and chemical composition are governed by standard specifications of the American Society for Testing and Materials (ASTM). Structural steel can be fabricated into numerous shapes for various construction purposes. See STEEL. [W.G.B.]

Structure (engineering) An arrangement of designed components that provides strength and stiffness to a built artifact such as a building, bridge, dam, automobile, airplane, or missile. The artifact itself is often referred to as a structure, even though its primary function is not to support but, for example, to house people, contain water, or transport goods. *See* AIRPLANE; AUTOMOBILE; BRIDGE; BUILDINGS; DAM.

The primary requirements for structures are safety, strength, economy, stiffness, durability, robustness, esthetics, and ductility. The safety of the structure is paramount, and it is achieved by adhering to rules of design contained in standards and codes, as well as in exercising strict quality control over all phases of planning, design, and construction. The structure is designed to be strong enough to support loads due to its own weight, to human activity, and to the environment (such as wind, snow, earthquakes, ice, or floods). The ability to support loads during its intended lifetime ensures that the rate of failure is insignificant for practical purposes. The design should provide an economical structure within the constraints of all other requirements. The structure is designed to be stiff so that under everyday conditions of loading and usage it will not deflect or vibrate to an extent that is annoying to the occupants or detrimental to its function. The materials and details of construction have durability, such that the structure will not corrode, deteriorate, or break under the effects of weathering and normal usage during its lifetime. A structure should be robust enough to withstand intentional or unintentional misuse (for example, fire, gas explosion, or collision with a vehicle) without totally collapsing. A structural design takes into consideration the community's esthetic sensibilities. Ductility is necessary to absorb the energy imparted to the structure from dynamic loads such as earthquakes and blasts. See CONSTRUCTION ENGINEERING; ENGINEERING DESIGN.

Common structural materials are wood, masonry, steel, reinforced concrete, aluminum, and fiber-reinforced composites. Structures are classified into the categories of frames, plates, and shells, frequently incorporating combinations of these. Frames consist of "stick" members arranged to form the skeleton on which the remainder of the structure is placed. Plated structures include roof and floor slabs, vertical shear walls in a multistory building, or girders in a bridge. Shells are often used as water or gas containers, in roofs of arenas, or in vehicles that transport gases and liquids. The connections between the various elements of a structure are made by bolting, welding or riveting. *See* COMPOSITE MATERIAL; CONCRETE; STEEL; STRUCTURAL MATERI-ALS. [T.V.G.]

Struthioniformes A small order of weak-flying, partridgelike birds and giant flightless ratite birds found in the southern continents. Their relationship to other birds is unknown. The struthioniforms are characterized by a palaeognathous palate which provides freedom of movement between segments of the palate and the jaw. For this reason they are frequently placed in a separate superorder, the Palaeognathae. However, that distinction places too much emphasis on their separation from other birds. Contrary to common opinion, no evidence supports the concept that the palaeognathous birds are primitive among living birds.

The order Struthioniformes can be divided into three suborders, Lithornithi, Tinami, and Struthioni, each of which includes both fossil and extant representatives.

The struthioniforms are medium-sized to giant birds. The ostriches are the largest extant birds, but some of the moas, elephant birds, and dromornithids were even larger. The head is small; a medium flattened bill is a common feature except in the kiwis, which have a long bill used for probing into the ground for worms. The wings are reduced in all forms except the lithionithids and tinamous, which can still fly, and the plumage is soft. Their legs are strong, and all forms run well. The struthioniforms eat a variety of foods, especially large fruits and other large food items. Breeding is polygamous, with two or more females laying eggs in a single nest. The males are responsible for incubation, and they assume the major or sole role in caring for the downy young, which leave the nest after hatching. *See* AVES; RATITES. [W.J.B.]

Strychnine alkaloids Alkaloid substances derived from the seeds and bark of plants of the genus *Strychnos* (family Loganiaceae). This genus serves as the source of poisonous, nitrogen-containing plant materials, such as strychnine (see structure; R = H). The seeds of the Asian species of *Strychnos* contain 2–3% alkaloids, of which about half is strychnine



and the rest is closely related materials; for example, brucine (see structure; $R = OCH_3$) is a more highly oxygenated relative. Strychnine and brucine are isolated by extraction of basified plant residue with chloroform and then, from the chloroform solution, by dilute sulfuric acid. Precipitation from the dilute acid is accomplished with ammonium hydroxide. Strychnine is separated from brucine by fractional crystallization from ethanol. See CRYSTALLIZATION; STRYCHNOS.

At one time strychnine was used as a tonic and a central nervous system stimulant, but because of its high toxicity (5 mg/kg is a lethal dose in the rat) and the availability of more effective substances, it no longer has a place in human medicine. [D.D.]

Strychnos A genus of tropical trees and shrubs belonging to the Logania family (Loganiaceae). *Strychnos nux-vomica*, a native of India and Ceylon, is the source of strychnine. The alkaloid, strychnine, has been used medicinally in the treatment of certain nervous disorders and paralysis. Curare, used by the Indians to poison arrows, is obtained from *S. toxifera* and *S. castelnaei* in Guiana and Amazonas and from *S. tieute* in the Sunda Islands. Curare paralyzes the motor nerve endings in striated muscles and is used in medical practice in cases in which a state of extreme muscular relaxation or even immobility is desirable. It has become an important drug in the field of anesthesiology. *See* GENTIANALES; STRYCHNINE ALKALOIDS. [PD.St.; E.L.C.]

Sturgeon Any of 10 species of large fish which comprise the family Acipenseridae in the order Acipenseriformes. These fish are found in North Temperate Zone waters, where they are almost exclusively bottom-living and feed on organisms such as mollusks, worms, and larvae. The body has five rows of bony plates, of which one is situated dorsally, two laterally, and two ventrally. The snout is elongate and there are four barbels on its lower surface; the mouth is ventrally located, and in the adult

Subduction zones 2137



Two species of sturgeons found in United States coastal waters. (a) Atlantic sturgeon (*Acipenser oxyrhynchus*). (b) Short-nosed sturgeon (*A. brevirostrus*).

the jaws lack teeth (see illustration). The skeleton is mainly cartilaginous.

Sturgeons are known primarily for the roe, which is processed as caviar. A single female can produce millions of eggs, which may be removed from dead fish or stripped from living fish. The smoked flesh of the sturgeon is a delicacy. See ACIPENSERIFORMES. [C.B.C.]

Stylasterina An order of the class Hydrozoa of the phylum Coelenterata, including several brightly colored branching or encrusting "corals" of warm seas. (True corals belong to a different class, the Anthozoa). The calcareous skeleton is covered by living tissue and is penetrated by ramifying tubes. Nutritive polyps, the gastrozooids, lie in cups on one surface or along certain edges of the skeletal substance. A spine, or style, at the base of each cup gives the order its name.

Some authorities combine the Stylasterina with the Milleporina in a single order, the Hydrocorallina. See ANTHOZOA; HY-DROZOA. [S.Cr.]

Stylolites Irregular surfaces occurring in certain rocks, mostly parallel to bedding planes, in which small toothlike projections on one side of the surface fit into cavities of like shape on the other side (see illustration). Stylolites are most common



Stylolite in limestone.

in limestones and dolomites but are also present in many other kinds of rock, including sandstones, gypsum beds, and cherts. *See* DOLOMITE; LIMESTONE; SEDIMENTARY ROCKS. [R.Si.]

Stylommatophora A superorder of the molluscan subclass Pulmonata containing about 20,500 species that are grouped into 56 families. Nearly all land snails without an operculum are stylommatophorans. They have eyespots on the tips of a pair of retractile tentacles, hermaphroditic reproduction with partial fusion of the male and female system, and, normally, a second pair of retractile tentacles that function as chemoreceptors. All are air-breathing and most are truly terrestrial. Three orders, based on excretory structures, are recognized. The more primitive Orthurethra and Mesurethra have less efficient water conservation devices than do the more specialized members of the order Sigmurethra, in which a closed ureter functions in water conservation. The latter specialization apparently is a prerequisite for evolution toward a sluglike structure, since all 16 families with slugs or sluglike taxa belong to the Sigmurethra. *See* PULMONATA. [G.A.S.]

Styrene A colorless, liquid hydrocarbon with the formula $C_6H_5CH=CH_2$. It boils at 145.2°C (293.4°F) and freezes at -30.6°C (-23.1°F). The ethylenic linkage of styrene readily undergoes addition reactions and under the influence of light, heat, or catalysts undergoes self-addition or polymerization to yield polystyrene.

The majority of the styrene used is converted into polystyrene, but other thermoplastic or even thermosetting resins are prepared from styrene by copolymerization with suitable comonomers. A smaller quantity of styrene goes into the manufacture of elastomers or synthetic rubbers.

Styrene is a skin irritant. Prolonged breathing of air containing more than 400 ppm of styrene vapor may be injurious to health. See POLYMERIZATION; POLYSTYRENE RESIN. [C.K.B.]

Subduction zones Regions where portions of the Earth's tectonic plates are diving beneath other plates, into the Earth's interior. Subduction zones are defined by deep oceanic trenches, lines of volcanoes parallel to the trenches, and zones of large earthquakes that extend from the trenches landward.

Plate tectonic theory recognizes that the Earth'ssurface is composed of a mosaic of interacting lithospheric plates, with the lithosphere consisting of the crust (continental or oceanic) and associated underlying mantle, for a total thickness of about 100 km (60 mi). Oceanic lithosphere is created by sea-floor spreading at mid-ocean ridges (divergent, or accretionary, plate boundaries) and destroyed at subduction zones (at convergent, or destructive, plate boundaries). At subduction zones, the oceanic lithosphere dives beneath another plate, which may be either oceanic or continental. Part of the material on the subducted plate is recycled back to the surface (by being scraped off the subducting plate and accreted to the overriding plate, or by melting and rising as magma), and the remainder is mixed back into the Earth's deeper mantle. This process balances the creation of lithosphere that occurs at the mid-ocean ridge system. The convergence of two plates occurs at rates of 1-10 cm/yr (0.4-4 in./yr) or 10-100 km (6–60 mi) per million years (see illustration).

During subduction, stress and phase changes in the upper part of the cold descending plate produce large earthquakes in the upper portion of the plate, in a narrow band called the Wadati-Benioff zone that can extend as deep as 700 km (420 mi). The plate is heated as it descends, and the resulting release of water leads to melting of the overlying mantle. This melt rises to produce the linear volcanic chains that are one of the most striking features of subduction zones. *See* LITHOSPHERE.

Subduction zones can be divided in two ways, based either on the nature of the crust in the overriding plate or on the age of the subducting plate. The first classification yields two broad categories: those beneath an oceanic plate, as in the Mariana or Tonga trenches, and those beneath a continental plate, as along the west coast of South America (see illustration). The first type is known as an intraoceanic convergent margin, and the second is known as an Andean-type convergent margin. *See* MID-OCEANIC RIDGE; PLATE TECTONICS.

Active volcanoes are highly visible features of subduction zones. The volcanoes that have developed above subduction zones in East Asia, Australasia, and the western Americas surround the Pacific Ocean in a so-called Ring of Fire. At

2138 Subgiant star



Principal subduction zones. Plate names are in all-capitals. Numbers at arrows indicate velocity of plate movement in kilometers (miles) per million years. (*After J. Gill, Orogenic Andesites and Plate Tectonics, Springer-Verlag, 1981*)

intraoceanic convergent margins, volcanoes may be the only component above sea level, leading to the name "island arc." The more general term "volcanic arc" refers to volcanoes built on either oceanic or continental crust. See VOLCANO.

An eventual consequence of subduction is orogeny, or mountain building. Subduction zones are constantly building new crust by the production of volcanic material or the accretion of oceanic sediments. However, the development of the greatest mountain ranges—the Alps or the Himalayas—occurs not during "normal" subduction but during the death of a subduction zone, when it becomes clogged with a large continent or volcanic arc. *See* OROGENY. [R.J.Ste.; S.H.B.]

Subgiant star An evolving star of luminosity class IV. Such a star is brighter than the main-sequence dwarfs and fainter than the true giants in its spectral class, lying between the two on the Hertzsprung-Russell diagram. The classic subgiants fall in a small region from class F to K (with effective temperatures ranging from 7000 to 4000 K or 12,000 to 7000°F). In class G they lie at absolute visual magnitude +3 with luminosities about five times the solar luminosity. Classic subgiants have masses around 1.3 times that of the Sun and violate the mass-luminosity relation as too bright for their masses. The concept is extended to the hot part of the Hertzsprung-Russell diagram from class F to O, in which the distinctions between subgiants and neighboring dwarfs and giants are much smaller, only a magnitude or less.

Main-sequence dwarfs run on the fusion of hydrogen to helium in their cores. The subgiant stage begins when the core hydrogen mass fraction drops to around 0.1, and then continues as the fraction goes to zero and the star evolves toward, but not onto, the red-giant branch with a contracting helium core. See DWARF STAR; GIANT STAR; HERTZSPRUNG-RUSSELL DIAGRAM; SPECTRAL TYPE; STAR; STELLAR EVOLUTION. [J.B.Ka.] **Sublimation** The process by which solids are transformed directly to the vapor state without passing through the liquid phase. Sublimation is of considerable importance in the purification of certain substances such as iodine, naphthalene, and sulfur.

Sublimation is a universal phenomenon exhibited by all solids at temperatures below their triple points. For example, it is a common experience to observe the disappearance of snow from the ground even though the temperature is below the freezing point and liquid water is never present. The rate of disappearance is low, of course, because the vapor pressure of ice is low below its triple point. Sublimation is a scientifically and technically useful phenomenon, therefore, only when the vapor pressure of the solid phase is high enough for the rate of vaporization to be rapid. See PHASE EQUILIBRIUM; TRIPLE POINT; VAPOR PRESSURE. [N.H.N.]

Submarine A ship which can operate completely submerged in the water. The term formerly applied to any ship capable of operating completely underwater, but now usually describes a ship built for military purposes. The term "submersible" usually is applied to small, underwater vehicles that are built for research, rescue, commercial work, or pleasure.

By the end of World War II, antisubmarine warfare had progressed significantly by exploiting the limited underwater endurance and speed of the diesel-electric designs of that era. The application of nuclear power to submarines after World War II reestablished the near-invulnerability of the submarine to antisubmarine warfare from surface ships and aircraft. Nuclear power depends on nuclear fission rather than the oxidation of fossil fuels and thus requires no oxygen source as do diesel engines, allowing the submarine to operate submerged for very long periods. However, advances in submarine technology and nonnuclear propulsion cause the nonnuclear submarine to remain highly attractive to the navies of many nations. See SHIP NUCLEAR PROPULSION.

Submarines can be classified by their primary military missions. Attack submarines are fast, long-range ships equipped with torpedo tubes or cruise missile launch tubes. They carry sensitive underwater sound receivers and transmitters (sonar) used to detect enemy submarines. They may be armed with torpedoes of various kinds, cruise missiles, mines, and equipment for deployment of small units of clandestine troops.

Ballistic-missile submarines carry long-range missiles fitted with nuclear warheads that can be launched while submerged. The submarine can remain submerged and undetected for many days and, on command, launch missiles on any target within range. The missiles are stowed in and launched from vertical tubes. See MISSILE.

Experimental submarines are occasionally built to test new designs of hull shape, deeper depth capability, power plants, or controls.

Submersibles are usually small, deep-diving vehicles. Their use is for exploration and study of the ocean depths, development of equipment, rescue, or commercial work. Some designs take advantage of the forces of gravity and buoyancy for vertical motion. Other designs use vertically oriented propellers to propel the craft up and down. Movement is restricted to short distances and slow speed because of small size and small battery capacity. See UNDERWATER VEHICLE.

Compared with surface ships, the submarine has features that enable it to submerge and resist great sea pressure. Submarines have a pressure hull and a nonpressure hull. The pressure hull is the watertight, pressure-proof envelope in which equipment operates and the officers and crew live. In certain areas of the submarine there is a nonpressure hull of lighter structure, forming the main ballast tanks. A nonwatertight superstructure provides a smooth, fair envelope to cover pipes, valves, and fittings on top of the hull. Above the superstructure the fairwater similarly encloses the bridge, the periscope, and multiple mast supports. *See* PERISCOPE; SHIP DESIGN.

The principal means of detecting the presence of a submerged submarine is to listen for sounds which may have been generated on board or by its movement through the water. Very small amounts of acoustic energy can be detected by sophisticated sonars. Therefore, modern submarines are designed with multiple features to greatly reduce the amount of noise they generate. [M.S.E; J.H.W.]

Submarine cable A cable, primarily for communications, laid on the ocean floor to provide international links. This term is sometimes applied to power cables in water, but these are not usually of great length.

Coaxial cables with 1840 circuits (each having 3-kHz bandwidth) are used extensively for shorter cables. Cables using digital transmission over optical fibers (lightguide transmission) offer a significant cost-per-channel advantage over coaxial systems. Voice signals are transmitted by laser-generated digital light pulses over single-mode fiber pairs. *See* COAXIAL CABLE; COMMU-NICATIONS CABLE; MULTIPLEXING AND MULTIPLE ACCESS; TRANSMIS-SION LINES.

SL280 digital lightguide transatlantic submarine cables utilize semiconductor laser diodes operating at a wavelength of 1.3micrometers and carries 280 megabits per second on six hairthin fibers, of which two pairs are active and the third is available as standby. Each fiber is a glass strand 125 μ m in diameter, plastic-coated to 250 μ m, and embedded in an elastomer that is part of the unit fiber structure. The fibers are helically wrapped around a central kingwire for strength during manufacture and an outer sheath of nylon to ensure dimensional stability. This structure is then helically wrapped with steel armor wires for strength, ensheathed in a copper conductor, and coated with a final protective layer of polyethylene insulation (see illustration).



Cutaway view of double-armored SL cable.

The fiber-optic digital repeaters regenerate the signal in its original form, unlike the earlier analog repeaters which were signal amplifiers. The repeaters detect the optical signal entering the repeater, transform it into an electrical signal and amplify it, regenerate it by putting it through a high-speed clocked decision circuit, and then convert the regenerated electrical signal into an optical one by using a 1.3- μ m laser transmitter. Transmission rates of 560 megabits per second are provided in the subsequent SL560 design, utilizing lasers operating at 1.5 μ m.

Virtually all analog (coaxial) submarine cable systems were configured as a single line from one landing point to another. Traffic requirements, however, are not identical for all countries, and it was realized to be advantageous to implement a tree configuration, where the various branches go to different countries. Digital fiber-optic technology makes branching possible by dedicating certain fiber pairs to certain subsystems.

With the increased volume and capacity of fiber-optic systems, customers and system owners cannot afford or support the potential loss of business and revenue associated with a cable failure. To reduce this risk, many systems employ a ringnetwork design architecture. These systems use a self-healing loop network with built-in backup and redundancy, which provide fiber-on-fiber restoration and connectivity. In the unlikely event of a cable failure of any nature, communications traffic can be shifted from one fiber cable to another so that digital voice, data, or video communications can continue without disruption. See LASER; OPTICAL COMMUNICATIONS; OPTICAL FIBERS; PULSE MODULATION. [G.J.S.]

Submarine canyon A steep-sided valley developed on the sea floor of the continental slope. Submarine canyons serve as major conduits for sediment transport from land and the continental shelf to the deep sea. Modern canyons are relatively

2140 Submillimeter astronomy

narrow, deeply incised, steeply walled, often sinuous valleys with predominantly V-shaped cross sections. Most canyons originate near the continental-shelf break, and generally extend to the base of the continental slope. Canyons often occur off the mouths of large rivers, such as the Hudson River or the Mississippi River, although many other canyons, such as the Bering Canyon in the southern Bering Sea, are developed along structural trends. *See* CONTINENTAL MARGIN.

Modern submarine canyons vary considerably in their dimensions. The average lengths of canyons has been estimated to be about 34 mi (55 km); although the Bering Canyon is more than 680 mi (1100 km) long and is the world's longest submarine canyon. The shortest canyons are those off the Hawaiian Islands, and average about 6 mi (10 km) in length. Submarine canyons are characterized by relatively steep gradients. The average slope of canyon floors is 309 ft/mi (58 m/km). In general, shorter canyons tend to have higher gradients. For example, shorter canyons of the Hawaiian group have an average gradient of 766 ft/mi (144 m/km), whereas the Bering Canyon has a slope of only 42 ft/mi (7.9 m/km).

In comparison to modern canyons, dimensions of ancient canyons are considerably smaller. Deposits of ancient canyons are good hydrocarbon reservoirs. This is because submarine canyons and channels are often filled with sand that has the potential to hold oil and gas. Examples of hydrocarbon-bearing canyon-channel reservoirs are present in the subsurface in California, Louisiana, and Texas.

Physical and biological processes that are common in submarine canyons are mass wasting, turbidity currents, bottom currents, and bioerosion, mass wasting and turbidity currents being the more important. Mass wasting is a general term used for the failure, dislodgement, and downslope movement of sediment under the influence of gravity. Common examples of mass wasting are slides, slumps, and debris flows. Major slumping events can lead to formation of submarine canyons. The Mississippi Canyon in the Gulf of Mexico is believed to have been formed by retrogressive slumping during the late Pleistocene fall in sea level and has been partially infilled during the Holocene rise in sea level. Turbidity currents are one of the most important erosional and depositional processes in submarine canyons. There is considerable evidence to suggest that turbidity currents flow at velocities of 11-110 in./s (28-280 cm/s) in submarine canyons. Therefore, turbidity currents play a major role in the erosion of canyons. See DEPOSITIONAL SYSTEMS AND ENVIRONMENTS; MARINE GEOLOGY; MARINE SEDIMENTS; MASS WASTING; TURBIDITY CURRENT. [G.Sh.]

Submillimeter astronomy Investigation of the universe by probing the electromagnetic spectrum at wavelengths from approximately 0.3 to 1.0 millimeter: the submillimeter waveband. This waveband is bounded at longer wavelengths by the millimeter waveband (1–10 mm), and at shorter wavelengths by the far-infrared waveband (20–300 micrometers). Astronomical objects with temperatures between about 10 kelvins and several hundred kelvins, typically in the interstellar medium of galaxies, emit radiation strongly in the submillimeter waveband. *See* ELECTROMAGNETIC RADIATION.

The submillimeter waveband is one of the last parts of the electromagnetic spectrum to be investigated. This is due both to the technical challenge of making sensitive detectors, which must be held at temperatures close to absolute zero, and to the effects of the Earth's atmosphere. Atmospheric molecules, primarily water vapor, both absorb (dim) the signal from astronomical sources and emit their own radiation that acts to mask the astronomical signals. The effects are most severe at the shortest wavelengths, and only high mountain sites and air or space-borne platforms can be used for submillimeter astronomy. From high mountains, observations are possible in only three atmospheric windows, at wavelengths of about 0.35, 0.45, and 0.85 mm. Radiation of 0.85 mm penetrates down to about 2 km (1.2 mi) above sea



James Clerk Maxwell Telescope, 49-ft-diameter (15-m) submillimeter telescope situated at Mauna Kea, Hawaii.

level; while on the driest nights, only about one-half of photons with wavelengths of 0.35 or 0.45 μ m that enter the upper atmosphere can be detected at the 4200-m (14,000-ft) summit of Mauna Kea, Hawaii.

Several submillimeter telescopes operate on high mountain tops, including the 15-m (49-ft) James Clerk Maxwell Telescope (JCMT; see illustration) and the 10.4-m (34-ft) Caltech Submillimeter Observatory (CSO) telescope, both on Mauna Kea, and the 15-m (49-ft) Swedish-ESO Submillimeter Telescope (SEST) in Chile. These telescopes look like satellite antennas, but their surfaces are smooth and precisely shaped to an accuracy of order 10 μ m. These ground-based telescopes are subject to more atmospheric absorption than airborne and space telescopes; but because they have larger apertures, they can collect more radiation, boosting their sensitivity. See TELESCOPE.

In 2005 the Stratospheric Observatory for Infrared Astronomy (SOFIA), a 2.5-m-aperture (8.2-ft) telescope in a 747 aircraft, was expected to be operating. Like the former Kuiper Airborne Observatory (KAO), SOFIA will fly at up to 14 km (9 mi) altitude. Stratospheric balloons, flying at altitudes of about 40 km (25 mi), can also be used to carry submillimeter telescopes; an example is the BOOMERANG cosmic microwave background experiment reported in 2000. See INFRARED ASTRONOMY.

The signals from several submillimeter telescopes can be combined together in an interferometer to provide a telescope with resolution as fine as one with an aperture as large as the greatest separation between the individual telescopes. Examples of this type of telescope are the Smithsonian Millimeter Array (SMA) on Mauna Kea, and the United States-Europe-Japan Atacama Large Millimeter Array (ALMA) under development in Chile. See RADIO TELESCOPE.

Submillimeter observations are used by astronomers with a wide range of interests to study matter in the universe at temperatures of about 10–100 K and the cosmic microwave background radiation. Interstellar chemistry in clouds of gas and dust, the formation of protostars embedded in stellar nurseries within the Milky Way, the properties of the interstellar medium in even the most distant galaxies, and the evolution of galaxies are all studied in the submillimeter waveband. Submillimeter observations are especially important for observations of regions that are rich in dust and gas. These are opaque to visible and infrared light, but submillimeter light can still escape, allowing these regions to be studied. [A.B.] **Submillimeter-wave technology** The generation, detection, and application of radiation in the submillimeter-wave region of the electromagnetic spectrum. In practical terms, this region corresponds to wavelengths between about 30 micrometers and 3 millimeters (see illustration). At its extremes the region overlaps with the microwave and infrared, and a unique feature is that the distinct propagation methods of these adjacent regions, guided-wave and free-space propagation, are used throughout it. *See* ELECTROMAGNETIC RADIATION; INFRARED RADIATION; MICROWAVE.

10 ¹⁰ 1	0 ¹¹ 1	0 ¹²	10 ¹³ 1	0 ¹⁴ frequency, Hz
10	1	0.1	0.01	wavelength, mm
microwave 1	submi 10	llimeter 100	infrared 1000	wave number, cm ⁻¹
0.1	1	10	100	energy, meV

Electromagnetic spectrum in and near the submillimeterwave region, showing its frequency, wavelength, wave number, and photon energy equivalents.

Widespread use of the submillimeter-wave region did not occur until two revolutionary developments in the late 1950s and early 1960s: (1) the practical realization of Fourier transform spectroscopy, which led to significant improvements in the quality of broad band submillimeter-wave spectroscopy; and (2) the discovery of two reasonably intense submillimeter-wave laser sources, the hydrogen cyanide and the water-vapor electrically excited lasers. *See* INFRARED SPECTROSCOPY; LASER.

Radiation sources. The main application of incoherent radiation sources is as broad band sources in Fourier transform spectrometers. The two most commonly used sources are the mercury-vapor lamp and the globar. *See* MERCURY-VAPOR LAMP; MICROWAVE TUBE.

There are several categories of coherent submillimeter-wave radiation sources: (1) nonrelativistic electron-beam tubes such as klystrons, extended interaction oscillators, orotrons, ledatrons, and carcinotrons or backward-wave oscillators, which oscillate with varying amounts of tunability and output power; (2) relativistic electron-beam devices such as gyrotrons, freeelectron lasers, and synchrotrons; and (3) solid-state devices such as Gunn and IMPATT oscillators. With the exception of the free-electron laser, these are long-wave devices. The final category, gas lasers, covers the entire submillimeter-wave region. *See* GYROTRON; KLYSTRON; MICROWAVE SOLID-STATE DEVICES; SYN-CHROTRON RADIATION.

Detectors. Most submillimeter-wave detectors are either thermal, rectifying, or photon detectors. The first respond to heat generated by absorbed radiation, the second by rectification of radiation-field-induced currents, the third by absorption of individual photons. They are commonly used in direct and heterodyne detection systems. Thermal detectors include room-temperature thermopile, Golay cell, and pyroelectric detectors, while more sensitive semiconductor bolometers operate at liquid-helium temperatures. Rectifying devices include metalsemiconductor and metal-metal point contacts and Schottky barrier diodes, all room-temperature devices, and Josephson-type devices operated at about 4.2 K $(-452^{\circ}F)$. Submillimeter-wave photon detectors are based on semiconductors and rely on the absorption of a photon changing the free-carrier density or mobility and hence the conductivity of the detector. See BOLOMETER; Pyroelectricity; RADIOMETRY; SUPERCONDUCTING DEVICES.

Atmospheric transparency. Between the visible and microwave regions the atmosphere is effectively opaque, with the exception of two windows in the near-infrared and three in the submillimeter-wave region, at 94, 140, and 220 GHz. The significant feature of the submillimeter-wave windows is that the levels of attenuation involved are not as seriously increased by

adverse atmospheric conditions as those of the near-infrared. This makes them attractive for many applications. The attenuation associated with these windows is, nonetheless, quite high and means that ground-level low-frequency submillimeter-wave systems have a relatively limited range. This can be advantageous for some applications.

Applications. In the past, applications of radiation in the submillimeter-wave region were distributed evenly over the region and largely concerned research activities. Now, however, as a result of developments toward integrated, inexpensive transmitter and receiver systems, applications include many radar- and telecommunications-related activities centered on the 94-GHz atmospheric window. These applications are mainly military, but spin-offs into the civil area are expected.

Photon energies in the submillimeter-wave region correspond to activation energies of many physical, chemical, and bioloical phenomena, and the spectroscopic techniques of this region have been widely applied to their study. Other spectroscopic applications have concerned the development of radiometric methods to meet the requirements of fusion plasma diagnostics and astronomy. See SUBMILLIMETER ASTRONOMY.

The submillimeter-wave region has been extensively used for fundamental frequency metrology, relating the frequencies of submillimeter-wave gas lasers directly back to the primary frequency standard. Important consequences have been the definition of the speed of light as a fixed quantity and the redefinition of the meter as a unit derived from it. *See* LIGHT; PHYSICAL MEA-SUREMENT.

The submillimeter-wave region impinges on fusion plasma research in two ways. First, in machines such as tokamaks it provides a way of studying the spatial and temporal evolution of the electron temperature. The second application is electron cyclotron heating, in which electron heating by microwave radiation at the electron cyclotron frequency has been demonstrated. High-power gyrotron oscillators are generally used as the radiation sources. *See* NUCLEAR FUSION; PLASMA (PHYSICS); PLASMA DIAGNOSTICS.

In the study and applications of fast electromagnetic pulse propagation, a number of novel techniques for the submillimeter-wave region have been developed. The generation and detection of such pulses can be done with ultrashort optical pulses driving photoconductive switches, or by allelectronic systems based on nonlinear transmission lines. The submillimeter-wave radiation used in these systems is variously known as T waves or terahertz waves. See OPTICAL PULSES. [J.R.Bi.]

Subsonic flight Movement of a vehicle through the atmosphere at a speed appreciably below that of sound waves. Subsonic flight extends from zero (hovering) to a speed approximately 85% of sonic speed corresponding to the ambient temperature. At higher vehicle velocities the local velocity of air passing over the vehicle surface may exceed sonic speed, and true subsonic flight no longer exists.

Vehicle type may range from a small helicopter, which operates at all times in the lower range of the velocity scale, to an intercontinental ballistic missile, which is operative throughout this and other velocity regimes, but is in subsonic flight for only a few seconds. The design of each is affected by the same principles of subsonic aerodynamics. Subsonic flow of a fluid such as air may be subdivided into a range of velocities in which the flow may be considered incompressible without appreciable error (below a velocity of approximately 300 mi/h or 135 m/s), and a higher range in which the compressible nature of the fluid becomes significant. In both cases the viscosity of the fluid is important. The theories which apply to compressible, inviscid fluids may be used almost without modification in some low-subsonic problems, and in other cases the results offered by these theories may be modified to account for the effects of viscosity and

2142 Substitution reaction

compressibility. See Compressible Flow; Transonic Flight; Viscosity.

A typical subsonic wing cross section (airfoil) has a rounded front portion (leading edge) and a sharp rear portion (trailing edge). Air approaching the leading edge comes to rest at some point on the leading edge, with flow above this point proceeding around the upper airfoil surface to the trailing edge, and flow below passing along the lower surface to the same point, where the flow again theoretically has zero velocity. The two points of zero local velocity are known as stagnation points. If the path from front to rear stagnation point is longer along the upper surface than along the lower surface, the mean velocity of flow along the upper surface must be greater than that along the lower surface. Thus, in accordance with the principle of conservation of energy, the mean static pressure must be less on the upper surface than on the lower surface. This pressure difference, applied to the surface area with proper regard to force direction, gives a net lifting force. Lift is defined as a force perpendicular to the direction of fluid flow relative to the body, or more clearly, perpendicular to the free-stream velocity vector. See AIRFOIL; BERNOULLI'S THEOREM

The wing, as the lifting device, whether fixed, as in the airplane, or rotating, as in the helicopter, is probably the most important aerodynamic part of an aircraft. However, stability and control characteristics of the subsonic airplane depend on the complete structure. Control is the ability of the airplane to rotate about any of the three mutually perpendicular axes meeting at its center of gravity. Static stability is the tendency of the airplane to return to its original flight attitude when disturbed by a moment about any of the axes. [J.E.Ma.]

Substitution reaction One of a class of chemical reactions in which one atom or group (of atoms) replaces another atom or group in the structure of a molecule or ion. Usually, the new group takes the same structural position that was occupied by the group replaced.

Substitution reactions involve the attack of a reagent, which is the source of the new atom or group, on the substrate, the molecule or ion in which the replacement occurs. They involve the formation of a new bond and the breaking of an old bond. Substitution reactions are classified according to the nature of the reagent (electrophilic, nucleophilic, or radical) and according to the nature of the site of substitution (saturated carbon atom or aromatic carbon atom). See ELECTROPHLIC AND NUCLEOPHILIC REAGENTS.

Systematic names for substitution reactions are composed of the parts: name of group introduced + de + name of group replaced + ation, with suitable elision or change of vowels for euphony. Thus, the replacement of bromine by a methoxy group is called methoxydebromination. See ORGANIC REACTION MECHA-NISM. [J.F.B.]

Subsynchronous resonance The resonance between a series-capacitor-compensated electric system and the mechanical spring-mass system of a turbine-generator at subsynchronous frequencies. Beginning about 1950, series capacitors were installed in long alternating-current transmission lines [250 km (150 mi) or more] to cancel part of the inherent inductive reactance of the line. Until 1971, up to 70% of the 60-Hz inductive reactance was canceled by series capacitors in some long lines with little concern for side effects. (If 70% of a line's inductive reactance is canceled, the line is said to have 70% series compensation.) In 1970, and again in 1971, a turbine-generator at the Mohave Power Plant in southern Nevada experienced shaft damage that required several months of repairs on each occasion, following switching events that placed the turbine-generator radial on a series-compensated transmission line. The shaft damage was due to torsional oscillations between the two ends of the generator-exciter shaft. Shortly after the second event, it was determined that the torsional oscillations were caused by torsional interaction, which is a type of subsynchronous resonance. There have been no reported occurrences of two other types of subsynchronous resonance, the induction generator effect and torque amplification. *See* ALTERNATING-CURRENT CIRCUIT THEORY; ALTERNATING-CURRENT GENERATOR; TRANSMISSION LINES; TURBINE.

It has been clearly established that subsynchronous resonance can be controlled with the use of countermeasures, thus making it possible to benefit from the distinct advantages of series capacitors. About a dozen countermeasures have been successfully applied such that there has been no reported subsynchronous resonance event since 1971. Subsynchronous resonance countermeasures protect turbine-generator shafts from harmful torsional oscillations by one of two methods. First, the turbinegenerator can be tripped when a subsynchronous resonance condition is detected. This limits the number of torsional oscillations experienced by the turbine-generator shafts. This type of countermeasure is relatively inexpensive but is not acceptable if the anticipated subsynchronous resonance conditions are expected to occur frequently. Generally, such a countermeasure will not be applied as the sole subsynchronous resonance protection if it is expected to cause a turbine-generator to be tripped more than once every 10 years. Three types of tripping countermeasures are applied: torsional motion relay, armature current relay, and the generator tripping logic scheme. The second method of protection does not involve turbine-generator tripping, but eliminates or limits harmful torsional oscillations. See ELECTRIC POWER SYSTEMS. [R G F]

Subtraction One of the four fundamental operations of arithmetic and algebra. Subtraction is often regarded as an operation inverse to addition, that is, if *a* and *b* are numbers, the number a - b is defined as that number which added to *b* gives *a*. The more modern viewpoint eliminates subtraction completely by considering the number a - b as the sum of *a* and that number (denoted by -b) which added to *b* gives 0. The number symbolized by -b is called the inverse of *b* (with respect to addition). Every real number has a unique inverse (the number 0 is its own inverse). In this sense "subtraction" may be performed on objects of many different kinds, and the original numerical operation greatly extended. *See* ADDITION; ALGEBRA; DIVISION; MULTIPLICATION; NUMBER THEORY. [L.M.B.]

Subway engineering The branch of transportation engineering that deals with feasibility study, planning, design, construction, and operation of subway (underground railway) systems. In addition to providing rapid and comfortable service, subways consume less energy per passenger carried in comparison with other modes of transportation such as automobiles and buses. They have been adopted in many cities as a primary mode of transportation to reduce traffic congestion and air pollution.

Subways are designed for short trips with frequent stops, compared to above-ground, intercity railways. Many factors considered in the planning process of subway systems are quite similar to those for railway systems. Subway system planning starts with a corridor study, which includes a forecast of ridership and revenues, an estimation of construction and operational costs, and a projection of the potential benefits from land development. *See* RAILROAD ENGINEERING; TRANSPORTATION ENGINEERING.

All subway systems have three major types of structures: stations, tunnels, and depots. The most important task in planning a new subway system or a new subway line is to locate stations and depots and to determine the track alignment. Subway lines are normally located within the right-of-way of public roads and as far away as possible from private properties and sites of importance. Because stations and entrances are usually located in densely populated areas, land acquisition is often a major problem. One solution is to integrate entrances into nearby developments such as parks, department stores, and public buildings, which lessens the visual impact of the entrances and reduces their impediment to pedestrian flow.

Design of the permanent works includes structural and architectural elements and electrical and mechanical facilities. There are two types of structures: stations and tunnels. For stations, space optimization and passenger flow are important. The major elements in a typical station are rails, platform, staircases, and escalators. For handicapped passengers, provisions should be made for the movement of wheelchairs in elevators and at fare gates, and special tiles should be available to guide the blind to platforms.

In both stations and tunnels, ventilation is essential for the comfort of the passengers and for removing smoke during a fire. Sufficient staircases are required for passengers to escape from the station platform to a point of safety in case of a fire. The electrical and mechanical facilities include the rolling stock, signaling, communication, power supply, automated fare collection, and environmental control (air-conditioning) systems. Corrosion has caused problems to structures in some subways; therefore, corrosion-resistant coatings may be required. To minimize noise and vibration from running trains, floating slabs can be used under rails or building foundations in sections of routes crossing densely populated areas and in commercial districts where vibration and secondary airborne noise inside buildings are unacceptable. *See* VENTILATION.

Underground stations are normally constructed by using an open-cut method. For open cuts in soft ground, the sides of the pits are normally retained by wall members and braced using struts. The pits are fitted with decks for maintaining traffic at the surface. For new lines that pass under existing lines, it is not possible to have open cuts. In such cases, stations have to be constructed using mining methods (underground excavation). *See* TUNNEL.

Many modern subway systems are fully automated and require only a minimal staff. Train movements are monitored and regulated by computers in a control center. Therefore, engineering is limited to the function and maintenance of the electrical and mechanical facilities. The electrical and mechanical devices requiring constant care include the rolling stock, signaling, communication and broadcasting systems, power supply, elevators and escalators, automated fare collection, and environmental control systems. Also included are depot facilities, and station and tunnel service facilities. *See* ELECTRIC DISTRIBUTION SYSTEMS; RAILROAD CONTROL SYSTEMS. [Z.C.M.; R.N.Hw.]

Sucrose An oligosaccharide, α -D-glucopyranosyl- β -D-fructofuranoside, also known as saccharose, cane sugar, or beet sugar. The structure is shown below. Sucrose is very soluble in



water and crystallizes from the medium in the anhydrous form. The sugar occurs universally throughout the plant kingdom in fruits, seeds, flowers, and roots of plants. Honey consists principally of sucrose and its hydrolysis products. Sugarcane and sugarbeets are the chief sources for the preparation of sucrose on a large scale. Another source of commercial interest is the sap of maple trees. *See* OLIGOSACCHARIDE; SUGARBEET; SUGARCANE. [W.Z.H.]

Suctoria A small specialized subclass of the protozoan class Ciliatea whose members were long considered entirely separate from the "true" ciliates. The sole order of this subclass is Suctorida. These forms show a number of highly specialized features. Most conspicuous are their tentacles, often numerous, which serve as mouths. These multiple organelles of ingestion



Endogenous budding in the suctorian *Podophrya*, a species which measure 10–28 micrometers.

fasten to the pellicle of prey organisms, generally passing ciliates. By forces not entirely understood, the tentacles are used to suck out the prey's protoplasm to provide sustenance for the suctorian. Nearly all species are stalked, and the sedentary, mature forms are devoid of any external ciliature. Young larval forms are produced by both endogenous and exogenous budding. These forms bear locomotor cilia and serve, as in the case of species of the Peritrichia, for dissemination (see illustration). See CILIATEA; PERITRICHIA. [J.O.C.]

Sudangrass An annual, warm-season grass of tropical origin said to have been grown in Egypt since early times, though its value was first recognized in Sudan only in 1909. In that same year it was introduced into the United States as a replacement for johnsongrass, which had become a noxious weed in many southern states. Sudangrass (*Sorghum bicolor var. sudanense*, also called *S. sudanense* and *S. vulgare var. sudanense*) and its hybrids are commonly used as pasture, greenchop, silage, or hay. They fill an important need in many regions of the United States, because they produce high-quality forage for cattle and sheep during the summer, when other pasture is in short supply or of low quality. Many of the varieties and hybrids produce forage until frost. *See* GRASS. [M.R.G.]

Sudden infant death syndrome (SIDS) The sudden and unexpected death of an apparently normal infant that remains unexplained after an adequate autopsy. Of a group of apparently healthy infants dying suddenly and unexpectedly, 15% will usually manifest pathologic evidence of a disease process which is sufficient to explain the death. The remaining 85% are unexplained and are classified as SIDS. In spite of the probable heterogeneity of diseases in SIDS cases, the consistent and distinctive characteristics of these infant deaths support the notion that many, if not the majority, represent a single disease process.

The incidence of SIDS in the United States is about 2.0 cases per 1000 live births, which makes SIDS the leading cause of death between the ages of 1 month and 1 year. Most SIDS deaths occur at 2–4 months of age, and about 90% occur by 6 months. SIDS is more common in males, prematurely born infants, multiple births, and the economically disadvantaged. SIDS is also increased in infants of teen-age or smoking mothers and in infants who have a history of a severe apparent life-threatening event, usually accompanied by marked cyanosis or pallor and limpness, and absence of breathing. SIDS also occurs more frequently during winter months. The rate among Native Americans is greater than among Blacks, which is greater than among Caucasians; Asians have the lowest rate. While there is slight familial clustering of SIDS, there is probably not a genetic predisposition to SIDS.

2144 Sugar

The cause of SIDS is unknown; leading hypotheses include respiratory, cardiac, and metabolic mechanisms. Much attention has been focused on the "apnea hypothesis," implicating a primary respiratory arrest due to chronic or transient insufficiency or irregularity of breathing. An imbalance between sympathetic and parasympathetic influences on cardiac activity, leading to potentially fatal cardiac arrhythmias, is a popular cardiac hypothesis.

While there is still no proof that SIDS can be prevented, electronic cardiorespiratory monitors have been prescribed for many infants in high-risk categories for SIDS. Home monitors are recommended only for infants at very high risk for SIDS. See CON-GENITAL ANOMALIES; HUMAN GENETICS. [A.St.; K.Wi.; J.G.Br.]

Sugar Usually, sucrose, the common sugar of commerce. This sugar is a disaccharide, $C_{12}H_{22}O_{11}$, which is split, as shown in the reaction below, by hydrolysis into two monosaccharides,

or simple sugars: glucose (dextrose) and fructose (levulose). Sucrose rotates the plane of polarized light to the right, as does glucose, but fructose is so strongly levorotatory that it overcomes the effect of glucose. Thus mixtures of equal amounts of glucose and fructose are levorotatory. The hydrolytic reaction is called inversion of sugar, and the product is invert sugar or, simply, invert. See CARBOHYDRATE; OPTICAL ACTIVITY.

Sucrose is widely distributed in nature, having been found in all green plants that have been carefully examined for its presence. The total quantity of all sugars formed each year has been estimated at a colossal 4×10^{11} tons (3.6×10^{11} metric tons). In spite of its availability in all green plants, sucrose is obtained commercially in substantial amounts from only two plants: sugarcane, which supplies about 56% of the world total, and the sugarbeet, which provides 44%.

Cane sugar manufacture. The manufacture of cane sugar is usually done in two series of operations. First, raw sugar of about 98% purity is produced at a location adjacent to the cane fields. The raw sugar is then shipped to refineries, where a purity that is close to 100% is achieved.

Raw cane sugar. The production of raw cane sugar begins with growing the cane in tropical or subtropical areas. The cane is harvested after a season varying from 7 months in subtropical areas to 12–22 months in the tropics. The cane stalks are harvested either mechanically or by heavy hand knives. The trend is toward mechanization. The stalks are transported to a mill by oxcart, rail, or truck. *See* SUGARCANE.

At the mill the stalks are crushed and macerated between heavy grooved iron rollers while being sprayed countercurrently with water to dilute the residual juice. The expressed juice contains 95% or more of the sucrose present. The fibrous residue, or bagasse, is usually burned under the boilers, although increasing amounts are being made into paper, insulating board, and hardboard as well as furfural, which is a chemical intermediate for the synthesis of furan and tetrahydrofuran.

Cane juice. The cane juice is treated with lime to bring its pH to about 8.2. This pH prevents the inversion reaction, which is favored by heat and acid and would lower the yield of crystallizable sugar. The juice is then heated to facilitate the precipitation of impurities, which are removed by continuous filtration.

The purified juice is concentrated by multiple-stage vacuum evaporation (usually four or five stages) and, when sufficiently concentrated, is boiled to grain or seeded with sucrose crystals in a single-stage vacuum pan. Usually three successive crops of crystals are grown, cooled, and centrifuged. The final mother liquor, which is resistant to further crystallization, is called blackstrap molasses. It is used principally as a feed for cattle. Relatively small amounts are still fermented to produce industrial alcohol and rum. *Raw cane sugar refining.* The refining of raw sugar begins with dissolution of the molasses which remains in a thin film on the crystals in spite of the centrifugation. This step, called affination, brings the purity from about 98 to about 99%. The crystals are dissolved in hot water and percolated through bone char columns to remove color by adsorption. The sucrose is finally concentrated by vacuum evaporation, crystallized by seeding, centrifuged, and dried.

A major step forward has been the use of bone char in a continuous countercurrent manner (Grosvenor patent). The bone char is washed, dried, and burned to remove impurities; it is reused until it wears out mechanically and is discarded as fines. Even the fines have value as fertilizer because of their high calcium phosphate content.

Bone char has been replaced in some refineries by granular carbon derived from coal or wood waste. It is used in columns and regenerated by a process similar to the one used for bone char, which unlike granular carbon is primarily calcium phosphate. Granular carbon is frequently used in combination with decolorizing resins to reduce the color of process liquors coming from the carbon columns. Pulverized activated vegetable carbon is used in a few instances to remove color from raw sugar liquors, but its use is declining. Ion-exchange resins are used to remove ash from raw sugar liquors, especially when the liquor is to be converted directly into refined liquid sugar without going through the crystallization step first. See ION EXCHANGE.

Beet sugar. In the United States, sugarbeets are grown under contract by farmers from seed supplied by a beet sugar company. Because sugarbeets, like the other temperate-zone crops, thrive best under crop rotation, they are not well adapted to one-crop agriculture. *See* SUGARBEET.

Beet processing. When beets are delivered to a factory, they are washed and sliced, and the slices are extracted countercurrently with hot water to remove the sucrose. The resulting solution is purified by repeatedly precipitating calcium carbonate, calcium sulfite, or both, in it. Colloidal impurities are entangled in the growing crystals of precipitate and removed by continuous filtration. The resulting solution is nearly colorless, and the sucrose is concentrated by multiple-effect vacuum evaporation. The syrup is seeded, cooled, and centrifuged, and the beet sugar crystals are washed with water and dried.

Beet molasses differs from cane molasses in having a much lower content of invert sugar. It is, therefore, relatively stable to the action of alkali and, in the United States, is usually treated with calcium oxide to yield a precipitate of calcium sucrate. This is a mixture of loose chemical aggregates of sucrose and calcium oxide which are relatively insoluble in water. The precipitate is filtered, washed, and added to the incoming crude sugar syrup, where it furnishes calcium for the precipitations of calcium carbonate and sulfite referred to above, which remove impurities. Carbon dioxide in the form of flue gas is the other reagent, and sulfur dioxide from burning sulfur is used to produce sulfite. Ion exclusion processes have been developed for recovery of sucrose from beet molasses.

Beet residue use. The beet tops and extracted slices, as well as the molasses, are valuable as feeds. More feed for cattle and other ruminants can be produced per acre-year from beets than from any other crop widely grown in the United States. This is independent of the food energy in the crystallized sucrose, which exceeds that available from any other temperate-zone plant. It is for these reasons that the densely populated countries of Europe have expanded their beet sugar production, in spite of the ready availability of cane sugar from the tropics.

The increased use of nitrogenous fertilizer has resulted in augmenting the protein content of the molasses and other beet byproducts.

Nutritional value. Sucrose has, in the past, been attacked by some nutritionists on the ground that it provides only "empty calories," without protein, minerals, or vitamins. This argument lost much of its force when it was shown that all the vitamins and minerals recommended by the National Research Council can be obtained by consuming any of a great variety of foods in amounts that yield a total of only one-half of the caloric requirements of an average person. The wide use by the public of vitamin supplements has caused some nutritionists to express an opposite worry over excessive vitamin consumption.

In addition to the charges that it provides empty calories, sugar has been blamed for such health problems as diabetes mellitus, coronary heart disease, dental caries, and obesity. These charges and the scientific evidence associated with them have been extensively reviewed by a special committee of the Federation of American Societies for Experimental Biology. A report of the committee findings, published by the Food and Drug Administration in 1976, includes the following: Sucrose is a contributor to the formation of dental caries when used at levels and in the manner prevailing in 1976; otherwise, there is no clear evidence in the available information that demonstrates a hazard to the public when sucrose is used at those levels.

Other sources of sugars. In Hawaii, the pineapple industry recovers both sucrose and citric acid from the rinds. The residue is fed to cattle. The total quantity of sucrose obtained from waste fruits is statistically negligible. On the other hand, a substantial fraction of the total sucrose consumed is that naturally present in a large number of fruits, vegetables, and nuts. *See* SUCROSE; SUGAR CROPS.

Lactose. Cow's milk, on a dry basis, is about 38% lactose or milk sugar. When milk is converted into cheese, the lactose remains in the whey, from which it may easily be isolated and purified. Lactose is a disaccharide which is split by hydrolysis into glucose and galactose. Lactose is about one-tenth as soluble in water as sucrose and it is one-sixth to one-half as sweet, depending on the concentration. Some of the major uses for lactose are as a direct-compression vehicle for the manufacture of pharmaceutical tablets and as a diluent for pharmaceuticals and synthetic sweeteners. *See* CHEESE; LACTOSE.

Starch. Starches can be hydrolyzed either by dilute acid or by enzymes. The product of acid hydrolysis varies with time and conditions but contains glucose, maltose, maltotriose, maltotetrose, and other sugars up to the dextrins. Only glucose, a monosaccharide, is readily isolated. Crystallized as the monohydrate, it is used in foods as a sweetener. *See* CORN; GLUCOSE.

Syrups high in maltose, a disaccharide, can be obtained by the action of amylases on starch. This hydrolysis has been of great importance for thousands of years in splitting starches for alcoholic fermentation. As yet, there is no large-scale production of pure maltose. See MALTOSE.

A major type of syrup was developed from starch hydrolyzates by the mid-1980s. An enzyme that isomerizes dextrose (glucose) was commercially developed to manufacture hydrolyzed starch syrups (corn syrups) containing varying amounts of fructose (levulose). The enzyme isomerizes dextrose into fructose theoretically to a 50-50 mixture of dextrose and fructose. In practice, the isomerization is allowed to progress until a syrup containing (on a solids basis) approximately 50% dextrose, 42% fructose, and 8% higher saccharide is achieved. Syrups containing more than 42% fructose (solid basis) can be made by subjecting the 42% fructose syrup to an ion-exclusion process that separates dextrose and fructose. The fructose-rich fraction from the separation is converted into a finished product, and the dextrose-rich fraction is returned to the isomerization process. Commercially produced syrups contain 42% or 55% fructose. The combined processes are also used to manufacture commercial fructose. The high-fructose corn syrups are competitive with sucrose and invert sugar in a variety of food products. The most widely used application for 55% high-fructose corn syrup is in soft drinks, displacing sucrose and invert sugar. See FRUCTOSE; STARCH.

Maple sugar. Before America was settled by Europeans, Indians were collecting and concentrating the juice of the hard maple (*Acer saccharum*), making maple syrup. The practice was copied by the new settlers, and production of maple syrup has been an industry ever since in the regions where hard maples are common, principally the northeastern United States.

The maple flavor does not exist in the sap but is developed by heating it. By additional heating at about $250^{\circ}F$ ($120^{\circ}C$), a flavor four or five times more intense can be developed. Maple syrup so produced is of special value for adding flavor to the less expensive products of the sucrose industry. Maple sugar is sucrose of about 95–98% purity; the characteristic flavor makes up only a small percent. Fairly satisfactory imitation maple flavors are available.

Honey. Honey is a form of relatively pure invert sugar dissolved in water to form a concentrated solution. However, honey also contains flavors derived from the nectar of the flowers from which it was obtained by the bee. Nutritionally, it is nearly equivalent in invert sugar but contains an excess of fructose over glucose. The sucrose found in the flowers is inverted by the enzyme honey invertase. Tupelo honey is remarkable in containing about twice as much fructose as glucose, and hence has little tendency to deposit glucose crystals.

The ready availability of the food energy in honey was known to athletes in ancient Greece. Only in recent times has it been discovered that, paradoxically, the energy of sucrose is still more quickly available. The flavors of various honeys run a wide gamut. That from the Mount Hymettus region, which is flavored by wild thyme, has been known and treasured since the poems of Homer.

Molasses. Virtually all molasses is distributed in the form of a concentrated viscous solution, but it can be reduced to a powder by means of spray-drying. It can then be handled without an investment by the consumer in tanks, pipes, and pumps. Unfortunately, later contact with moisture converts the dried molasses to a gummy mass. The availability of vaporproof bags (for example, those lined with polyethylene) has provided one solution to this problem. There are also various additives which, when mixed with molasses, reduce its tendency to pick up moisture. So far, dried molasses has made little headway against the practice of handling concentrated solutions.

Syrups. Syrups are relatively concentrated, somewhat viscous, solutions of various sugars, frequently in admixture to hinder crystallization. Large amounts of sucrose have been distributed in the form of high-purity syrups. These syrups consist of sucrose, invert sugar, or both. In some instances syrups include mixtures with dextrose and various corn syrups. The so-called liquid sugars are important for several reasons: the requirement for the use of syrups in a number of food processes (conversion of granular sucrose to a syrup was eliminated at the point of use); economy of handling, being moved with pumps, pipes, tank trucks, and tank cars; the high degree of sanitation inherent in the storage and distribution of sugar within the plant in a closed system, including the elimination of dust from the dumping of bags; and the availability of a combination of sugars with differing characteristics in a single product.

The quantity of sucrose distributed in syrup form has been reduced drastically because of the inroads made by high-fructose corn syrups, which have a composition somewhat comparable to that of invert sugar. These products are available only as syrups. They do not have taste characteristics identical to sucrose and invert sugar syrups, but lower prices catalyze their substitution for sucrose.

It is the aim of the sugar refiner and the beet processor to eliminate color and all flavors other than sweetness. The manufacture of table syrups, which are widely used on waffles and pancakes, aims at a broader spectrum of flavor. Corn syrup, which is somewhat lacking in sweetness, ordinarily has about 15% sucrose added. The high viscosity of the corn syrup, resulting from the content of dextrins, tends to hinder crystallization and is an advantage in the manufacture of certain candies.

Some sugar refiners manufacture so-called refiners' syrups for use in the manufacture of foods and table syrups. In addition to these syrups, a group of products generally referred to as cane

2146 Sugarbeet

syrup and edible molasses are manufactured from cane juice. Cane syrup is concentrated whole cane juice, while edible molasses is concentrated cane juice from which some sugar has been removed by crystallization. These syrups and molasses contain both sucrose and invert sugar, and have a relatively dark color and a distinctive flavor. They also contain many of the nonsugars found in sugarcane juice. They are used principally by food processors, but substantial amounts are packaged for home use. Sorghum syrup is made by extracting juice from the stalk of sweet sorghum and evaporating the juice to syrup consistency. It is a dark, pungently flavored product that is not widely distributed. *See* FOOD ENGINEERING; FOOD MANUFACTURING; SUGAR CROPS. [H.B.H; Ch.B.B.]

Sugarbeet The plant *Beta vulgaris*, developed in modern times to fill the need for a sugar crop that could be grown in temperate climates. The sugarbeet was the source of only 5% of the world's commercial sugar in 1840, but by 1890 it supplied almost 50%, and since about 1920 this has dropped to about 40% with 60% derived from sugarcane. European countries produce most of their sugar from sugarbeet, with some countries having surplus sugar for export. *See* SUGARCANE.

The sugarbeet, fodder beet, garden beet, and leaf beet (Swiss chard) are cultivars of *B. vulgaris* and are genetically related. In addition to *B. vulgaris*, which includes all cultivated forms of beet, there are 12 species of wild beet in the Mediterranean and Middle Eastern regions. These wild beets are of interest in sugarbeet improvement as sources of disease resistance. Most varieties of sugarbeet in the United States are hybrid. [D.Ste.]

Sugarcane Saccharum officinarum, a member of the grass family. This crop originated in New Guinea about 15,000– 8000 B.C., and was later moved by primitive peoples westward into Southeast Asia and India and eastward into Polynesia. Most current commercial varieties are interspecific hybrids involving primarily two or more of the following species: *S. officinarum*, *S. robustum*, and *S. spontaneum*.

Sugar is generally removed from sugarcane in large factories by either milling or diffusion. Milling crushes the stalks as they pass between a series of large metal rollers separating the fiber (bagasse) from the juice laden with sugar. The diffusion process separates sugar from finely cut stalks by dissolving it in hot water or hot juice. Processing of the juice is completed by clarification to remove nonsugar components, by evaporation to remove water, and by removal of molasses in high-speed centrifuges to produce centrifugal sugar. Bagasse and molasses are the principal byproducts from processing sugarcane. [R.E.Co.]

Sulfide phase equilibria The chemistry of the sulfides is rather simple inasmuch as most sulfide minerals involve only two major elements, although some involve three, and a few four. Understanding of the phase relations among these minerals is important both to the geologist, whose task it is to locate and exploit ore deposits, and to the metallurgist, whose task it is to extract the metals from the ores for industrial use.

Sulfide ore deposits are the most important sources of numerous metals such as lead, zinc, copper, nickel, cobalt, and molybdenum. In addition, sulfide ores provide substantial amounts of noble metals such as platinum, gold, and silver, and of other industrially important elements such as cadmium, rhenium, and selenium. Although iron sulfides usually are the most common minerals in such deposits, most commercial iron is mined from iron oxide ores and most sulfur from elemental sulfur deposits.

Some of the more common sulfide minerals are listed in the table. The first eight sulfide minerals listed may be plotted in the ternary system copper-iron-sulfur. Similarly, the first two and the last two minerals may be plotted in the ternary system ironnickel-sulfur. Thus, by studying in detail the phase equilibria in two ternary systems a great deal of information can be obtained about many of the common sulfides occurring in ore deposits.

common sulfide minerals				
Mineral name	Chemical formula			
Pyrite Pyrrhotites (three or more varieties) Covellite Digenite Chalcocite Bornite Chalcopyrite Cubanite Galena Sphalerite and wurtzite Metacinnabar and cinnabar Argentite and acanthite Molybdenite Millerite Pentlandite	$\begin{array}{c} FeS_2\\ Fe_{1-x}S\\ CuS\\ Cu_9S_5\\ Cu_2S\\ Cu_5FeS_4\\ CuFeS_2\\ CuFe_2S_3\\ PbS\\ ZnS\\ HgS\\ Ag_2S\\ MoS_2\\ NiS\\ (Fe,Ni)_9S_8 \end{array}$			

However, before the ternary systems can be explored in a systematic way, the binary systems bounding the ternaries have to be studied in detail. Similarly, it is necessary that the four bounding ternary systems be fully understood before a quaternary system can be systematically investigated. Thus, it is seen that before a systematic study of, for example, the immensely important quaternary system copper-iron-nickel-sulfur can proceed, much preliminary information is required. The prerequisite data include complete knowledge of the four ternary systems Cu-Fe-S, Cu-Fe-Ni, Fe-Ni-S, and Cu-Ni-S. In turn, the phase relations in these ternary systems cannot be systematically studied before the six binary systems Fe-Cu, Fe-Ni, Cu-Ni, Fe-S, Cu-S, and Ni-S have been thoroughly explored. See PHASE EQUILIBRIUM.

The enormous differences in the vapor pressures over the different phases occurring in sulfide systems add complications to the diagrammatic representation. For instance, in the Fe-S system the vapor pressure over pure iron is about 10^{-25} atm $(10^{-20} \text{ pascal})$ at 450°C (840°F), whereas that over pure sulfur is a little more than 1 atm (10^5 pascals) at the same temperature. A complete diagrammatic representation of the relations in such a system, therefore, requires coordinates for composition and temperature, as well as for pressure. In a two-component system, such as the Fe-S system, such a representation is feasible because only three coordinates are necessary. However, in ternary (where such diagrams involve four-dimensional space) and in multicomponent systems, this type of diagrammatic representation is not possible. For this reason it is customary to use composition and temperature coordinates only for the diagrammatic representation of sulfide systems. The relations as shown in such diagrams in reality represent a projection from composition-temperature-pressure space onto a two-dimensional composition-temperature plane or onto a three-dimensional prism, depending upon whether the system contains two or three components.

Pyrite or pyrrhotite, or both, occur almost ubiquitously, not only in ore deposits but in nearly all kinds of rocks. Of the binary systems mentioned above, therefore, the iron-sulfur system is of the most importance to the economic geologist. See PyrITE. [G.Kul.]

Sulfonamide One of the group of organosulfur compounds, RSO_2NH_2 . Many sulfonamides, which are the amides of sulfonic acids, have the marked ability to halt the growth of bacteria. The therapeutic drugs of this group are known as sulfa drugs. *See* ORGANOSULFUR COMPOUND.

The antibacterial spectrum of sulfonamides comprises a wide variety of gram-positive and gram-negative bacteria, including staphylococci, streptococci, meningococci, and gonococci, as well as the gangrene, tetanus, coli, dysentery, and cholera bacilli. They have only slight activity against *Mycobacterium tuberculosis*, while certain closely related sulfones are quite active against *M. leprae.* The use of sulfonmethoxine has proved to be effective in the treatment of chloroquine-resistant malaria. The relative potency of sulfonamides against the different microorganisms varies, and their action is bacteriostatic rather than bactericidal.

The antibacterial effect, both in patients and in test tube cultures, is antagonized by *p*-aminobenzoic acid (PABA) and PABAcontaining natural or synthetic products, such as folic acid and procaine. Accordingly, the mode of action of sulfonamides is considered to be an antimetabolite activity, dependent upon the inhibition of enzyme systems involving the essential PABA. *See* BACILLARY DYSENTERY; CHOLERA; LEPROSY; MALARIA; MENINGO-COCCUS; STAPHYLOCOCCUS; STREPTOCOCCUS; TETANUS; TUBERCU-LOSIS.

Bacterial resistance has developed to all known sulfonamides, and many sulfonamide-resistant strains are encountered among the gram-positive and gram-negative bacteria. The emergence of resistance to sulfonamides, however, seems less rapid and less widespread than resistance to most antibiotics.

Sulfonamides are used today mostly as auxiliary drugs or in combination with antibiotics. In certain infectious diseases, however (for instance, in meningococcal infections and most infections of the urinary tract), sulfonamides deserve preference over antibiotics. See ANTIMICROBIAL AGENTS; DRUG RESISTANCE. [R.L.M.; E.Gru.]

Sulfonation and sulfation Sulfonation is a chemical reaction in which a sulfonic acid group, —SO₃H, is introduced into the structure of a molecule or ion in place of a hydrogen atom. Sulfation involves the attachment of the —OSO₂OH group to carbon, yielding an acid sulfate, ROSO₂OH, or of the —SO₄—group between two carbons, forming the sulfate, ROSO₂OR.

Sulfonation of aromatic compounds is the most important type of sulfonation. This is accomplished by treating the aromatic compound with sulfuric acid, as in the reaction below. The product of sulfonation is a sulfonic acid.



Sulfonation may also be defined as any chemical process by which the sulfonic acid group, — SO_2OH , or the corresponding salt or sulfonyl halide group, for example, — SO_2CI , is introduced into an organic compound. These groups may be bonded to either a carbon or a nitrogen atom. The latter compounds are designated *N*-sulfonates or sulfamates.

Most sulfonates are employed as such in acid or salt form for applications where the strongly polar hydrophilic —SO₂OH group confers needed properties on a comparatively hydrophobic nonpolar organic molecule. Some sulfonates, such as methanesulfonic and toluenesulfonic acids, are used as catalysts. The major quantity of sulfonates and sulfates is both marketed and used in salt form. This category includes detergents; emulsifying, demulsifying, wetting, and solubilizing agents; lubricant additives; and rust inhibitors. *See* ORGANOSULFUR COM-POUND; SUBSTITUTION REACTION. [P.H.G./R.S.K.]

Sulfonic acid A derivative of sulfuric acid (HOSO₂OH) in which an OH has been replaced by a carbon group, as shown in the structure below. Sulfonic acids are strongly acidic,



water-soluble, nonvolatile, and hygroscopic; they do not act as oxidizing agents and are typically highly stable compounds. Sulfonic acids rarely occur naturally. An exception, taurine, $NH_2CH_2CH_2SO_3H$, occurs in bile.

The aliphatic sulfonic acids are generally made by oxidation of thiols. Several have unique properties. For example, trifluoromethanesulfonic acid, CF_3SO_3H , is such a strong acid that it will protonate sulfuric acid. A compound derived from natural camphor, 10-camphorsulfonic acid, is used extensively in the optical resolution of amines.

Aromatic sulfonic acids are much more important than those of the aliphatic series. Aromatic sulfonic acids are produced by sulfonation of aromatic compounds with sulfuric acid or fuming sulfuric acid. Sulfonation of aromatic hydrocarbons is a reversible process; treatment of an aromatic sulfonic acid with superheated steam removes the —SO₃H group. This process can be used in purifying aromatic hydrocarbons. Aromatic sulfonic acids and their derivatives, especially metal salts, are important industrial chemicals. See SULFURIC ACID.

The most extensive use of the sulfonation reaction is in the production of detergents. The most widely used synthetic detergents are sodium salts of straight-chain alkylbenzenesulfonic acids. *See* DETERGENT.

Sulfonated polymers, particularly sulfonated polystyrenes, act as ion-exchange resins which have important applications in water softening, ion-exchange chromatography, and metal separation technology. Both sulfonated polymers and simple aromatic sulfonic acids, particularly *p*-toluenesulfonic acid, are frequently used as acid catalysts in organic reactions such as esterification and hydrolysis. *See* HOMOGENEOUS CATALYSIS; ION EXCHANGE.

The sulfonic group, in either acid or salt form, is capable of making many substances water soluble, increasing their usefulness. This application is particularly significant in the dying industry, in which a majority of the dyes are complex sodium sulfonates. Many acid-base indicators are soluble due to the presence of a sodium sulfonate moiety. Some pigments used in the paint and ink industry are insoluble metal salts or complexes of sulfonic acid derivatives. Most of the brighteners used in detergents compounded for laundering are sulfonic acid derivatives of heterocyclic compounds. *See* ACID-BASE INDICATOR; DETERGENT; DYE; INK; ORGANOSULFUR COMPOUND; PAINT. [C.R.J.]

Sulfur A chemical element, S, atomic number 16, and atomic weight 32.064. The atomic weight reflects the fact that sulfur is composed of the isotopes 32 S (95.1%), 33 S (0.74%), 34 S (4.2%), and 36 S (0.016%). The ratios of the various isotopes vary slightly but measurably according to the history of the sample. By virtue of its position in the periodic table, sulfur is classified as a main-group element. See PERIODIC TABLE.

The chemistry of sulfur is more complex than that of any other elemental substance, because sulfur itself exists in the largest variety of structural forms. At room temperature, all the stable forms of sulfur are molecular; that is, the individual atoms aggregate into discrete molecules, which in turn pack together to form the solid material. In contrast, other elements near sulfur in the periodic table normally exist as polymers (silicon, phosphorus, arsenic, selenium, tellurium) or as diatomic molecules (oxygen, nitrogen, chlorine). Selenium and phosphorus can exist as molecular solids, but the stable forms of these elements are polymeric.

At room temperature the most stable form of sulfur is the cyclic molecule S_8 . The molecule adopts a crownlike structure, consisting of two interconnected layers of four sulfur atoms each. The S—S bond distances are 0.206 nanometer and the S—S—S bond angles are 108°. Three allotropes are known for cyclo- S_8 . The most common form is orthorhombic α -sulfur, which has a density of 2.069 g/cm³ (1.200 oz/in.³) and a hardness of 2.5 on the Mohs scale. It is an excellent electrical insulator, with a room temperature conductivity of 10^{18} ohm⁻¹ cm⁻¹. Sublimed sulfur and "flowers" of sulfur are generally composed of α -S₈. Sulfur is quite soluble in carbon disulfide (CS₂; 35.5/100 g or 1.23 oz/3.52 oz at 25°C or 77°F), poorly soluble in alcohols, and

practically insoluble in water. At 95.3°C (203°F), sulfur changes into the monoclinic β allotrope. This form of sulfur also consists of cyclic S₈ molecules, but it has a slightly lower density at 1.94– 2.01 g/cm³ (1.12–1.16 oz/in.³). A third allotrope containing S₈ is triclinic γ -sulfur. The β and γ allotropes of sulfur slowly revert to the α form at room temperature. Crystals of sulfur are yellow and have an absorption maximum in the ultraviolet at 285 nm, which shifts to higher energy as the temperature decreases. At low temperatures, S₈ is colorless. Even at room temperature, however, finely powdered sulfur can appear to be nearly white.

The best-studied system is α -S₈, which converts to the β form at 90°C (194°F), which then melts at 120° C (248°F) to give a golden yellow liquid. If this melt is quickly recooled, it refreezes at 120° C (248°F), thus indicating that it consists primarily of S₈ molecules. If the melt is maintained longer at 120°C (248°F), then the freezing point is lowered about $5^{\circ}C$ (9°F), indicating the formation of about 5% of other rings and some polymer. At $159.4^{\circ}C$ (318.9°F), the melt suddenly assumes a red-brown color. Over the range 159.4-195°C (318.9-383°F), the viscosity of the melt increases 10,000-fold before gradually decreasing again. This behavior is very unusual, since the viscosity of most liquids decreases with increasing temperature. The strong temperature dependence of the viscosity is due to the polymerization and eventual depolymerization of sulfur. Polymeric sulfur retains its elastomeric character even after being cooled to room temperature. There are several polymeric forms of sulfur, but all of them revert to α -S₈ after a few hours.

Sublimination of S₈ occurs when it is maintained in a vacuum at a temperature below its melting point. It vaporizes at 444.61°C (832.30°F). Below 600°C (1110°F), the predominant species in the gas are S₈ followed by S₇ and S₆. Above 720°C (1328°F), violet S₂ is the major species.

Principal inorganic compounds. Hydrogen sulfide (H₂S) is the most important compound that contains only sulfur and hydrogen. It is a gas at room temperature with a boiling point of $-61.8^{\circ}C$ ($-79.2^{\circ}F$) and a freezing point of $-82.9^{\circ}C$ ($-117^{\circ}F$). The low boiling point of hydrogen sulfide is attributed to the weakness of intermolecular S…H hydrogen bonding; the O…H hydrogen bond is much stronger, as evidenced by the high boiling point of water. Gaseous hydrogen sulfide is 1.19 times more dense than air, and air-H₂S mixtures are explosive. Hydrogen sulfide has a strong odor similar to that of rotten eggs; its odor is detectable at concentrations below 1 microgram/m³. At high concentrations, H₂S has a paralyzing effect on the olfactory system, which is very hazardous because H₂S is even more toxic than carbon monoxide (CO).

The most common compound that contains only carbon and sulfur is carbon disulfide (CS₂). Carbon disulfide molecules are linear, consisting of two sulfur atoms connected to a central carbon atom. Carbon disulfide is a toxic, highly flammable, and volatile liquid that melts at $-111^{\circ}C$ ($-168^{\circ}F$) and boils at $46^{\circ}C$ (115°F). Commercial carbon disulfide has a strong unpleasant odor due to impurities. It is manufactured from methane and elemental sulfur and is used for the production of carbon tetrachloride, rayon, and cellophane. Structurally related to carbon disulfide is carbonyl sulfide (SCO), which forms from carbon monoxide and elemental sulfur. The chlorination of CS₂ gives Cl₃CSCl, which can be reduced by H₂S to thiophosphene, CSCl₂. Thiophosgene (CSCl₂) [boiling point 73°C or 163°F] is a planar molecule with the carbon at the center of a triangle defined by the sulfur and two chlorine atoms. Thiocyanate, the linear anion NCS⁻, is prepared by the reaction of cyanide (--CN) with elemental sulfur.

Several sulfur oxides exist, but the dioxide and trioxide are of preeminent importance. Sulfur dioxide (SO₂) is a colorless gas that boils at -10.02° C (113.97°F) and freezes at -75.46° C (-103.8° F). The density of liquid sulfur dioxide at -10° C (14°F) is 1.46 g/cm³ (0.84 oz/in.³). Liquid sulfur dioxide is an excellent solvent. The sulfur dioxide molecule is bent, with an O—S—O angle of 119°.

Sulfur trioxide (SO₃) is a planar molecule that is a liquid at room temperature that exists in equilibrium with a cyclic trimeric structure known as β -SO₃. When β -SO₃, actually S₃O₉, is treated with traces of water, it converts to either of two polymeric forms referred to as γ - and α -sulfur trioxide. These are fibrous materials, proposed to have the formula (SO₃)_xH₂, where *x* is in the thousands. Sulfur trioxide is prepared by the oxidation of sulfur dioxide, although at very high temperatures this reaction reverses. Exposure of sulfur trioxide to water yields sulfuric acid (H₂SO₄); exposure of SO₃ to sulfuric acid yields disulfuric acid (H₂S₂O₇). See SULFURIC ACID.

Chlorine and sulfur react to give a family of compounds with the general formula S_xCl_2 , several members of which have been obtained in pure form. The structures of these compounds are based on an atom or chain of sulfur atoms terminated with Cl. Sulfur monochloride (S_2Cl_2), also known as sulfur monochloride, is the most widely available of the series. It is a yellow oil that boils at 138°C (280°F), and reacts with chlorine in the presence of iron(III) chloride (FeCl₃) catalyst to give sulfur dichloride (Sc_2), which is a red volatile liquid with a boiling point of 59°C (138°F). Treatment of sulfur dichloride with sodium fluoride (NaF) gives SF_4 .

Thionyl chloride (OSCl₂) is a colorless reactive compound with a boiling point of 76°C (169°F); it is used to convert hydroxy compounds to chlorides. Important applications include the preparation of anhydrous metal halides and alkyl halides. Sulfuryl chloride (O₂SCl₂; boiling point 69°C or 156°F) is used as a source of chlorine.

Organosulfur compounds. This family of compounds contains carbon, hydrogen, and sulfur, and it is a particularly vast area of sulfur chemistry. Thiols, also known as mercaptans, feature the linkage C-S-H. Mercaptans are foul-smelling compounds. They are the sulfur analogs of alcohols, but they are more volatile. They can be prepared by the action of hydrogen sulfide (H₂S) on olefins. Deprotonation of thiols gives thiolate anions, which form stable compounds with many heavy metals. Thiols and especially thiolates can be oxidized to form disulfides (persulfides), which have the connectivity of C-S-S-C. The organic persulfides are also related to organic polysulfides, which have chains of sulfur atoms terminated with carbon. The introduction of such mono-, di-, and polysulfide linkages is the basis of the vulcanization process, which imparts desirable mechanical properties to natural or synthetic polyolefin rubbers. This is accomplished by heating the polymer with sulfur in the presence of a zinc catalyst. See RUBBER.

Thioethers, also known as organic sulfides, feature the connectivity C—S—C and are often prepared from the reaction of thiolates and alkyl halides. Like mercaptans, thioethers often have strong unpleasant odors, but they are also responsible for the pleasant odors of many foods and perfumes. They are intentionally introduced at trace levels in order to impart an odor to gaseous hydrocarbon fuels. The reaction of alkyl dihalides and sodium polysulfides affords organic polysulfide polymers known as thiokols.

There are many organic sulfur oxides; prominent are sulfonic acids (RSO_3H), which are the organic derivatives of sulfuric acid. These compounds are prepared by the oxidation of thiols as well as by treatment of benzene derivatives with sulfuric acid, for example, benzene sulfonic acid. Most detergents are salts of sulfonic acids. *See* DETERGENT.

Biochemistry. Sulfur is required for life. Typical organisms contain 2% sulfur dry weight. Three amino acids contain sulfur, as do many prosthetic groups in enzymes. Some noteworthy sulfur compounds include the disulfide lipoic acid, the thioethers biotin and thiamine (vitamin B_1), and the thiol coenzyme A. Sulfide ions, S^{2-} , are found incorporated in metalloproteins and metalloenzymes such as the ferredoxins, nitrogenases, and hydrogenases. See AMINO ACIDS; ENZYME; PROTEIN.

Many bacterial species obtain energy by the oxidations of sulfides. Bacteria of the genus *Thiobacillus* couple the conversion of carbon dioxide (CO_2) to carbohydrates to the aerobic oxidation of mineral sulfides to sulfuric acid. This activity can be turned to good use for leaching low-grade mineral ores. Often, however, the sulfuric acid runoff (such as in mines or sewers) has negative environmental consequences. The purple and green bacteria as well as the blue-green algae are remarkable because they are photosynthetic but anaerobic; they oxidize sulfide, not water (as do most photosynthetic organisms). Depending on the species, the sulfur produced in this energy-producing pathway can accumulate inside or outside the cell wall. See BACTERIAL PHYSIOLOGY AND METABOLISM; PHOTOSYNTHESIS.

Minerals. Sulfide minerals are among the most important ores for several metals. These compounds are two- or threedimensional polymers containing interconnected metal cations and sulfide S^{2-} or persulfido S_2^{2-} anions. In general, metal sulfides are darkly colored, often black, and they are not soluble in water. They can sometimes be decomposed by using strong acids, with liberation of hydrogen sulfide. Certain sulfides will also dissolve in the presence of excess sulfide or polysulfide ions.

Pyrites (FeS₂), also known as iron pyrites or fool's gold, are the most common sulfide minerals and can be obtained as very large crystals that have a golden luster. Sphalerite (zinc blende; ZnS) and galena (PbS) are major sources of zinc and lead. Orange cinnabar (HgS) and yellow greenockite (CdS) are the major ores for mercury and cadmium, respectively. Molybdenite (MoS₂) is the major ore of molybdenum.

The sulfur content of fossil fuels results from the sulfur in the ancient organisms as well as from subsequent incorporation of mineral sulfur into the hydrocarbon matrix. Gaseous fossil fuels are often contaminated with hydrogen sulfide, which is an increasingly important source of sulfur. Organic derivatives containing the C—S—C linkage are primarily responsible for the sulfur content of petroleum and coal. The so-called organic sulfur in petroleum can be removed by hydrodesulfurization catalysis, involving reaction with hydrogen over a molybdenum catalyst, to give hydrocarbons and hydrogen sulfide. [T.B.R.]

Sulfuric acid A strong mineral acid with the chemical formula H_2SO_4 . It is a colorless, oily liquid, sometimes called oil of vitriol or vitriolic acid. The pure acid has a density of 1.834 at 25°C (77°F) and freezes at 10.5°C (50.90°F). It is an important industrial commodity, used extensively in petroleum refining and in the manufacture of fertilizers, paints, pigments, dyes, and explosives.

Sulfuric acid is produced on a large scale by two commercial processes, the contact process and the lead-chamber process. In the contact process, sulfur dioxide, SO_2 , is converted to sulfur trioxide, SO_3 , by reaction with oxygen in the presence of a catalyst. Sulfuric acid is produced by the reaction of the sulfur trioxide with water. The lead-chamber process depends upon the oxidation of sulfur dioxide by nitric acid in the presence of water, the reaction being carried out in large lead rooms.

Sulfuric acid reacts vigorously with water to form several hydrates. The concentrated acid, therefore, acts as an efficient drying agent, taking up moisture from the air and even abstracting the elements of water from such compounds as sugar and starch. The concentrated acid also acts as a strong oxidizing agent. It reacts with most metals upon heating to produce sulfur dioxide. *See* SULFUR. [F.J.]

Sum rules Formulas in quantum mechanics for transitions between energy levels, in which the sum of the transition strengths is expressed in a simple form. Sum rules are used to describe the properties of many physical systems, including solids, atoms, atomic nuclei, and nuclear constituents such as protons and neutrons. The sum rules are derived from quite general principles, and are useful in situations where the behavior of individual energy levels is too complex to describe by a precise quantum-mechanical theory. *See* ENERGY LEVEL (QUANTUM MECHANICS). In general, sum rules are derived by using Heisenberg's quantum-mechanical algebra to construct operator equalities, which are then applied to particles or the energy levels of a system. *See* QUANTUM MECHANICS. [G.F.Be.]

Sun The star around which the Earth revolves, and the planet's source of light and heat, hence life. The Sun is a globe of gas, 1.4×10^6 km (8.65×10^5 mi) in diameter with a mass 333,000 times the Earth, held together by its own gravity. The surface temperature of the Sun is about 6000 K ($10,000^\circ$ F); since solids and liquids do not exist at these temperatures, the Sun is entirely gaseous. Almost all the gas is in atomic form, although a few molecules exist in the coolest surface regions, such as sunspots.

The Sun is a typical member of the spectral class dG2, stars of surface temperature 6000 K. The d stands for a dwarf, a normal star of that class. *See* SPECTRAL TYPE.

Solar structure. The interior of the Sun can be studied only by inference from the observed properties of the entire star. The great mass of the Sun presses down on the center, requiring a gas with a central density of near 90 g/cm³ and 2×10^7 K (3.6 $\times 10^{6}$ °F) temperature to support it. At these huge temperatures and densities, nuclear reactions take place. The radiation produced flows outward till it is radiated into space by the surface (photosphere) at 6000 K (10,000°F).

Energy production. The energy of the Sun is produced by the conversion of hydrogen into helium. For each hydrogen atom converted, one neutrino is produced. These neutrinos are detected, but less than the expected number. *See* NEUTRINO; SOLAR NEUTRINOS.

The material at the center of the Sun is so dense that a few millimeters are opaque, so the photons created by nuclear reactions are continually absorbed and reemitted and thus make their way to the surface by a random walk. The atoms in the center of the Sun are entirely stripped of their electrons by the high temperatures, and most of the absorption is by continuum processes, such as the scattering of light by electrons. Because there are so many absorption and emission processes along the way, it can take as long as a million years to complete the random walk to the surface.

Convection. In the outer regions of the solar interior, the temperature is low enough for ions and even neutral atoms to form and, as a result, atomic absorption becomes very important. The high opacity makes it very difficult for the radiation to continue outward, so steep temperature gradients are established that result in convective currents. Most of the outer envelope of the Sun is in such convective equilibrium. These large-scale mass motions are responsible for the complex phenomena observed at the surface. *See* CONVECTION (HEAT).

Radiation. Electromagnetic energy is produced by the Sun in essentially all wavelengths. However, more than 95% of the energy is concentrated in the relatively narrow band between 290 and 2500 nm and is accessible to routine observation from ground stations on Earth. The maximum radiation is in the green region, and the eyes of human beings have naturally evolved to be sensitive to this range of the spectrum. The total radiation is called the solar constant. It is not exactly constant but varies slightly ($\pm 0.1\%$) with the solar cycle. The ultraviolet flux, however, varies by substantial factors depending on the exact wavelength, and this affects the Earth's upper atmosphere. See ELECTROMAGNETIC RADIATION; SOLAR CONSTANT.

Atmosphere. Although the Sun is gaseous, it can be seen only to the point at which the density is so high that the material is opaque. This layer, the visible surface of the Sun, is termed the photosphere. Light from father down reaches the Earth by repeated absorption and emission by the atoms, but the deepest layers cannot be seen directly. The surface is actually not sharp, but the Sun is so far away that the smallest distance that can be resolved with the best telescope is about 300 km (200 mi). Since

2150 Sun

the density e-folding height (scale height) is less than 200 km (120 mi), the edge appears sharp. See PHOTOSPHERE.

Above the photosphere the atmosphere is transparent, and its density falls off much more slowly because magnetic fields support the ionized particles. The atmosphere can be seen by using a narrow-band filter or a spectrograph to pick out the isolated wavelengths absorbed by the atmospheric gases. In the upper photosphere it is cooler, and the lines are dark. If the light is imaged in the strongest lines, such as those of hydrogen, a region higher still is seen, called the chromosphere. The light from this region is dominated by the red hydrogen alpha (level $2 \rightarrow 3$ transition) line, which gives it a rosy color seen at a solar eclipse. The chromosphere is a rapidly fluctuating region of jets and waves coming up from the surface. When all the convected energy coming up from below reaches the surface, it is concentrated in the thin material and produces considerable activity. Where the magnetic field is stronger, these waves are absorbed, and raise the temperature to 7000–8000 K ($12,000-14,000^{\circ}$ F). The scale height of the chromosphere is 1000 km (600 mi) or more, so there no longer is a sharp edge. See CHROMOSPHERE; ECLIPSE.

When the Moon obscures the Sun at a total solar eclipse, the vast extended atmosphere of the Sun called the corona can be seen. The corona is a million times fainter than the photosphere, so it is visible only when seen against the dark sky of an eclipse or with very special instruments. Its density is low, but its temperature is high (more than 10^6 K or 1.8×10^6 °F). The hot gas evaporating out from the corona flows steadily to the Earth and farther in the solar wind. See SOLAR CORONA; SOLAR WIND.

Coronal holes. Early coronal observations showed that the corona was occasionally absent over certain regions. In particular, at sunspot minimum it was quite weak over the poles. X-ray pictures revealed great bands of the solar surface essentially devoid of corona for many months. These proved to be regions where the local magnetic fields were connected to quite distant places, so the fields actually reached out to heights from which the solar wind could sweep the gas outward. Analysis of solar wind data showed that equatorial coronal holes were associated with high-velocity streams in the solar wind, and recurrent geomagnetic storms were associated with the return of these holes. Thus the relative intensity of the corona over sunspot regions is partly due to their strong, closed magnetic fields which trap the coronal gas.

Solar activity. There are a number of transient phenomena known collectively as solar activity. These are all connected with sunspots.

Sunspots. Sunspots were discovered around 1610. Heinrich Schwabe announced in 1843 that their number rose and fell with a 10-year period. Subsequent study of the old records revealed an 11-year period since the original discovery.

The number of sunspots peaks soon after the beginning of each cycle and decays to a minimum in 11 years. The first spots of a number cycle always occur at higher latitudes, between 20° and 35° , and the latitude of occurrence decreases as the cycle unfolds (Spörer's law). Almost no spots are observed outside the latitude range of 5-35°. The great majority are small and last a few days, but some last for two rotations. In 1908, George Ellery Hale discovered that sunspots had strong magnetic fields. Each spot group contains positive and negative magnetic polarity (monopoles are forbidden by Maxwell's laws). Hale found that the polarities were mirrored, with the same polarity generally leading in one hemisphere and following in the other. He found that with each new number cycle the lead polarity switches, so that the complete magnetic cycle lasts 22 years. But each new number cycle starts a few years before the end of the previous one, so the average duration of a half-cycle is nearly 14 years. See MAGNETISM.

The darkness of sunspots (Fig. 1) is probably due to the intense magnetic fields (3000 gauss or 0.3 tesla), which cool the surface



Fig. 1. Large symmetric sunspot photographed in $H\alpha$ light. Clock indicates time of photograph. (*Big Bear Solar Observatory*)

by suppressing the normal convective energy flow from below. It takes several days for the darkening to occur.

Although the sunspot is cool, its neighborhood is the scene of the hottest and most intense activity, generally referred to as an active region. Magnetic energy is continually released there. The corona above an active region is hot and dense, roughly three times hotter and denser than in quiet regions.

Prominences. The term "prominence" is used for any cloud of cool gas in the corona, where it appears bright against the sky. Because these clouds absorb the chromospheric light and scatter it, they appear dark against the solar disk in H α and other strong lines. In continuous light they are transparent. At the limb we see the chromospheric light they scatter against the dark sky. Since they are much denser than the corona, something must hold them up against gravity. Prominences are found only in regions of horizontal magnetic fields that support them. Thus filaments on the disk, which may last for weeks, are good markers of the magnetic boundaries. When the magnetic structure changes, prominences become unstable and erupt, always upward. They also may be ejected by solar flares or appear as graceful loops raining from the corona after flares. Erupting prominences are probably the source of coronal mass ejections, in which a bubble of coronal material erupts outward at several hundred kilometers per second and flows out into interplanetary space.

Plages. Just as prominences occur when the magnetic field changes from one sign to the other, plages occur whenever the magnetic field is vertical and relatively strong but not strong enough to form a sunspot. They are bright regions in any strong spectrum line, because the chromosphere is heated there. In a typical active region, the preceding magnetic field is clumped in a sunspot and the following field spread out in a plage. In H α light, the plage is seen to be connected to the sunspot by dark fibrils outlining the lines of force.

Flares. The most spectacular activity associated with sunspots is the solar flare (Fig. 2). A flare is defined as an abrupt increase in the H α emission from the sunspot region. The brightness of the flare may be up to eight times that of the chromosphere; the rise time is seldom longer than a few minutes. The H α brightening results from heating of the chromosphere at the foot points of the magnetic field by a tremendous energy release in the atmosphere. While flares are usually visible only in chromospheric lines, the foot points of big flares can be seen in white light. From the foot points, a cloud of hot material, up to 3×10^7 K (5.4 \times $10^{7\circ}$ F) arises and concentrates at the arch tops. This cloud condenses out in an array of loop prominences. An active sunspot



Fig. 2. The great "sea horse" flare of August 7, 1972, late in the flare, photographed in the blue wing of the H α line. The neutral line between two bright strands is crossed by an arcade of bright loop prominences raining down from the corona. (*Big Bear Solar Observatory*)

group produces a hierarchy of flares, a few big and many small ones.

Flares are often associated with the eruption of filaments. A few minutes after the eruption begins, there is an abrupt acceleration and a storm of energetic particles is produced, heating the corona to flare brightness.

The flare produces a huge stream of solar energetic particles (SEP) as well as shock waves. A huge magnetohydrodynamic shock wave flies out at about 1000 km/s (600 mi/s) and continues into interplanetary space, often reaching the Earth. The wave produces a huge radio burst in the meter-wavelength range as it excites the coronal layers. The energetic nuclei produce gamma-ray lines from nuclear reactions as they penetrate to the photosphere. If they are sufficiently numerous, they heat the photosphere faster than it can reemit energy and a white light flare is observed, usually in the form of bright transient flashes at the foot points of the flare loops. The particles reach the Earth in a great particle storm.

Sun dog A bright spot of light that sometimes appears on either side of the Sun, the same distance above the horizon as the Sun, and separated from it by an angle of about 22° (see illustration). For higher Sun elevations, the angle increases slightly. These spots are known by many common names: sun



Sun dogs on either side of the Sun. Also visible are the 22° halo and the parhelic circle. (*Courtesy of Robert Greenler*)

dogs, mock suns, false suns, or the 22° parhelia. They usually show a red edge on the side closest to the Sun. On some occasions the entire spectrum of colors can be spread out in the sun-dog spot but, commonly, the red edge is followed by an orange or yellow band that merges into a diffuse white region. The effects result from the refraction by sunlight through small, flat, hexagonal-shaped ice crystals falling through the air such that their flat faces are oriented nearly horizontally. See HALO; METEOROLOGICAL OPTICS. [R.Gr.]

Sundew Any plant of the genus *Drosera* (90 species) of the family Droseraceae. Sundews are small, herbaceous, insectivorous plants that grow on all the continents, especially Australia. Numerous glandular hairs (tentacles) on the leaf secrete a viscous fluid which traps a visiting insect. The tentacles then bend inward about the victim, bringing it into contact with the surface of the leaf where it is digested. The droplets secreted by the glands on the leaves glitter like dewdrops in the morning sunlight; hence, the name, sundew. *See* INSECTIVOROUS PLANTS; NEPENTHALES; SECRETORY STRUCTURES (PLANT). [P.D.St./E.I.C.]

Sundial An instrument for telling time by the Sun. It is composed of a style that casts a shadow and a dial plate, which is the surface upon which hour lines are marked and upon which the shadow falls. The style lies parallel to Earth's axis. The construction of the hour lines is based on the assumption that the apparent motion of the Sun is always on the celestial equator. The most widely used form is the horizontal dial that indicates local apparent time (Sun time). Other forms of the sundial indicate local mean time, and standard time. See TIME. [R.N.M.]

Sunfish A name given to certain members of the freshwater family Centrarchidae as well as to members of the marine family Molidae. There are two species of oceanic sunfishes in the genus *Mola*; both are large, stout, deep-bodied fishes with rough skin and are found in warm seas, where they are frequently seen near the surface. These fishes appear to be tailless.

The fresh-water sunfishes are grouped together with the crappies. The warmouth sunfish (*Chaenobryttus gulosus*) is the only species found in the United States, where it occurs in the Mississippi River drainage and in streams along the Atlantic coast. It differs from other sunfishes in having teeth on the tongue.

The crappies are game fish of considerable importance. Crappies are a favorite species for pond culture. They can be readily transplanted and, under favorable conditions, multiply prodigiously. *See* PERCIFORMES. [C.B.C.]

Sunflower Helianthus annuus, the most widely distributed of the 50 native North American species of this genus of the family Compositae. It is an extremely variable species, with two main divisions. The first involves wild weedy plants found along roadways and other recently disturbed areas; the second, domesticated plants grown in fields and gardens. See ASTERALES.

Within the domestic type there are two categories of plants: the ornamental, which has a few branches with larger heads than the wild, and the crop type, which has only a single stem and the largest head of all sunflowers (see illustration). Crop types are either oil or nonoil. Plant breeders have modified the plant for adaptation to modern, large-scale farming and have increased the oil content of the seeds. The present worldwide interest in growing sunflowers as a crop is due to the increased yield of the new commercially available oilseed hybrids.

Sunflowers are grown on all the continents and in many countries throughout the world. Russia is the major producer, followed by Argentina, the United States, and Canada. Sunflower oil (sunoil) is the second most important vegetable crop oil. It is



Maturing sunflower. (U.S. Department of Agriculture)

a high-quality oil and is high in linoleic fatty acid. The oil is used in cooking, salad dressing, mayonnaise, margarine, and soap. [B.H.B.]

Sunlamp A special form of mercury arc discharge lamp designed to produce ultraviolet radiation. These lamps also produce some radiant energy in the visible region of the spectrum, thus having a light output as well as an ultraviolet output. The lamps are principally used for producing a skin tan on the human



Structure of sunlamps. (a) FS-40. (b) RS.

body. The less common uses include therapeutically producing vitamin D in the body for the treatment of rickets and causing fluorescence or photochemical reactions.

Sunlamps have ultraviolet radiation at wavelengths above 280 nanometers. The lower limit is set by the fact that the quartz and high-silica glass used for lamp envelopes do not transmit below 280 nm. The two lamps designed for tanning are the FS-40 and the RS. The FS-40 tubular fluorescent lamp (illus. *a*) operates with a low-pressure mercury arc that causes a special chemical phosphor coating inside the tube to radiate ultraviolet energy. The RS sunlamp (illus. *b*) is a reflector unit containing a high-pressure mercury arc tube for generating ultraviolet energy plus a tungsten filament (similar to incandescent lamp filaments) in series with the arc tube to serve as a ballast. *See* MERCURY-VAPOR LAMP; ULTRAVIOLET LAMP. [G.R.P]

Superacid An acid which has an extremely great protondonating ability. It has proved convenient to define a superacid somewhat arbitrarily as an acid, or more generally, an acidic medium, which has a proton-donating ability equal to or greater than that of anhydrous (100%) sulfuric acid.

Superacids belong to the general class of proton or Brönsted acids. A proton acid is defined as any species which can act as a source of protons and which will therefore protonate a suitable base, as in reaction (1).

$$HA + B \rightleftharpoons BH^+ + A^- \tag{1}$$

The strengths of acids are often compared by measuring the extent of their ionization in water, that is, the extent to which they can protonate the base water, as in reaction (2).

$$HA + H_2 0 \rightleftharpoons H_3 0^+ + A^- \tag{2}$$

However, all strong acids are fully ionized in dilute aqueous solution, and they therefore appear to have the same strength. Their strengths are said to be reduced or leveled to that of the hydronium ion (H_3O^+) , which is the most highly acidic species that can exist in water. In any case, many of the superacids react with and are destroyed by water. For these reasons, the strengths of superacids cannot be measured by the conventional means of utilizing their aqueous solutions. The acidities of superacids can, however, be conveniently measured in terms of the Hammett acidity function. See ACID AND BASE.

The Hammett acidity function is a method of measuring acidity based on the determination of the ionization ratios of suitable weak bases (indicators), usually by means of the change in absorption spectrum that occurs on protonation of the base, although the nuclear magnetic resonance (NMR) spectrum has also been used. The Hammett acidity function (H_0) is defined by Eq. (3), where K_{BH+} is the dissociation constant of the acid

$$H_0 = pK_{BH+} - \log\frac{[BH^+]}{[B]}$$
(3)

form of the indicator and $[BH^+]/[B]$ is the ionization ratio of the indicator. Hammett acidity function (H_0) values for a number of superacids are given in the table. In each case the value refers to the 100% (anhydrous) acid. Each of the superacids in the table

Hammett acidity function values for several superacids				
Superacid	Formula	$-H_0$		
Sulfuric acid Chlorosulfuric acid Trifluoromethane sulfonic acid Disulfuric acid Fluorosulfuric acid Hydrogen fluoride	H ₂ SO ₄ HSO ₃ Cl HSO ₃ CF ₃ H ₂ S ₂ O ₇ HSO ₃ F HF	11.9 13.8 14.0 14.4 15.1 15.1		

is a liquid at room temperature, and each forms the basis of a solvent system. See IONIC EQUILIBRIUM; SOLUTION. [R.J.Gi.]

Supercharger An air pump or blower in the intake system of an internal combustion engine. Its purpose is to increase the air-charge weight and therefore the power output from an engine of a given size. In an aircraft engine, the supercharger counteracts the power loss that results from the decrease of atmospheric pressure with increase of altitude. Various types of pumps and compressors may be used as superchargers, which are either mechanically driven by the engine crankshaft or powered by the engine exhaust gas. See COMPRESSOR; PUMP; TURBOCHARGER.

Some production automobile engines use a spiral-type supercharger, while others use a pressure-wave supercharger. However, automotive, marine, and stationary engines generally use a positive-displacement Roots blower driven from the engine crankshaft. See INTERNAL COMBUSTION ENGINE.

In a supercharged diesel engine, the increased air charge allows the engine to burn more fuel and produce greater power without creating excessive pressure inside the cylinder. Supercharging the diesel engine makes ignition of the fuel easier without requiring a fuel of better quality. See DIESEL ENGINE.

To enable a reciprocating aircraft engine to develop its rated sea-level power at altitude, a supercharger must be used to increase the pressure and weight of the intake air charge. Centrifugal compressors are generally used because of their relatively small size for a given capacity, and are driven either by a gear drive from the crankshaft or by the engine exhaust gas. With gear drive, a lower ratio is commonly used at low and medium altitudes, with a change to a higher ratio at high altitudes. *See* AIRCRAFT ENGINE. [J.A.B.; D.L.An.]

Supercomputer A computer which, among existing general-purpose computers at any given time, is superlative, often in several senses: highest computation rate, largest memory, or highest cost. Predominantly, the term refers to the fastest "number crunchers," that is, machines designed to perform numerical calculations at the highest speed that the latest electronic device technology and the state of the art of computer architecture allow.

The demand for the ability to execute arithmetic operations at the highest possible rate originated in computer applications areas collectively referred to as scientific computing. Large-scale numerical simulations of physical processes are often needed in fields such as physics, structural mechanics, meteorology, and aerodynamics. A common technique is to compute an approximate numerical solution to a set of partial differential equations which mathematically describe the physical process of interest but are too complex to be solved by formal mathematical methods. This solution is obtained by first superimposing a grid on a region of space, with a set of numerical values attached to each grid point. Large-scale scientific computations of this type often require hundreds of thousands of grid points with 10 or more values attached to each point, with 10 to 500 arithmetic operations necessary to compute each updated value, and hundreds of thousands of time steps over which the computation must be repeated before a steady-state solution is reached. See COMPU-TATIONAL FLUID DYNAMICS; NUMERICAL ANALYSIS; SIMULATION.

Two lines of technological advancement have significantly contributed to what roughly amounts to a doubling of the fastest computers' speeds every year since the early 1950s—the steady improvement in electronic device technology and the accumulation of improvements in the architectural designs of digital computers.

Computers incorporate very large-scale integrated (VLSI) circuits with tens of millions of transistors per chip for both logic and memory components. A variety of types of integrated circuitry is used in contemporary supercomputers. Several use high-speed complementary metallic oxide semiconductor (CMOS) technology. Throughout most of the history of digital computing, supercomputers generally used the highest-performance switching circuitry available at the time—which was usually the most exotic and expensive. However, many supercomputers now use the conventional, inexpensive device technology of commodity microprocessors and rely on massive parallelism for their speed. *See* COMPUTER STORAGE TECHNOLOGY; CONCURRENT PROCESSING; INTEGRATED CIRCUITS; LOGIC CIRCUITS; SEMICONDUCTOR MEMORIES.

Increases in computing speed which are purely due to the architectural structure of a computer can largely be attributed to the introduction of some form of parallelism into the machine's design: two or more operations which were performed one after the other in previous computers can now be performed simultaneously. *See* COMPUTER SYSTEMS ARCHITECTURE.

Pipelining is a technique which allows several operations to be in progress in the central processing unit at once. The first form of pipelining used was instruction pipelining. Since each instruction must have the same basic sequence of steps performed, namely instruction fetch, instruction decode, operand fetch, and execution, it is feasible to construct an instruction pipeline, where each of these steps happens at a separate stage of the pipeline. The efficiency of the instruction pipeline depends on the likelihood that the program being executed allows a steady stream of instructions to be fetched from contiguous locations in memory.

The central processing unit nearly always has a much faster cycle time than the memory. This implies that the central processing unit is capable of processing data items faster than a memory unit can provide them. Interleaved memory is an organization of memory units which at least partially relieves this problem.

Parallelism within arithmetic and logical circuitry has been introduced in several ways. Adders, multipliers, and dividers now operate in bit-parallel mode, while the earliest machines performed bit-serial arithmetic. Independently operating parallel functional units within the central processing unit can each perform an arithmetic operation such as add, multiply, or shift. Array processing is a form of parallelism in which the instruction execution portion of a central processing unit is replicated several times and connected to its own memory device as well as to a common instruction interpretation and control unit. In this way, a single instruction can be executed at the same time on each of several execution units, each on a different set of operands. This kind of architecture is often referred to as single-instruction stream, multiple-data stream (SIMD).

Vector processing is the term applied to a form of pipelined arithmetic units which are specialized for performing arithmetic operations on vectors, which are uniform, linear arrays of data values. It can be thought of as a type of SIMD processing, since a single instruction invokes the execution of the same operation on every element of the array. See COMPUTER PROGRAMMING; PRO-GRAMMING LANGUAGES.

A central processing unit can contain multiple sets of the instruction execution hardware for either scalar or vector instructions. The task of scheduling instructions which can correctly execute in parallel with one another is generally the responsibility of the compiler or special scheduling hardware in the central processing unit. Instruction-level parallelism is almost never visible to the application programmer.

Multiprocessing is a form of parallelism that has complete central processing units operating in parallel, each fetching and executing instructions independently from the others. This type of computer organization is called multiple-instruction stream, multiple-data stream (MIMD). See MULTIPROCESSING. [D.W.M.]

Superconducting devices Devices that perform functions in the superconducting state that would be difficult or impossible to perform at room temperature, or that contain components which perform such functions. The superconducting state involves a loss of electrical resistance and occurs in many metals and alloys at temperatures near absolute zero. An enormous impetus was provided by the discovery in 1986 of a new class of ceramic, high-transition-temperature (T_c)

superconductors, which has resulted in a new superconducting technology at liquid nitrogen temperature. Superconducting devices may be conveniently divided into two categories: small-scale thin-film devices, and large-scale devices which employ zero-resistance superconducting windings made of type II superconducting materials. *See* SUPERCONDUCTIVITY.

Small-scale devices. A variety of thin-film devices offer higher performance than their nonsuperconducting counterparts. The prediction and discovery in the early 1960s of the Josephson effects introduced novel opportunities for ultrasensitive detectors, high-speed switching elements, and new physical standards. Niobium-based devices, patterned on silicon wafers using photolithographic techniques taken over from the semiconductor industry, have reached a high level of development, and a variety of such devices are commercially available. These devices operate at or below 4.2 K ($-452^{\circ}F$), the temperature of liquid helium boiling under atmospheric pressure. See LIQUID HELIUM.

The discovery of the high-transition-temperature superconductors has enabled the operation of devices in liquid nitrogen at 77 K (-321° F). Not only is liquid nitrogen much cheaper and more readily available than liquid helium, but it also boils away much more slowly, enabling the use of simpler and more compact dewars or simpler, relatively inexpensive refrigerators. Of the new ceramic superconductors, only $YBa_2Cu_3O_{7-x}$ (YBCO) has been developed in thin-film form to the point of practical applications, and several devices are available. Intensive materials research has resulted in techniques. notably laser-ablation and radio-frequency sputtering, for the epitaxial growth of highquality films with their crystalline planes parallel to the surface of the substrate. Most of the successful Josephson-junction devices have been formed at the interface between two grains of YBCO. These so-called grain-boundary junctions are made by depositing the film either on a bicrystal in which the two halves of the substrate have a carefully engineered in-plane misalignment of the crystal axes, or across a step-edge patterned in the substrate. See CRYOGENICS; GRAIN BOUNDARIES.

Two types of superconducting quantum interference device (SQUID) detect changes in magnetic flux: the dc SQUID and the rf SQUID. The dc SQUID, which operates with a dc bias current, consists of two Josephson junctions incorporated into a superconducting loop. The maximum dc supercurrent, known as the critical current, and the current-voltage (I-V) characteristic of the SQUID oscillate when the magnetic field applied to the device is changed. The oscillations are periodic in the magnetic flux Φ threading the loop with a period of one flux quantum, $\Phi_0 = h/2e \approx 2.07 \times 10^{-15}$ weber, where h is Planck's constant and *e* is the magnitude of the charge of the electron. Thus, when the SQUID is biased with a constant current, the voltage is periodic in the flux. The SQUID is almost invariably operated in a flux-locked loop. A change in the applied flux gives rise to a corresponding current in the coil that produces an equal and opposite flux in the SQUID. The SQUID is thus the null detector in a feedback circuit, and the output voltage is linearly proportional to the applied flux. See JOSEPHSON EFFECT; SQUID.

The rf SQUID consists of a single Josephson junction incorporated into a superconducting loop and operates with an rf bias. The SQUID is coupled to the inductor of an *LC*-resonant circuit excited at its resonant frequency, typically 30 MHz. The characteristics of rf voltage across the tank circuit versus the rf current depends on applied flux. With proper adjustment of the rf current, the amplitude of the rf voltage across the tank circuit oscillates as a function of applied flux. The rf SQUID is also usually operated in a feedback mode.

SQUIDs are mostly used in conjunction with an input circuit. For example, magnetometers are made by connecting a superconducting pickup loop to the input coil to form a flux transformer. A magnetic field applied to the pickup loop induces a persistent current in the transformer and hence a magnetic flux in the SQUID. These magnetometers have found application in geophysics, for example, in magnetotellurics. *See* GEOPHYSICAL EXPLORATION; MAGNETOMETER.

Low-transition-temperature SQUIDs are widely used to measure the magnetic susceptibility of tiny samples over a wide temperature range. Another application is a highly sensitive voltmeter, used in measurements of the Hall effect and of thermoelectricity. Low-transition-temperature SQUIDs are used as ultrasensitive detectors of nuclear magnetic and nuclear quadrupole resonance, and as transducers for gravitationalwave antennas. So-called scanning SQUIDs are used to obtain magnetic images of objects ranging from single-flux quanta trapped in superconductors to subsurface damage in two metallic sheets riveted together. See HALL EFFECT; MAGNETIC SUSCEPTIBIL-ITY; NUCLEAR MAGNETIC RESONANCE (NMR); NUCLEAR QUADRUPOLE RESONANCE; THERMOELECTRICITY; VOLTMETER.

Perhaps the single largest area of application is biomagnetism, notably to image magnetic sources in the human brain or heart. In these studies an array of magnetometers or gradiometers is placed close to the subject, both generally being in a magnetically shielded room. The fluctuating magnetic signals recorded by the various channels are analyzed to locate their source. These techniques have been used, for example, to pinpoint the origin of focal epilepsy and to determine the function of the brain surrounding a tumor prior to its surgical removal. *See* BIOMAG-NETISM.

The most sensitive detector available for millimeter and submillimeter electromagnetic radiation is the superconductorinsulator-superconductor (SIS) guasiparticle mixer. In this tunnel junction, usually niobium-aluminum oxide-niobium, the Josephson supercurrent is quenched and only single electron tunneling occurs. The current-voltage characteristic exhibits a very sharp onset of current at a voltage $2\Delta/e$, where Δ is the superconducting energy gap. The mixer is biased near this onset where the characteristics are highly nonlinear and used to mix the signal frequency with the frequency of a local oscillator to produce an intermediate frequency that is coupled out into a low-noise amplifier. These mixers are useful at frequencies up to about 750 GHz (wavelengths down to 400 micrometers). Such receivers are of great importance in radio astronomy, notably for airborne, balloon-based, or high-altitude, ground-based telescopes operating above most of the atmospheric water vapor. See RADIO ASTRONOMY.

The advent of high-transition-temperature superconductors stimulated major efforts to develop passive radio-frequency and microwave components that take advantage of the low electrical losses offered by these materials compared with normal conductors in liquid nitrogen. The implementation of thin-film YBCO receiver coils has improved the signal-to-noise ratio of nuclear magnetic resonance (NMR) spectrometers by a factor of 3 compared to that achievable with conventional coils. This improvement enables the data acquisition time to be reduced by an order of magnitude. These coils also have potential applications in low-frequency magnetic resonance imaging (MRI). High-transition-temperature bandpass filters have application in cellular communications. See ELECTRIC FILTER; MOBILE RADIO. [J.C.]

Large-scale devices. Large-scale applications of superconductivity comprise medical, energy, transportation, high-energy physics, and other miscellaneous applications such as highgradient magnetic separation. When strong magnetic fields are needed, superconducting magnets offer several advantages over conventional copper or aluminum electromagnets. Most important is lower electric power costs because once the system is energized only the refrigeration requires power input, generally only 5–10% that of an equivalent-field resistive magnet. Relatively high magnetic fields achievable in unusual configurations and in smaller total volumes reduce the costs of expensive forcecontainment structures. *See* MAGNET.

Niobium-titanium (NbTi) has been used most widely for large-scale applications, followed by the A15 compounds, which include niobium-tin (Nb₃Sn), niobium-aluminum (Nb-Al),

niobium-germanium (Nb-Ge), and vanadium-gallium (Va₃Ga). Niobium-germanium held the record for the highest critical field (23 K; -418.5° F) until the announcement of high-temperature ceramic superconductors. See A15 PHASES.

Significant advances have been made in high-temperature superconducting wire development. Small coils have been wound that operate at 20 K (-410° F). Current leads are in limited commercial use. Considerable development remains necessary to use these materials in very large applications.

MRI dominates superconducting magnet systems applications. Most of the MRI systems are in use in hospitals and clinics, and incorporate superconducting magnets. *See* MEDICAL IMAGING.

Some of the largest-scale superconducting magnet systems are those considered for energy-related applications. These include magnetic confinement fusion, superconducting magnetic energy storage, magnetohydrodynamic electrical power generation, and superconducting generators. *See* MAGNETOHYDRODY-NAMIC POWER GENERATOR; NUCLEAR FUSION.

In superconducting magnetic energy storage superconducting magnets are charged during off-peak hours when electricity demand is low, and then discharged to add electricity to the grid at times of peak demand. The largest systems would require large land areas, for example, an 1100m-diameter (3600-ft) site for a 5000-MWh system. However, intermediate-size systems are viable. A 6-T peak-field solenoidal magnet system designed for the Alaskan power network stores 1800 megajoules (0.5 MWh). High-purity-aluminum-stabilized niobium-titanium alloy conductor carrying 16 kiloamperes current is used for the magnet winding.

Superconducting magnets have potential applications for transportation, such as magnetically levitated vehicles. In addition, superconducting magnets are used in particle accelerators and particle detectors. *See* MAGNETIC LEVITATION; PARTICLE ACCELERATOR; PARTICLE DETECTOR. [A.M.Da.]

Superconductivity A phenomenon occurring in many electrical conductors, in which the electrons responsible for conduction undergo a collective transition into an ordered state with many unique and remarkable properties. These include the vanishing of resistance to the flow of electric current, the appearance of a large diamagnetism and other unusual magnetic effects, substantial alteration of many thermal properties, and the occurrence of quantum effects otherwise observable only at the atomic and subatomic level.

Superconductivity was discovered by H. Kamerlingh Onnes in Leiden in 1911 while studying the temperature dependence of the electrical resistance of mercury within a few degrees of absolute zero. He observed that the resistance dropped sharply to an unmeasurably small value at a temperature of $4.2 \text{ K} (-452^{\circ}\text{F})$. The temperature at which the transition occurs is called the transition or critical temperature, T_c . The vanishingly small resistance (very high conductivity) below T_c suggested the name given the phenomenon.

In 1933 W. Meissner and R. Ochsenfeld discovered that a metal cooled into the superconducting state in a moderate magnetic field expels the field from its interior. This discovery demonstrated that superconductivity involves more than simply very high or infinite electrical conductivity, remarkable as that alone is. *See* MEISSNER EFFECT.

In 1957, J. Bardeen, L. N. Cooper, and J. R. Schrieffer reported the first successful microscopic theory of superconductivity. The Bardeen-Cooper-Schrieffer (BCS) theory describes how the electrons in a conductor form the ordered superconducting state. The BCS theory still stands as the basic explanation of superconductivity, even though extensive theoretical work has embellished it.

There are a number of practical applications of superconductivity. Powerful superconducting electromagnets guide elementary particles in particle accelerators, and they also provide the magnetic field needed for magnetic resonance imaging. Ultrasensitive superconducting circuits are used in medical studies of the human heart and brain and for a wide variety of physical science experiments. A completely superconducting prototype computer has even been built. See MEDICAL IMAGING; PARTICLE ACCELERATOR; SUPERCONDUCTING DEVICES.

Transition temperatures. It was realized from the start that practical applications of superconductivity could become much more widespread if a high-temperature superconductor, that is, one with a high T_c , could be found. For instance, the only practical way to cool superconductors with transition temperatures below 20 K $(-424^{\circ}F)$ is to use liquid helium, which boils at a temperature of 4.2 K $(-452^{\circ}F)$ and which is rather expensive. On the other hand, a superconductor with a transition temperature of 100 K (-280° F) could be cooled with liquid nitrogen, which boils at 77 K $(-321^{\circ}F)$ and which is roughly 500 times less expensive than liquid helium. Another advantage of a high- T_c material is that, since many of the other superconducting properties are proportional to T_c , such a material would have enhanced properties. In 1986 the discovery of transition temperatures possibly as high as 30 K (-406° F) was reported in a compound containing barium, lanthanum, copper, and oxygen. In 1987 a compound of yttrium, barium, copper, and oxygen was shown to be superconducting above 90 K (-298° F). In 1988 researchers showed that a bismuth, strontium, calcium, copper, and oxygen compound was superconducting below 110 K (-262° F), and transition temperatures as high as $135 \text{ K} (-216^{\circ}\text{F})$ were found in a mercury, thallium, barium, calcium, copper, and oxygen compound.

Occurrence. Some 29 metallic elements are known to be superconductors in their normal form, and another 17 become superconducting under pressure or when prepared in the form of thin films. The number of known superconducting compounds and alloys runs into the thousands. Superconductivity is thus a rather common characteristic of metallic conductors. The phenomenon also spans an extremely large temperature range. Rhodium is the element with the lowest transition temperature (370 μ K), while Hg_{0.2}Tl_{0.8}Ca₂Ba₂Cu₃O is the compound with the highest (135 K or -216° F).

Despite the existence of a successful microscopic theory of superconductivity, there are no completely reliable rules for predicting whether a metal will be a superconductor. Certain trends and correlations are apparent among the known superconductors, however—some with obvious bases in the theory—and these provide empirical guidelines in the search for new superconductors. Superconductors with relatively high transition temperatures tend to be rather poor conductors in the normal state.

The ordered superconducting state appears to be incompatible with any long-range-ordered magnetic state: Usually the ferromagnetic or antiferromagnetic metals are not superconducting. The presence of nonmagnetic impurities in a superconductor usually has very little effect on the superconductivity, but the presence of impurity atoms which have localized magnetic moments can markedly depress the transition temperature even in concentrations as low as a few parts per million. See ANTIFERRO-MAGNETISM; FERROMAGNETISM.

Some semiconductors with very high densities of charge carriers are superconducting, and others such as silicon and germanium have high-pressure metallic phases which are superconducting. Many elements which are not themselves superconducting form compounds which are.

Certain organic conductors are superconducting. For instance, brominated polymeric chains of sulfur and nitrogen, known as $(\text{SNBr}_{0,4})_x$, are superconducting below 0.36 K. Other more complicated organic materials have T_c values near 10 K (-442°F). See ORGANIC CONDUCTOR.

Although nearly all the classes of crystal structure are represented among superconductors, certain structures appear to be especially conducive to high-temperature superconductivity. The so-called A15 structure, shared by a series of intermetallic compounds based on niobium, produced several superconductors with T_c values above 15 K (-433°F) as well as the record holder, NbGe, at 23 K (-418°F). Indeed, the robust applications of superconductivity that depend on the ability to carry high current in the presence of high magnetic fields still exclusively use two members of this class: NbTi with $T_c = 8$ K (-445°F), and Nb₃Sn with $T_c = 18.1$ K (-427°F). See A15 PHASES.

After 1986 the focus of superconductivity research abruptly shifted to the copper-oxide-based planar structures, due to their significantly higher transition temperatures. Basically there are three classes of these superconductors, all of which share the common feature that they contain one or more conducting planes of copper and oxygen atoms. The first class is designated by the chemical formula $La_{2-x}A_xCuO_4$, where the A atom can be barium, strontium, or calcium. Superconductivity was originally discovered in the barium-doped system, and systematic study of the substitutions of strontium, calcium, and so forth have produced transition temperatures as high as 40 K ($-388^{\circ}F$).

The second class of copper-oxide superconductor is designated by the chemical formula $Y_1Ba_2Cu_3O_{7-\delta}$, with $\delta<1.0$. Here, single sheets of copper and oxygen atoms straddle the rare-earth yttrium ion and chains of copper and oxygen atoms thread among the barium ions. The transition temperature, 92 K (-294°F), is quite insensitive to replacement of yttrium by many other rare-earth ions. See RARE-EARTH ELEMENTS.

The third class is the most complicated. These compounds contain either single thallium-oxygen layers, represented by the chemical formula $Tl_1Ca_{n-1}Ba_2Cu_nO_{2n+3}$, where *n* refers to the number of copper-oxygen planes, or double thallium-oxygen layers, represented by the chemical formula $Tl_2Ca_{n-1}Ba_2Cu_nO_{2n+4}$. The number of copper-oxygen planes may be varied, and as many as three planes have been included in the structure. Thallium may be replaced by bismuth, thus generating a second family of superconductors. In all of these compounds, the transition temperature appears to increase with the number of planes, but T_c decreases for larger values of *n*.

The spherical molecule comprising 60 carbon atoms (C₆₀), known as a buckyball, can be alloyed with various alkaline atoms which contribute electrons for conduction. By varying the number of conductors in C₆₀, it is possible to boost T_c to a maximum value of 52 K (-366° F). See FULLERENE.

Superconductivity was discovered in magnesium diboride (MgB_2) in January 2001 in Japan. This material may be a good alternative for some of the applications envisioned for high- T_c superconductivity, since this compound has T_c of 39 K (-389° F), is relatively easy to make, and consists of only two elements.

Magnetic properties. The existence of the Meissner-Ochsenfeld effect, the exclusion of a magnetic field from the interior of a superconductor, is direct evidence that the superconducting state is not simply one of infinite electrical conductivity. Instead, it is a true thermodynamic equilibrium state, a new phase which has lower free energy than the normal state at temperatures below the transition temperature and which somehow requires the absence of magnetic flux.

The exclusion of magnetic flux by a superconductor costs some magnetic energy. So long as this cost is less than the condensation energy gained by going from the normal to the superconducting phase, the superconductor will remain completely superconducting in an applied magnetic field. If the applied field becomes too large, the cost in magnetic energy will outweigh the gain in condensation energy, and the superconductor will become partially or totally normal. The manner in which this occurs depends on the geometry and the material of the superconductor. The geometry which produces the simplest behavior is that of a very long cylinder with field applied parallel to its axis. Two distinct types of behavior may then occur, depending on the type of superconductor—type I or type II.

Below a critical field H_c which increases as the temperature decreases below T_c , the magnetic flux is excluded from a type I superconductor, which is said to be perfectly diamagnetic. For a type II superconductor, there are two critical fields, the lower crit-

ical field H_{c1} and the upper critical field H_{c2} . In applied fields less than H_{c1} , the superconductor completely excludes the field, just as a type I superconductor does below H_c . At fields just above H_{c1} , however, flux begins to penetrate the superconductor, not in a uniform way, but as individual, isolated microscopic filaments called fluxoids or vortices. Each fluxoid consists of a normal core in which the magnetic field is large, surrounded by a superconducting region in which flows a vortex of persistent supercurrent which maintains the field in the core. See DIAMAGNETISM.

Thermal properties. The appearance of the superconducting state is accompanied by rather drastic changes in both the thermodynamic equilibrium and thermal transport properties of a superconductor.

The heat capacity of a superconducting material is guite different in the normal and superconducting states. In the normal state (produced at temperatures below the transition temperature by applying a magnetic field greater than the critical field), the heat capacity is determined primarily by the normal electrons (with a small contribution from the thermal vibrations of the crystal lattice) and is nearly proportional to the temperature. In zero applied magnetic field, there appears a discontinuity in the heat capacity at the transition temperature. At temperatures just below the transition temperature, the heat capacity is larger than in the normal state. It decreases more rapidly with decreasing temperature, however, and at temperatures well below the transition temperature varies exponentially as $e^{-\Delta/kT}$, where Δ is a constant and k is Boltzmann's constant. Such an exponential temperature dependence is a hallmark of a system with a gap Δ in the spectrum of allowed energy states. Heat capacity measurements provided the first indications of such a gap in superconductors, and one of the key features of the macroscopic BCS theory is its prediction of just such a gap. See SPECIFIC HEAT OF SOLIDS.

Ordinarily a large electrical conductivity is accompanied by a large thermal conductivity, as in the case of copper, used in electrical wiring and cooking pans. However, the thermal conductivity of a pure superconductor is less in the superconducting state than in the normal state, and at very low temperatures approaches zero. Crudely speaking, the explanation for the association of infinite electrical conductivity with vanishing thermal conductivity is that the transport of heat requires the transport of disorder (entropy). The superconducting state is one of perfect order (zero entropy), and so there is no disorder to transport and therefore no thermal conductivity. *See* ENTROPY; THERMAL CONDUCTION IN SOLIDS.

Two-fluid model. C. J. Gorter and H. B. G. Casimir introduced in 1934 a phenomenological theory of superconductivity based on the assumption that in the superconducting state there are two components of the conduction electron "fluid" (hence the name given this theory, the two-fluid model). One, called the superfluid component, is an ordered condensed state with zero entropy; hence it is incapable of transporting heat. It does not interact with the background crystal lattice, its imperfections, or the other conduction electron component and exhibits no resistance to flow. The other component, the normal component, is composed of electrons which behave exactly as they do in the normal state. It is further assumed that the superconducting transition is a reversible thermodynamic phase transition between two thermodynamically stable phases, the normal state and the superconducting state, similar to the transition between the liquid and vapor phases of any substance. The validity of this assumption is strongly supported by the existence of the Meissner-Ochsenfeld effect and by other experimental evidence. This assumption permits the application of all the powerful and general machinery of the theory of equilibrium thermodynamics. The results tie together the observed thermodynamic properties of superconductors in a very satisfying way.

Microscopic (BCS) theory. The key to the basic interaction between electrons which gives rise to superconductivity was provided by the isotope effect. It is an interaction mediated by the background crystal lattice and can crudely be pictured as follows: An electron tends to create a slight distortion of the elastic lattice as it moves, because of the Coulomb attraction between the negatively charged electron and the positively charged lattice. If the distortion persists for a brief time (the lattice may ring like a struck bell), a second passing electron will see the distortion and be affected by it. Under certain circumstances, this can give rise to a weak indirect attractive interaction between the two electrons which may more than compensate their Coulomb repulsion.

The first forward step was taken by Cooper in 1956, when he showed that two electrons with an attractive interaction can bind together to form a "bound pair" (often called a Cooper pair) if they are in the presence of a high-density fluid of other electrons, no matter how weak the interaction is. The two partners of a Cooper pair have opposite momenta and spin angular momenta. Then, in 1957, Bardeen, Cooper, and Schrieffer showed how to construct a wave function in which all of the electrons (at least, all of the important ones) are paired. Once this wave function is adjusted to minimize the free energy, it can be used as the basis for a complete microscopic theory of superconductivity.

The successes of the BCS theory and its subsequent elaborations are manifold. One of its key features is the prediction of an energy gap. Excitations called quasiparticles (which are something like normal electrons) can be created out of the superconducting ground state by breaking up pairs, but only at the expense of a minimum energy of Δ per excitation; Δ is called the gap parameter. The original BCS theory predicted that Δ is related to T_c by $\Delta = 1.76kT_c$ at T = 0 for all superconductors. This turns out to be nearly true, and where deviations occur they are understood in terms of modifications of the BCS theory. The manifestations of the energy gap in the low-temperature heat capacity and in electromagnetic absorption provide strong confirmation of the theory. [D.N.L.; R.J.So.; M.O.]

Supercontinent The six major continents today are Africa, Antarctica, Australia, Eurasia, North America, and South America. Prior to the formation of the Atlantic, Indian, and Southern ocean basins over the past 180 million years by the process known as sea-floor spreading, the continents were assembled in one supercontinent called Pangea (literally "all Earth"). Pangea came together by the collision, about 300 million years ago (Ma), of two smaller masses of continental rock, Laurasia and Gondwanaland. Laurasia comprised the combined continents of ancient North America (known as Laurentia), Europe, and Asia. Africa, Antarctica, Australia, India, and South America made up Gondwanaland (this name comes from a region in southern India). The term "supercontinent" is also applied to Laurasia and Gondwanaland; hence it is used in referring to a continental mass significantly bigger than any of today's continents. A supercontinent may therefore incorporate almost all of the Earth's continental rocks, as did Pangea, but that is not implied by the word. See CONTINENT; CONTINENTS, EVOLUTION OF.

Laurasia, Gondwanaland, and Pangea are the earliest supercontinental entities whose former existence can be proven. Evidence of older rifted continental margins, for example surrounding Laurentia and on the Pacific margins of South America, Antarctica, and Australia, point to the existence of older supercontinents. The hypothetical Rodinia (literally "the mother of all continents") may have existed 800–1000 Ma, and Pannotia (meaning "the all-southern supercontinent") fleetingly around 550 Ma. Both are believed to have included most of the Earth's continental material. There may have been still earlier supercontinents, because large-scale continents, at least the size of southern Africa or Western Australia, existed as early as 2500 Ma at the end of Archean times. See ARCHEAN; CONTINENTAL MARGIN.

The amalgamation and fragmentation of supercontinents are the largest-scale manifestation of tectonic forces within the Earth. The cause of such events is highly controversial. See PLATE TEC-TONICS. [I.W.O.D.] **Supercritical fields** Static fields that are strong enough to cause the normal vacuum, which is devoid of real particles, to break down into a new vacuum in which real particles exist. This phenomenon has not yet been observed for electric fields, but it is predicted for these fields as well as others such as gravitational fields and the gluon field of quantum chromodynamics.

Vacuum decay in quantum electrodynamics. The original motivation for developing the new concept of a charged vacuum arose in the late 1960s in connection with attempts to understand the atomic structure of superheavy nuclei expected to be produced by heavy-ion linear accelerators. *See* PARTICLE ACCELERATOR.

The best starting point for discussing this concept is to consider the binding energy of atomic electrons as the charge *Z* of a heavy nucleus is increased. If the nucleus is assumed to be a point charge, the total energy *E* of the $1s_{1/2}$ level drops to 0 when *Z* = 137. This so-called *Z* = 137 catastrophe had been well known, but it was argued loosely that it disappears when the finite size of the nucleus is taken into account. However, in 1969 it was shown that the problem is not removed but merely postponed, and reappears around *Z* = 173. Any level *E*(*nj*) can be traced down to a binding energy of twice the electronic rest mass if the nuclear charge is further increased. At the corresponding charge number, called *Z*_{cr}, the state dives into the negative-energy continuum of the Dirac equation (the so-called Dirac sea). The overcritical state acquires a width and is spread over the continuum. See ANTIMATTER; RELATIVISTIC QUANTUM THEORY.

When Z exceeds $Z_{\rm cr}$ a K-shell electron is bound by more than twice its rest mass, so that it becomes energetically favorable to create an electron-positron pair. The electron becomes bound in the $1s_{1/2}$ orbital and the positron escapes. The overcritical vacuum state is therefore said to be charged. See POSITRON.

Clearly, the charged vacuum is a new ground state of space and matter. The normal, undercritical, electrically neutral vacuum is no longer stable in overcritical fields: it decays spontaneously into the new stable but charged vacuum. Thus the standard definition of the vacuum, as a region of space without real particles, is no longer valid in very strong external fields. The vacuum is better defined as the energetically deepest and most stable state that a region of space can have while being penetrated by certain fields.

Superheavy quasimolecules. Inasmuch as the formation of a superheavy atom of Z > 173 is very unlikely, a new idea is necessary to test these predictions experimentally. That idea, based on the concept of nuclear molecules, was put forward in 1969: a superheavy guasimolecule forms temporarily during the slow collision of two heavy ions. It is sufficient to form the quasimolecule for a very short instant of time, comparable to the time scale for atomic processes to evolve in a heavy atom, which is typically of the order 10^{-18} to 10^{-20} . Suppose a uranium ion is shot at another uranium ion at an energy corresponding to their Coulomb barrier, and the two, moving slowly (compared to the K-shell electron velocity) on Rutherford hyperbolic trajectories, are close to each other (compared to the K-shell electron orbit radius). Then the atomic electrons move in the combined Coulomb potential of the two nuclei, thereby experiencing a field corresponding to their combined charge of 184. This happens because the ionic velocity (of the order of c/10) is much smaller than the orbital electron velocity (of the order of c), so that there is time for the electronic molecular orbits to be established, that is, to adjust to the varying distance between the charge centers, while the two ions are in the vicinity of each other. See QUASI-ATOM.

Giant nuclear systems. The energy spectrum for positrons created in, for example, a uranium-curium collision consists of three components: the induced, the direct, and the spontaneous, which add up to a smooth spectrum. The presence of the spontaneous component leads only to 5–10% deviations for normal nuclear collisions along Rutherford trajectories. This situation raises the question as to whether there is any way to get a clear

2158 Supercritical fluid

qualitative signature for spontaneous positron production. Suppose that the two colliding ions, when they come close to each other, stick together for a certain time Δt before separating again. The longer the sticking, the better is the static approximation. For Δt very long, a very sharp line should be observed in the positron spectrum with a width corresponding to the natural lifetime of the resonant positron-emitting state. The observation of such a sharp line will indicate not only the spontaneous decay of the vacuum but also the formation of giant nuclear systems (Z > 180). See LINEWIDTH; NUCLEAR MOLECULE.

Search for spontaneous positron emission. The search for spontaneous positron emission in heavy-ion collisions began in 1976. Of special interest are peak structures in the positron energy distribution. However, the issue of spontaneous positron production in strong fields remains open. If line structures have been observed at all, they are most likely due to nuclear conversion processes. The observation of vacuum decay very much depends on the existence of sufficiently long-lived (at least 10^{-20} s) giant nuclear molecular systems. Therefore, the investigation of nuclear properties of heavy nuclei encountering heavy nuclei at the Coulomb barrier is a primary task.

Other field theories. The idea of overcriticality also has applications in other field theories, such as those of pion fields, gluon fields (quantum chromodynamics), and gravitational fields (general relativity).

A heavy meson may be modeled as an ellipsoidal bag, with a heavy quark Q and antiquark \overline{Q} located at the foci of the ellipsoid. The color-electric or glue-electric field lines do not penetrate the bag surface. The Dirac equation may be solved for light quarks q and antiquarks \overline{q} in this field of force. In the spherical case the potential is zero, and the solutions with different charges degenerate. As the source charges Q and \overline{Q} are pulled apart, the wave functions start to localize. At a critical deformation of the bag, positive and negative energy states cross; that is, overrcriticality is reached and the color field is strong enough that the so-called perturbative vacuum inside the bag rearranges so that the wave functions are pulled to opposite sides and the color charges of the heavy quarks are completely shielded. Hence two new mesons of types $\overline{Q}q$ and $Q\overline{q}$ appear; the original meson fissions. See GLUONS; QUANTUM CHROMODYNAMICS; QUARKS. [W.Gr.]

Supercritical fluid Any fluid at a temperature and a pressure above its critical point; also, a fluid above its critical temperature regardless of pressure. Below the critical point the fluid can coexist in both gas and liquid phases, but above the critical point there can be only one phase. Supercritical fluids are of interest because their properties are intermediate between those of gases and liquids, and are readily adjustable. *See* CRITICAL PHENOMENA; PHASE EQUILIBRIUM.

In a given supercritical fluid the thermodynamic and transport properties are a function of density, which depends strongly on the fluid's pressure and temperature. The density may be adjusted from a gaslike value of 0.1 g/ml to a liquidlike value as high as 1.2 g/ml. Furthermore, as conditions approach the critical point, the effect of temperature and pressure on density becomes much more significant. Increasing the density of supercritical carbon dioxide from 0.2 to 0.5 g/ml, for example, requires raising the pressure from 85 to 140 atm (8.6 to 14.2 megapascals) at 158°F (70°C), but at 95°F (35°C) the required change is only from 65 to 80 atm (6.6 to 8.1 MPa).

For a given fluid, the logarithm of the solubility of a solute is approximately proportional to the solvent density at constant temperature. Therefore, a small increase in pressure, which causes a large increase in the density, can raise the solubility a few orders of magnitude. While almost all of a supercritical fluid's properties vary with density, some of these properties are more like those of a liquid while others are more like those of a gas. See SUPERCRITICAL-FLUID CHROMATOGRAPHY.

In most supercritical-fluid applications the fluid's critical temperature is less than $392^{\circ}F$ ($200^{\circ}C$) and its critical pressure is less than 80 atm (8.1 MPa). High critical temperatures require operating temperatures that can damage the desired product, while high critical pressures result in excessive compression costs. In addition to these pure fluids, mixed solvents can be used to improve the solvent strength. [K.Jo.; R.Len.]

Spercritical-fluid chromatography Any separation technique in which a supercritical fluid is used as the mobile phase. For any fluid, a phase diagram can be constructed to show the regions of temperature and pressure at which gases and liquids, gases and solids, and liquids and solids can coexist. For the gas-liquid equilibrium, there is a certain temperature and pressure, known as the critical temperature and pressure, below which a gas and a liquid can coexist but above which only a single phase (known as a supercritical fluid) can form.

For example, the density of supercritical fluids is usually between 0.25 and 1.2 g/ml and is strongly pressure-dependent. Their solvent strength increases with density, so that molecules that are retained on the column can often be eluted simply by increasing the pressure under which the fluid is compressed.

Diffusion coefficients of solutes in supercritical fluids are tenfold greater than the corresponding values in liquid solvents (although about three orders of magnitude less than the corresponding values in gases). The high diffusivity of solutes in supercritical fluids decreases their resistance to mass transfer in a chromatographic column and hence allows separations to be made either very quickly or at high resolution. Molecules can be separated in a few minutes or less by using packed columns of the type used for high-performance liquid chromatography (HPLC). See DIFFUSION IN GASES AND LIQUIDS.

Despite the relatively large number of separations made using packed-column supercritical-fluid chromatography that have been described since 1962, supercritical-fluid chromatography became popular only in the late 1980s with the commercial introduction of capillary (open-tubular) supercriticalfluid chromatographs. While supercritical-fluid chromatography may never be as widely used as gas chromatography or highperformance liquid chromatography, it provides a useful complement to these chromatographies. *See* CHEMICAL SEPARATION TECHNIQUES; CHROMATOGRAPHY; FLUIDS; GAS CHROMATOGRAPHY; LIQUID CHROMATOGRAPHY. [PR.G.]

Supercritical wing A wing with special streamwise sections, or airfoils, which provide substantial delays in the onset of the adverse aerodynamic effects which usually occur at high subsonic flight speeds.

When the speed of an aircraft approaches the speed of sound, the local airflow about the airplane, particularly above the upper surface of the wing, may exceed the speed of sound. Such a condition is called supercritical flow. On previous aircraft, this supercritical flow resulted in the onset of a strong local shock wave above the upper surface of the wing (illus. *a*). This local



Comparison of airflow about airfoils. (a) Conventional airfoils, Mach number = 0.7. (b) Supercritical airfoils. Mach number = 0.8.

wave caused an abrupt increase in the pressure on the surface of the wing, which may cause the surface boundary-layer flow to separate from the surface, with a resulting severe increase in the turbulence of the flow. The increased turbulence leads to a severe increase in drag and loss in lift, with a resulting decrease in flight efficiency. The severe turbulence also caused buffet or shaking of the aircraft and substantially changed its stability or flying qualities. *See* AERODYNAMIC FORCE; AERODYNAMIC WAVE DRAG; LOW-PRESSURE GAS FLOW; SHOCK WAVE; TRANSONIC FLIGHT.

Supercritical airfoils are shaped to substantially reduce the strength of the shock wave and to delay the associated boundarylayer separation (illus. *b*). Since the airfoil shape allows efficient flight at supercritical flight speeds, a wing of such design is called a supercritical wing. *See* AIRPLANE; WING. [R.T.Wh.]

Superfluidity The frictionless flow of liquid helium at low temperature; also, the flow of electric current without resistance in certain solids at low temperature (superconductivity).

Both helium isotopes have a superfluid transition, but the detailed properties of their superfluid states differ considerably because they obey different statistics. ⁴He, with an intrinsic spin of 0, is subject to Bose-Einstein statistics, and ³He, with a spin of $1/_2$, to Fermi-Dirac statistics. There are two distinct superfluid states in ³He called A and B.

The term "superfluidity" usually implies He II or the A and B phases of ³He, but the basic similarity between these and the "fluid" consisting of pairs of electrons in superconductors is sufficiently strong to designate the latter as a charged superfluid. Besides flow without resistance, superfluid helium and superconducting electrons display quantized circulating flow patterns in the form of microscopic vortices. *See* BOSE-EINSTEIN STATISTICS; LIQUID HELIUM; QUANTIZED VORTICES; SECOND SOUND; SUPERCONDUCTIVITY. [L.J.C.]

Supergiant star A member of a class of evolved stars that occupy the top of the Hertzsprung-Russell diagram to the right of the main sequence. The absolute visual magnitudes (M_V) of supergiants range approximately between -4 and -10, and they are the largest and brightest stars. They are recognized by their spectroscopic characteristics. For example, those in class A have narrow hydrogen lines. Supergiants are subdivided into classes Ib (M_V about -5) and Ia (M_V about -7). A "hypergiant class zero" was later added near $M_V = -10$; the use of transition class Ia-0 at absolute visual magnitudes near -9 is now common. Red supergiants are the largest of all stars, and at maximum can reach diameters approaching that of the orbit of Saturn.

In an evolutionary sense, supergiants are stars above about 10 solar masses and absolute visual magnitude –6 that began mainsequence life hotter than class B1 and cannot evolve to become white dwarfs. As progeny of O stars, supergiants are exceedingly rare. Supergiant masses allow further burning of carbon and oxygen to oxygen, neon, and magnesium; thence to silicon and sulfur; and thence to iron. The iron cores collapse to produce type II supernovae, the condensed remnants at the centers becoming either neutron stars or, from the most massive stars, black holes. *See* ASTRONOMICAL SPECTROSCOPY; BLACK HOLE; HERTZSPRUNG-RUSSELL DIAGRAM; NEUTRON STAR; SPECTRAL TYPE; STAR; STELLAR EVOLUTION; SUPERNOVA. [J.B.Ka.]

Supergranulation A system of convective cells, with typical diameters of 20,000 km (12,000 mi), that cover the Sun's surface. Solar convection cells are invisible in ordinary photographs.

High-resolution photographs of the Sun's visible surface reveal the granulation, a closely packed cellular grid having bright (hot) centers surrounded by dark (cool) lanes. Granules have lifetimes of 10–30 minutes and average diameters of 1000 km (620 mi). In the 1930s, L. Biermann suggested that, in the outer 30% of the Sun, heat from the Sun's interior is transported to the surface by convection, just like hot rising bubbles in a pot

of boiling soup heated from below. The granules are the boiling bubbles (plumes) of this convective process. Theorists in the 1950s proposed that the bubble sizes are approximately equal to the local scale height *H* (the distance in which density or pressure changes by a factor $e \approx 2.7$). Since *H* varies from about 1000 km (620 mi) at the Sun's surface to 100,000 km (62,000 mi) at the base of the convection zone, a very large range of plume sizes was hypothesized. See CONVECTION (HEAT).

In 1959, R. Leighton modified the spectroheliograph to image the Sun using the Doppler and Zeeman effects. Images of line-ofsight (approaching and receding) gas motions (Dopplergrams) and of magnetic fields (magnetograms) could be obtained. The Dopplergrams showed a new cell structure with area 400 times that of the granulation, having a mainly horizontal flow pattern. Lifetimes of most supergranules are 1–2 days, but observations in 1998 with the Michelson Doppler Imager (MDI) on the Solar and Heliospheric Observatory (SOHO) spacecraft showed that some live longer than 4 days. The cells form an irregular polygonal structure, with most of the down-flow occurring at the polygon vertices. The supergranules fit neatly within an essentially identical network (grid) structure seen in magnetograms. The kinetic energy of the supergranular motions at the Sun's surface exceeds the magnetic field's ability to resist such motions, and thus the magnetic field is dragged forcibly to the supergranule boundaries until the two patterns coincide. This magnetic field, in turn, causes local heating of the upper solar atmosphere (chromosphere), producing a similar chromospheric network pattern seen in high-temperature spectral lines. See DOPPLER EFFECT; FRAUNHOFER LINES; SOLAR MAGNETIC FIELD; SPECTROGRAPH; SPEC-TROHELIOGRAPH; SUN; ZEEMAN EFFECT. [G.W.Si.]

Supergravity A theory that attempts to unify gravitation with the other fundamental interactions. The first, and only, completely successful unified theory was constructed by James Clerk Maxwell, in which the up-to-then unrelated electric and magnetic phenomena were unified in his electrodynamics. *See* FUNDAMENTAL INTERACTIONS; MAXWELL'S EQUATIONS.

Electroweak theory. The second stage of unification concerns the unification of electromagnetic and weak interactions, using Maxwell's theory as a guide. This was accomplished making use of the nonabelian gauge theories invented by C. N. Yang and R. L. Mills, and of spontaneous symmetry breaking. The symmetry of Maxwell's theory is very similar to spatial rotations about an axis, rotating the vector potentials while leaving the electric and magnetic fields unchanged. It is a local invariance because the rotations about a fixed axis can be made by different amounts at different points in spacetime. Thus, Maxwell's theory is invariant under a one-parameter group of transformations U(1). In Yang-Mills theory this local invariance was generalized to theories with larger symmetry groups such as the three-dimensional rotation group $SO(3) \simeq SU(2)$ which has three parameters. The number of parameters of the local symmetry (gauge) group is also equal to the number of 4-vector potentials in the gauge theory based on that group. A detailed analysis of weak and electromagnetic forces shows that their description requires four 4-vector potentials (gauge fields), so that the gauge group must be a fourparameter group. In fact, it is the product $SU(2) \cdot U(1)$. See ELEC-TROWEAK INTERACTION; GAUGE THEORY; GROUP THEORY; SYMMETRY BREAKING

Grand unified theories. In the third stage of unification, electroweak and strong forces are regarded as different components of a more general force which mediates the interactions of particles in a grand unified model. Strong forces are responsible for the interactions of hadrons and for keeping quarks confined inside hadrons. They are described by eight massless 4-vector potentials (gluons), the corresponding eight-parameter group being SU(3). This local symmetry is called color, and the corresponding theory quantum chromodynamics (QCD). The combination SU(3) \cdot SU(2) \cdot SU(1) has strong experimental

2160 Superheater

support, and has come to be known as the standard model. Thus the gauge group of any grand unified model must include the standard model as a subsymmetry. The most dramatic prediction of these theories is the decay of protons. *See* GLUONS; GRAND UNI-FICATION THEORIES; PROTON; QUANTUM CHROMODYNAMICS; QUARKS; STANDARD MODEL.

Supersymmetry and supergravity theories. A still higher and more ambitious stage of unification deals with the possibility of combining grand unified and gravity theories into a superunified theory, also known as supergravity. To achieve this, use is made of the dual role played by local internal symmetry groups. On the one hand, they describe the behavior of forces. On the other hand, they classify the elementary particles (fields) of the theory into multiplets: spin-zero fields in one multiplet, spin-1/2 fields in another multiplet, and so forth, but never fermions and bosons in a single irreducible multiplet of internal symmetry. This last restriction used to be a major obstacle on the way to superunification. This is because, of all the elementary particles, only the guanta of the gravitational field (gravitons) have spin 2, so that a multiplet of elementary particles including the graviton must of necessity involve particles of different spin. But then by an internal symmetry transformation, which is by definition distinct from space-time (Lorentz) transformations, it is possible to "rotate" particles of different spin into one another, thus altering their space-time transformation properties. This apparent paradox can be circumvented if both the internal symmetry and Lorentz transformations are part of a larger (supersymmetric) transformation group which also includes the spin-changing transformations. The irreducible multiplets of such supergroups naturally contain both fermions and bosons. This is how supersymmetry makes its appearance in supergravity theories. See GRAVITON; LORENTZ TRANSFORMATIONS; RELATIVITY; SUPERSYMME-TRY; SYMMETRY LAWS (PHYSICS).

Effective theory. If supergravity models are regarded not as fundamental theories but as effective theories describing the lowenergy behavior of superstring theories, it is possible to make a strong case for their usefulness. In that case, since supergravity is no longer a fundamental theory, it is no longer crucial that supergravity satisfy very stringent physical requirements such as renormalizability. In its role as an effective theory, supergravity has been used in a number of problems in particle physics. *See* ELEMENTARY PARTICLE; QUANTUM FIELD THEORY; SUPERSTRING THEORY. [FM.]

Superheater A component of a steam-generating unit in which steam, after it has left the boiler drum, is heated above its saturation temperature. The amount of superheat added to the steam is influenced by the location, arrangement, and amount of superheater surface installed, as well as the rating of the boiler. The superheater may consist of one or more stages of tube banks arranged to effectively transfer heat from the products of combustion. *See* STEAM-GENERATING UNIT. [G.W.K.]

Superluminal motion Proper motion of an astronomical object apparently exceeding the velocity of light, *c*. This phenomenon is relatively common in the nuclei of quasars, many of which exhibit systematic changes in images of their radio-frequency emission over periods of months to years. In some cases, features in the image appear to separate at a speed inferred to be more than 10 times the speed of light, given the great distance of the quasars from Earth.

Superluminal motion was one of the most exciting discoveries to emerge from a technique in radio astronomy first developed in the late 1960s and called very long baseline interferometry (VLBI). This method involves the tape recording of radio signals from large antennas at up to 10–15 locations across the Earth, and the combination of these signals in a computer to form a radio image of the quasar at extremely high resolution (less than 0.001 arcsecond). See QUASAR; RADIO ASTRONOMY; RADIO TELESCOPE.

Superluminal motion is seen mostly in quasars but also in some other active galactic nuclei. This rapid motion is confined to within a few tens of parsecs of the nucleus, whose power source is believed to be a massive black hole. At least 30 examples of superluminal motion are now known. Most show apparent speeds less than 10c, but examples of speeds above 20c have been found. A very few objects in the Milky Way Galaxy also show superluminal motion. An example is GRS 1915 + 105, a relativistic jet source, which emits strongly in the x-ray as well as in the radio spectrum. Because the object is within the Milky Way Galaxy, its apparent speed of 1.25c is detectable in less than a day. See GALAXY, EXTERNAL; X-RAY ASTRONOMY.

Announcement of the discovery of superluminal motion in 1972 caused widespread concern because of the apparent violation of Albert Einstein's special theory of relativity, even though the basic explanation still favored now was in fact predicted some years before the announcement. Many explanations were proposed (besides Einstein's theory being incorrect), but only the relativistic jet model has stood the test of time. Superluminal motion is explained in this model as primarily a geometric effect. A feature (perhaps a cloud of relativistic plasma) moves away from the nucleus of the quasar at high (relativistic) speed (but less than *c*) at a small angle to the line of sight to the Earth. Radio waves from the moving feature arrive only slightly later than waves from the nucleus, whereas the feature took a much longer time to reach its current position. The motion appears superluminal because the speed is calculated using this much shorter time interval. As the speed approaches *c* and the angle to the line of sight decreases, the apparent speed can be arbitrarily large. In this explanation, no material speeds faster than c are required, so there is no conflict with special relativity. See RELATIVITY. [S.C.U.]

Supermassive stars Hypothetical objects with masses exceeding 60 solar masses, the mass of the largest known ordinary stars (1 solar mass equals 4.4×10^{30} lbm or 2×10^{30} kg). The term is most often used in connection with objects larger than 10^4 solar masses that might be the energy source in quasars and active galaxies. These objects probably do not exist. However, their nonexistence is one of the major assumptions that makes the case for giant black holes, rather than supermassive stars, being the central engines of quasars. *See* BLACK HOLE; QUASAR; STAR. [H.L.Sh.]

Supermultiplet A generalization of the concept of a multiplet. A multiplet is a set of quantum-mechanical states, each of which has the same value of some fundamental quantum number and differs from the other members of the set by another quantum number which takes values from a range of numbers dictated by the fundamental quantum number. The number of states in the set is called the multiplicity or dimension of the multiplet. The concept was originally introduced to describe the set of states in a nonrelativistic quantum-mechanical system with the same value of the orbital angular momentum, L, and different values of the projection of the angular momentum on an axis, M. The values that M can take are the integers between -L and L, 2L + 1 in all. This is the dimension of the multiplet. If the hamiltonian operator describing the system is rotationally invariant, all states of the multiplet have the same energy. A supermultiplet is a generalization of the concept of multiplet to the case when there are several quantum numbers that describe the quantum-mechanical states. See ANGULAR MOMENTUM; SYMMETRY LAWS (PHYSICS).

Both concepts, multiplet and supermultiplet, acquire a precise mathematical meaning by the use of the theory of group transformations. A multiplet is an irreducible representation of a group, G. The quantum number called fundamental in the paragraph above labels the representation of the group. The other quantum number labels the representation of a subgroup G' of G. For angular momentum, the group G is the rotation group, called

special orthogonal group in three dimensions, SO(3), and its subgroup G' is the group of special orthogonal transformations in two dimensions, SO(2). A supermultiplet is a generalization to the case in which the group G is not a group of rank one but has larger rank. A group of rank one has only one quantum number to label its representations. The concept of a multiplet or supermultiplet is particularly useful in the classification of states of physical systems. See GROUP THEORY; NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS; QUANTUM NUMBERS.

The term supermultiplet was first used by E. P. Wigner in 1932 in order to classify the quantum-mechanical states of light atomic nuclei. The constituents of these are protons, *p*, and neutrons, *n*. Each proton and neutron has an intrinsic spin, S, of 1/2 in units of \hbar , which is Planck's constant divided by 2π . The projection of the intrinsic spin on an axis, S_z , is then $S_z = \frac{1}{2}$ or $-\frac{1}{2}$ (spin up or down). In addition to having the same spin, the proton and neutron have essentially the same mass but differ in that the proton is charged whereas the neutron is not. They can thus be regarded as different charge states of the same particle, a nucleon. The distinction can be made formal by introducing a quantum number called isotopic spin, T, which has the value 1/2. The two charge orientations, T_z , are taken to be 1/2 for the proton and -1/2 for the neutron. There are thus four constituents of nuclei, protons and neutrons with spin up and down, that is, $p \uparrow$, $p \downarrow$, $n \uparrow$, and $n \downarrow$. The set of transformations among these constituents forms a group called SU(4), the special unitary group in four dimensions. This is the group G for Wigner's theory. The representations of SU(4), that is, Wigner supermultiplets, are characterized by three quantum numbers (λ_1 , λ_2 , λ_3) with $\lambda_1 \geq$ $\lambda_2 \geq \lambda_3$. See I-SPIN; NUCLEAR STRUCTURE.

Supernova The catastrophic, explosive death of a star, accompanied by the sudden, transient brightening of the star to an optical luminosity comparable to that of an entire galaxy.

A supernova shines typically for several weeks to several months with a luminosity between 2×10^8 and 5×10^9 times that of the Sun, then gradually fades away. Each explosion ejects from one to several tens of solar masses at speeds ranging from thousands to tens of thousands of kilometers per second. The total kinetic energy, 10^{44} joules (2.5×10^{28} megatons of high explosive), is about 100 times the total light output, making supernovae some of the highest-energy explosions in the universe. Unlike its fainter relative, the nova, a supernova does not recur for the same object. See NOVA.

Supernovas may be grouped according to either their observational characteristics or their explosion mechanism. Basically, type I supernovae have no hydrogen in their spectrum; type II supernovae do. Two mechanisms are involved: thermonuclear explosion in white dwarfs and gravitational collapse in massive stars. Type I supernovae of different subclasses can occur by either mechanism, but it is thought that most type II supernovae are powered by gravitational collapse.

During the last thousand years, there have been approximately seven supernovae visible to the unaided eye, in 1006, 1054, 1181, 1408, 1572, 1604, and 1987. SN 1006 may have been as bright as the quarter moon. The first six of these occurred in the Earth's vicinity of the Milky Way Galaxy. But the last, and only, naked-eye supernova since the invention of modern instrumentation occurred in the Large Magellanic Cloud, a small satellite galaxy of the Milky Way about 160,000 light-years away. Supernovae are discovered in other galaxies at a rate of about 150 per year. Most supernovae in the Milky Way Galaxy are obscured by dust, but various arguments suggest that about two type II supernovae per century and one type Ia every other century occur in the Milky Way Galaxy. See MAGELLANIC CLOUDS; MILKY WAY GALAXY.

Type la supernovae. Type la supernovae may be regarded as nature's largest thermonuclear bombs. They occur when an accreting white dwarf, composed of carbon and oxygen, grows to a mass 1.38 times that of the Sun, almost the critical mass that

can be supported by electron degeneracy pressure, and ignites carbon fusion near its center. Ignition occurs when carbon fusion at the center releases energy faster than neutrinos can carry it away. Because the pressure is insensitive to the temperature, a nuclear runaway occurs. Fusion releases energy, which raises the temperature, which makes fusion go faster, but the gas cannot expand and cool. The nuclear runaway spreads in about 1 second through the star. The energy released by this nuclear burning is more than enough to completely blow the white dwarf apart with high velocity. Nothing remains—no neutron star, no black hole, and no burst of neutrino emission. See BINARY STAR; THERMONUCLEAR REACTION.

Type II supernovae. A typical type II supernova results from a star somewhat over 8 solar masses, on the main sequence, that spends its last years as a red supergiant burning progressively heavier fuels in its center. The radius of the star, after hydrogen has burned and the star is part way through helium burning, is roughly 500 solar radii, and its luminosity is already about 100,000 times that of the Sun. Each burning stage is shorter than the previous one. The last stage turns silicon and sulfur into a ball of roughly 1.4 solar masses of iron. Once iron has been produced, no more nuclear energy is available. *See* SUPERGIANT STAR.

A combination of instabilities now leads to the implosion of the iron core to a neutron star. When the density at the center reaches several times that of the atomic nucleus, the collapse halts and briefly springs back owing to the short-range repulsive component of the nuclear force. But the energy of this bounce is soon dissipated, and a hot young neutron star remains which, over the next few seconds, radiates away its heat and binding energy as neutrinos. *See* NEUTRINO; STRONG NUCLEAR INTERAC-TIONS.

The energy output in these neutrinos is enormous, about $3 \times$ 10^{46} joules or 15% of the rest mass of the Sun converted to energy; rivaling the luminosity of the rest of the observable universe in light. A small fraction of these neutrinos, about 0.3%, are absorbed in reactions with neutrons and protons in the regions just outside the neutron star and deposit their energy. Even this small amount of energy is much greater than the gravitational binding of the remaining part of the star external to the newly formed neutron star. A bubble of radiation is inflated by the neutrino energy deposition, the outer boundary of which expands supersonically, driving a shock wave through the rest of the star and ejecting it with high velocity. The main energy of the explosion, though, is carried away as neutrinos. This general picture was confirmed when a neutrino burst of the predicted energy and duration was detected February 23, 1987, from the Large Magellanic Cloud in conjunction with SN 1987A. See NEUTRINO ASTRONOMY; SHOCK WAVE.

Nucleosynthesis. Supernovae are major element factories, responsible for producing most of the elements in nature heavier than nitrogen. The largest yields are of the more abundant elements, including oxygen, silicon, magnesium, neon, iron, and a portion of carbon, but dozens of other elements are also made. *See* NUCLEOSYNTHESIS.

Type la cosmological applications. Because of their brightness and the regularity of their light curves, type Ia supernovae have long been used as standard candles to survey cosmological distances. More recently it has been realized that the relatively small variation that occurs in the peak brilliance of such supernovae may be correlated with their decline rates. Use of this so-called Phillips relation allows even greater precision in distance determination. Using type Ia supernovae in this fashion reveals a surprising result. Two independent analyses show that the expansion rate of the universe is not slowing as might be expected long after the big bang, but is actually accelerating. The pull of gravity can only cause deceleration, so the acceleration is attributed to an invisible form of dark energy that enters into the cosmological equations as a repulsive term. See ASTROPHYSICS, HIGH-ENERGY; COSMOLOGY; STAR; STELLAR EVOLUTION; UNIVERSE; VARIABI E STAR [S.E.Wo.]

2162 Superoxide chemistry

Superoxide chemistry A branch of chemistry that deals with the reactivity of the superoxide ion (O_2^-) , a oneelectron (e^-) adduct of molecular oxygen (dioxygen; O_2) formed by the combination of O_2 and e^- . Because 1–15% of the O_2 that is respired by mammals goes through the O_2^- oxidation state, the biochemistry and reaction chemistry of the species are important to those concerned with oxygen toxicity, carcinogenesis, and aging. Although the name superoxide has prompted many to assume an exceptional degree of reactivity for O_2^- , the use of the prefix in fact was chosen to indicate stoichiometry. Superoxide was the name given in 1934 to the newly synthesized potassium salt (KO₂) to differentiate its two-oxygens-per-metal stoichiometry from that of most other metal–oxygen compounds (NA₂O, Na₂O₂, NaOH, Fe₂O₃).

Ionic salts of superoxide (yellow-to-orange solids), which form from the reaction of dioxygen with metals such as potassium, rubidium, or cesium, are paramagnetic, with one unpaired electron per two oxygen atoms.

In 1969, by means of electron spin resonance (ESR) spectroscopy, superoxide ion was detected as a respiratory intermediate, and metalloproteins were discovered that catalyze the disproportionation of superoxide, that is, superoxide dismutases (SODs), as shown in the reaction below.

$$2O_2^- + 2H^+ \xrightarrow{SOD} O_2 + H_2O_2$$

The biological function of superoxide dismutases is believed to be the protection of living cells against the toxic effects of superoxide. The possibility that superoxide might be an important intermediate in aerobic life provided an impetus to the study of superoxide reactivity.

The most general and universal property of O_2^- is its tendency to act as a strong BrØnsted base. Its strong proton affinity manifests itself in any media. Another characteristic of O_2^- is its ability to act as a moderate one-electron reducing agent. See ACID AND BASE; OXIDATION-REDUCTION.

In general, superoxide ion chemistry does not appear to be sufficiently robust to make superoxide ion a toxin. However, it can interact with protons, halogenated carbons, and carbonyl compounds to yield peroxy radicals that are toxic. See OXYGEN TOXICITY; REACTIVE INTERMEDIATES.

Superoxide does not appear to have exceptional reactivity. Nevertheless, superoxide will continue to be an interesting species for study because of the multiplicity of its chemical reactions and because of its importance as an intermediate in reactions that involve dioxygen and hydrogen peroxide. *See* BIOINOR-GANIC CHEMISTRY; OXYGEN. [D.T.S.]

Superplastic forming A process for shaping superplastic materials, a unique class of crystalline materials that exhibit exceptionally high tensile ductility. Superplastic materials may be stretched in tension to elongations typically in excess of 200% and more commonly in the range of 400–2000%. There are rare reports of higher tensile elongations reaching as much as 8000%. The high ductility is obtained only for superplastic materials and requires both the temperature and rate of deformation (strain rate) to be within a limited range. The temperature and strain rate required depend on the specific material. A variety of forming processes can be used to shape these materials; most of the processes involve the use of gas pressure to induce the deformation under isothermal conditions at the suitable elevated temperature. The tools and dies used, as well as the superplastic material, are usually heated to the forming temperature. The forming capability and complexity of configurations producible by the processing methods of superplastic forming greatly exceed those possible with conventional sheet forming methods, in which the materials typically exhibit 10-50% tensile elongation. See SUPERPLASTICITY.

There are a number of commercial applications of superplastic forming and combined superplastic forming and diffusion bonding, including aerospace, architectural, and ground transportation uses. Examples are wing access panels in the Airbus A310 and A320, bathroom sinks in the Boeing 737, turbo-fanengine cooling-duct components, external window frames in the space shuttle, front covers of slot machines, and architectural siding for buildings. *See* METAL FORMING. [C.H.Ha.]

Superplasticity The unusual ability of some metals and alloys to elongate uniformly thousands of percent at elevated temperatures, much like hot polymers or glasses. Under normal creep conditions, conventional alloys do not stretch uniformly, but form a necked-down region and then fracture after elongations of only 100% or less. The most important requirements for obtaining superplastic behavior include a very small metal grain size, a well-rounded (equiaxed) grain shape, a deformation temperature greater than one-half the melting point, and a slow deformation rate. See ALLOY; CREEP (MATERIALS); EUTECTICS.

Superplasticity is important to technology primarily because large amounts of deformation can be produced under low loads. Thus, conventional metal-shaping processes (for example, rolling, forging, and extrusion) can be conducted with smaller, and cheaper equipment. Nonconventional forming methods can also be used; for instance, vacuum-forming techniques, borrowed from the plastics industry, have been applied to sheet metal to form car panels, refrigerator door linings, and TV chassis parts. *See* METAL FORMING; PLASTICITY. [E.E.U.]

Superposition principle The principle, obeyed by many equations describing physical phenomena, that a linear combination of the solutions of the equation is also a solution.

An effect is proportional to a cause in a variety of phenomena encountered at the level of fundamental physical laws as well as in practical applications. When this is true, equations which describe such a phenomenon are known as linear, and their solutions obey the superposition principle. Thus, when f, g, h, \cdots , solve the linear equation, then s ($s = \alpha f + \beta g + \gamma h + \cdots$, where $\alpha, \beta, \gamma, \cdots$, are coefficients) also satisfies the same equation. See LINEAR ALGEBRA; LINEARITY.

For example, an electric field is proportional to the charge that generates it. Consequently, an electric force caused by a collection of charges is given by a superposition—a vector sum of the forces caused by the individual charges. The same is true for the magnetic field and its cause—electric currents. Each of these facts is connected with the linearity of Maxwell's equations, which describe electricity and magnetism. *See* ELECTRIC FIELD; MAXWELL'S EQUATIONS.

The superposition principle is important both because it simplifies finding solutions to complicated linear problems (they can be decomposed into sums of solutions of simpler problems) and because many of the fundamental laws of physics are linear. Quantum mechanics is an especially important example of a fundamental theory in which the superposition principle is valid and of profound significance. This property has proved most useful in studying implications of quantum theory, but it is also a source of the key conundrum associated with its interpretation.

Its effects are best illustrated in the double-slit superposition experiment, in which the wave function representing a quantum object such as a photon or electron can propagate toward a detector plate through two openings (slits). As a consequence of the superposition principle, the wave will be a sum of two wave functions, each radiating from its respective slit. These two waves interfere with each other, creating a pattern of peaks and troughs, as would the waves propagating on the surface of water in an analogous experimental setting. However, while this pattern can be easily understood for the normal (for example, water or sound) waves resulting from the collective motion of vast numbers of atoms, it is harder to understand its origin in quantum mechanics, where the wave describes an individual quantum, which can be detected, as a single particle, in just one spot along the detector (for example, photographic) plate. The interference pattern will eventually emerge as a result of many such individual quanta, each of which apparently exhibits both wave (interference-pattern) and particle (one-by-one detection) characteristics. This ambivalent nature of quantum phenomena is known as the wave-particle duality. *See* INTERFERENCE OF WAVES; QUANTUM MECHANICS. [W.H.Z.]

Superposition theorem (electric networks)

Essentially, that it is permissible, if there are two or more sources of electromotive force in a linear electrical network, to compute at any element of the network the response of voltage or of current that results from one source alone, and then the response resulting from another source alone, and so on for all sources, and finally to compute the total response to all sources acting together by adding these individual responses.

Thus, if a load of constant resistance is supplied with electrical energy from a linear network containing two batteries, two generators, or one battery and one generator, it would be correct to find the current that would be supplied to the load by one source (the other being reduced to zero), then to find the current that would be supplied to the load by the second source (the first source now being reduced to zero), and finally to add the two currents so computed to find the total current that would be produced in the load by the two sources acting simultaneously.

By means of the principle of superposition, effects are added instead of causes. This principle seems so intuitively valid that there is far greater danger of applying superposition where it is incorrect than of failing to apply it where it is correct. It must be recognized that for superposition to be correct the relation between cause and effect must be linear. [H.H.Sk.]

Supersaturation A solution is at the saturation point when dissolved solute in its crystallizes from it at the same rate at which it dissolves. Under prescribed experimental conditions of temperature and pressure, a solution can contain at saturation only one fixed amount of dissolved solute. However, it is possible to prepare relatively stable solutions which contain a quantity of a dissolved solute greater than that of the saturation value provided solute phase is absent. Such solutions are said to be supersaturated. They can be prepared by changing the experimental conditions of a system so that greater solubility is obtained, perhaps by heating the solution, and then carefully returning the system to or near its original state. The addition of solute phase will immediately relieve supersaturation. Solutions in which there is no spontaneous formation of solute phase for extended periods of time are said to be metastable. There is no sharp line of demarcation between an unstable and metastable solution. The process whereby initial aggregates within a supersaturated solution develop spontaneously into particles of new stable phase is known as nucleation. The greater the degree of supersaturation, the greater will be the number of nuclei formed. See NUCLEATION; PHASE EQUILIBRIUM; PRECIPITATION (CHEMISTRY). [L.Go./R.W.Mu.]

Supersonic diffuser A passive compressor (or shaped duct) in which gas enters at a velocity greater than the speed of sound, is decelerated in a contracting section, and reaches sonic speed at a throat.

Supersonic compression systems can be categorized to three basic types (see illustration). External-compression inlets have the supersonic diffusion taking place at or ahead of the cowl lip (or throat station) and generally employ one or more oblique waves ahead of the normal shock. Internal-compression inlets accomplish supersonic diffusion internally downstream of the cowl lip. Deceleration of the flow is produced by a number of weak reflecting waves in a gradually convergent channel. The



Basic supersonic compression systems. (a) External compression. (b) Internal compression. (c) Combination of external and internal compression.

third system is a combination of external and internal compression and appears to represent an effective compromise. [J.E.C.; L.J.O.]

Supersonic flight Relative motion of a solid body and a gas at a velocity greater than that of sound propagation under the same conditions. The general characteristics of supersonic flight can be understood by considering the laws of propagation of a disturbance or pressure impulse, in a compressible fluid.

If the fluid is at rest, the pressure impulse propagates uniformly with the velocity of sound in all directions, the effect always acting along an ever-increasing spherical surface. If, however, the source of the impulse is placed in a uniform stream, the impulse will be carried by the stream simultaneously with its propagation at sonic velocity relative to the stream. Hence the resulting propagation is faster in the direction of the stream and slower against the stream. If the velocity of the stream past the source of disturbance is supersonic, the effect of the impulse and whose vertex angle decreases from 90° (corresponding to Mach number equal to 1) to smaller and smaller values as the Mach number of the stream increases (see illustration). If the source of the pressure impulse travels through the air at rest, the conditions are analogous. See MACH NUMBER.

Consider the supersonic motion of a wing moving into air at rest. Because signals cannot propagate ahead of the wing, the presence of the wing has no effect on the undisturbed air until the wing passes through it. Hence there must be an abrupt change in the properties of the undisturbed air as it begins to flow over the wing. This abrupt change takes place in a shock wave which



Generation of Mach wave by body at supersonic velocity; zones of action and silence are separated.
2164 Supersonic flow

is attached to the leading edge of the wing, provided that the leading edge is sharp and the flight Mach number is sufficiently large. As the air passes through the shock wave, its pressure, temperature, and density are markedly increased.

Further aft of the leading edge, the pressure of the air is decreased as the air expands over the surface of the wing. Hence the pressure acting on the front part of the wing is higher than the ambient pressure, and the pressure acting on the rear part of the wing is lower than the ambient pressure. The pressure difference between front and rear parts produces a drag, even in the absence of skin friction and flow separation. The wing produces a system of compression and expansion waves which move with it. This phenomenon is similar to that of a speedboat moving with a velocity greater than the velocity of the surface waves. Because of this analogy, supersonic drag is called wave drag. It is peculiar to supersonic flight, and it may represent the major portion of the total drag of a body. See HYPERSONIC FLIGHT; SUBSONIC FLIGHT; TRANSONIC FLIGHT. [J.E.Sc.]

Supersonic flow Fluid motion in which the Mach number M, defined as the speed of the fluid relative to the sonic speed in the same medium, is more than unity. It is, however, common to call the flow transonic when 0.8 < M < 1.4, and hypersonic when M > 5. See MACH NUMBER.

Mach waves. A particle moving in a compressible medium, such as air, emits acoustic disturbances in the form of spherical waves. These waves propagate at the speed of sound (M = 1). If the particle moves at a supersonic speed, the generated waves cannot propagate upstream of the particle. The spherical waves are enveloped in a circular cone called the Mach cone. The generators of the Mach cone are called Mach lines or Mach waves.

Shock waves. When a fluid at a supersonic speed approaches an airfoil (or a high-pressure region), no information is communicated ahead of the airfoil, and the flow adjusts to the downstream conditions through a shock wave. Shock waves propagate faster than Mach waves, and the flow speed changes abruptly from supersonic to less supersonic or subsonic across the wave. Similarly, other properties change discontinuously across the wave. A Mach wave is a shock wave of minimum strength. A normal shock is a plane shock normal to the direction of flow, and an oblique shock is inclined at an angle to the direction of flow. The velocity upstream of a shock wave is always



Typical normal shock, oblique shock, and Mach wave pattern in supersonic flow past a blunt body. *M* is the Mach number and *V* is the particle speed. The curved line parallel to normal and oblique shock waves indicates the end of the velocity vectors.

supersonic. Downstream of an oblique shock, the velocity may be subsonic resulting in a strong shock, or supersonic resulting in a weak shock. The downstream velocity component normal to any shock wave is always subsonic. There is no change in the tangential velocity component across the shock.

In a two-dimensional supersonic flow around a blunt body (see illustration), a normal shock is formed directly in front of the body, and extends around the body as a curved oblique shock. At a sufficient distance away, the flow field is unaffected by the presence of the body, and no discontinuity in velocity occurs. The shock then reduces to a Mach wave. *See* COMPRESSIBLE FLOW; FLUID FLOW; SUPERSONIC FLIGHT. [M.A.S.]

Superstring theory A proposal for a unified theory of all interactions, including gravity. At present, the strong, weak, and electromagnetic interactions are accounted for within the framework of the standard model. This model correctly describes experiments up to the highest energies performed so far, and gives a complete description of the elementary particles and their interactions down to distances of the order of 10^{-18} m. Nevertheless, it has serious limitations, and attempts to overcome them and to unify the forces of nature have been only partly successful. Moreover, these attempts have left standing fundamental difficulties in reconciling gravitation and the laws of guantum mechanics. Superstring theory represents an ambitious program to unify all of the interactions observed in nature, including gravitation, in a theory with no unexplained parameters. In other words, this theory, if successful, should be able to account for all of the particles observed in nature and their interactions. See ELEMENTARY PARTICLE; FUNDAMENTAL INTERACTIONS.

String concept. In string theory, the fundamental objects are not point particles, as in standard theories of elementary particles, but one-dimensional extended objects, the open and closed strings. In such a theory, what are usually called the elementary particles are simply particular quantum states of the string. In superstring theories, space-time is ten-dimensional (space is nine-dimensional). If such theories are to describe nature, six dimensions must be "curled up" or "compact." The main consequence of such extra dimensions is the existence of certain very massive particles. *See* SPACE-TIME.

The essential features of string theories can be understood by analogy with the strings of a musical instrument. Such strings vibrate at a characteristic frequency, as well as any integer multiple of that frequency. Each of these modes of vibration (so-called normal modes) can be excited by plucking or striking the string. In classical physics, the amplitudes of vibration of each mode can take on a continuum of values. If there were a string of atomic dimensions, subject to the laws of quantum mechanics, the energies of this quantum string could take on only discrete values, corresponding to particular quantum states. *See* QUANTUM ME-CHANICS; VIBRATION.

The strings of superstring theory are quite similar. The main difference is that they obey Einstein's principles of special relativity. As a result, since each quantum state has a particular energy, it has a definite mass. Thus, each state of the string behaves as a particle of definite mass. Because it is possible, in principle, to pump an arbitrarily large amount of energy into the string, the theory contains an infinity of different types of particles of arbitrarily large mass. The interactions of these particles are governed by the ways in which the strings themselves interact. To be consistent with the principles of relativity, a string can interact only by splitting into two strings or by joining together with another string to form a third string. As a result, the interactions of strings are nearly unique. This geometric picture of string interactions translates into a precise set of rules for calculating the interaction of individual string states, that is, particles. See RELATIVITY.

Classical solutions. Obtaining a description of superstring theory analogous to quantum field theory is an active topic of research. However, even though the equations that describe this

field theory are not completely known at present, it is known how to find classical solutions of these equations, and by various techniques, an enormous number of such solutions have been found. These include states in which space-time has any dimension between one and ten, and states with many bizarre symmetries and spectra. Each of these solutions then corresponds to a possible ground state of the system. The theories built around some of these states look very much like the real world. Not only are four dimensions flat while six are compact, but they possess gauge symmetries close to that of the standard model. Some have three or four generations of quarks and leptons, as well as light Higgs particles, which are of crucial importance in the standard model. Many of these solutions possess space-time supersymmetry. *See* GAUGE THEORY; HIGGS BOSON; LEPTON; QUARKS.

However, if the theory does describe nature, it must have some mechanism that chooses one of the possible ground states. Because the masses and couplings of the elementary particles depend only on the choice of ground state, determining this true ground state will yield a set of predictions for these quantities. If string theory is a correct theory, these predictions must agree with the experimental values. [M.Di.]

Supersymmetry A conjectured enhanced symmetry of the laws of nature that would relate two fundamental observed classes of particles, bosons and fermions.

All particles can be classified as fermions, such as the electron and quarks, or bosons, such as the photon and graviton. A fundamental characteristic distinguishing these two classes is that they carry different quantum-mechanical spin. If the amount of spin of an elementary particle is measured in terms of the fundamental quantum unit of angular momentum— \hbar , equal to Planck's constant divided by 2π —then bosons always have integer amounts of spin (that is, 0, 1, 2 . . .), while fermions have odd half-integer amounts of spin (that is, 1/2, 3/2, 5/2, . . .). See SPIN (QUANTUM MECHANICS).

There is seemingly a fundamental distinction between particles with differing amounts of spin. For example, bosons like to act collectively (Bose-Einstein statistics), producing such distinctive behavior as the laser, while, conversely, fermions obey the Pauli exclusion principle (and the Pauli-Dirac statistics), which disallows two identical fermions to be in the same state, and explains the stability of matter. Moreover, all the symmetries that are observed in the world relate different particles of the same spin. *See* BOSE-EINSTEIN STATISTICS; FERMI-DIRAC STATISTICS; QUANTUM STATISTICS; SYMMETRY LAWS (PHYSICS).

In contrast, supersymmetry would relate bosons and fermions. This would be a remarkable step forward in understanding the physical world. However, if supersymmetry were realized as an exact symmetry, the particles so related should have almost all their characteristics, such as mass and charge, preserved. Explicitly, any fermion of spin 1/2 should have a boson superpartner of spin 0, while any gauge boson of spin 1 should have a fermion superpartner of spin 1/2. This is apparently a disaster for the idea of supersymmetry since it predicts, for instance, that there should exist a spin-0 boson partner of the electron, the selectron, with electric charge and mass equal to that of the electron. Such a particle would be easy to detect and is certainly ruled out by very many experiments.

The crucial caveat to this negative result is the condition that supersymmetry be realized as an exact symmetry. A fundamental concept of modern physics is spontaneously broken symmetry. Physics displays many examples of symmetries that are exact symmetries of the fundamental equations describing a system, but not of their solutions. In particle physics the spontaneous breaking of a symmetry usually results in a difference in the masses of the particles related by the symmetry; the amount of breaking can be quantified by this mass difference. See SYM-METRY BREAKING.

If supersymmetry is broken by a large amount, then all the superpartners have masses much greater than the particles that are currently observed, and there is little hope of seeing evidence for supersymmetry. However, evidence that supersymmetry is broken by only a moderate amount comes from examination of the properties of the fundamental forces at high energy.

Of the four fundamental forces, the three excluding gravity are very similar in their basic formulation; they are all described by gauge theories, generalizations of the quantum theory of electromagnetism, and quantum electrodynamics (QED). The strength of electrical interaction between two electrons can be quantified in terms of a number, the coupling constant α_1 . However, the quantity α_1 is actually not a constant, but depends on the energies at which the interaction strength is measured. In fact, the interaction strengths, α_1 , α_2 , and α_3 , of the three forces (excluding gravity) all depend on energy, μ . The couplings $\alpha_{1,2,3}$ satisfy differential equations-renormalization group equations-that depend on the types of elementary particles that exist with mass at or below the energy scale μ and that are charged with respect to each of the three interactions. If the fundamental particles include not only the observed particles but also their superpartners, taken to have masses not greater than 1000 GeV heavier than their (observed) partners, then from the renormalization group equations, the couplings α_i are predicted to meet (unify) at a huge energy of 2 \times 10 16 GeV. In contrast, if either supersymmetry is not an underlying symmetry of the world, or it is very badly broken so that the superpartners are very massive, the couplings fail to unify at a single point. See FUNDAMENTAL INTERACTIONS; GAUGE THEORY; QUANTUM ELECTRODYNAMICS; RENORMALIZATION.

Although the unification of couplings is the most significant indication that supersymmetry is a new law of nature, there are a number of other hints in this same direction. By observing the large-scale motions of the galaxies, the average density of large volumes of the universe can be deduced, resulting in a value that is substantially greater than that directly observed in luminous matter such as stars and hot gas. Therefore, a substantial fraction of the mass of the universe must be composed of some form of nonluminous or dark matter. Remarkably, many attractive models of supersymmetry predict that the lightest of all the superpartners is a weakly interacting massive particle with just the right characteristics to be this dark matter. See COSMOL-OGY; UNIVERSE; WEAKLY INTERACTING MASSIVE PARTICLE (WIMP). [J.M.R.]

Suppression (electricity) The process or technique of reducing electrical interference to acceptable levels or to situations having no adverse effect. Suppression techniques may be applied to the interference source, the intervening path, the victim or receptor, or any combination. Normal strategy for interference control is to first suppress the source, if possible, since it may disturb many victims.

For intentional transmitters, suppressing interference may include reducing or eliminating harmonic radiations, restricting the bandwidth, or restricting levels of unnecessary or excessive modulation sidebands. These are usually accomplished by radio frequency filters. *See* ELECTRIC FILTER.

For many devices involving incidental radiators, such as brushtype motors and fluorescent lights, interference suppression may require both filtering and shielding. Electrical filtering may take the form of transient surge suppressors, feed-through capacitors, electromagnetic interference (EMI) filters, ferrite absorbers, isolation transformers, or Faraday shielded isolation transformers. *See* ELECTRIC PROTECTIVE DEVICES.

Shielding to control radiation involves using metal boxes, cases, cabinet housings, or metalized plastic versions thereof. Since the interconnecting cables between equipment offer the greatest threat as an "antenna farm," the dominant suppression technique is to shield the cables, wires, or harnesses. For the best protection from electromagnetic interference, the cable shield should be designed as an extension of the box or equipment shield. Other forms of interference hardening of cables include twisting parallel wire pairs, multiple-layer shields, and

2166 Supramolecular chemistry

absorbing jackets. Where possible, best electromagnetic compatibility performance is obtained by replacing the signal cables with fiber-optic links. *See* ELECTRICAL SHIELDING.

From a system point of view, the radiated emission propensity or pickup susceptibility of interconnected equipment is proportional to the maximum length or dimension and frequency until this length corresponds to an electrical length of about one-half wavelength. Thus, to help suppress electromagnetic interference, the corresponding frequency should not fall within the passband of victims or receptors. *See* ELECTRICAL INTERFERENCE; ELECTRO-MAGNETIC COMPATIBILITY. [D.R.J.W.]

Supramolecular chemistry A highly interdisciplinary field covering the chemical, physical, and biological features of complex chemical species held together and organized by means of intermolecular (noncovalent) bonding interactions. *See* CHEMICAL BONDING; INTERMOLECULAR FORCES.

When a substrate binds to an enzyme or a drug to its target, and when signals propagate between cells, highly selective interactions occur between the partners that control the processes. Supramolecular chemistry is concerned with the study of the basic features of these interactions and with their implementation in biological systems as well as in specially designed nonnatural ones. In addition to biochemistry, its roots extend into organic chemistry and the synthetic procedures for receptor construction, into coordination chemistry and metal ion-ligand complexes, and into physical chemistry and the experimental and theoretical studies of interactions. *See* BIOINORGANIC CHEM-ISTRY; ENZYME; LIGAND; PHYSICAL ORGANIC CHEMISTRY; PROTEIN.

The field started with the selective binding of alkali metal cations by natural as well as synthetic macrocyclic and macropolycyclic ligands, the crown ethers and cryptands. This led to the emergence of molecular recognition as a new domain of chemical research that, by encompassing all types of molecular components and interactions as well as both oligo and polymolecular entities, became supramolecular chemistry. It underwent rapid growth with the development of synthetic receptor molecules of numerous types for the strong and selective binding of cationic, anionic, or neutral complementary substrates of organic, inorganic, or biological nature by means of various interactions (electrostatic, hydrogen bonding, van der Waals, and donor-acceptor). Molecular recognition implies the (molecular) storage and the (supramolecular) retrieval and processing of molecular structural (geometrical and interactional) information. See Hydrogen Bond; Macrocyclic compound; Molecular RECOGNITION.

Many types of receptor molecules have been explored (crown ethers, cryptands, cyclodextrins, calixarenes, cavitands, cyclophanes, cryptophanes, and so on), and many others may be imagined for the binding of complementary substrates of chemical or biological significance. They allow, for instance, the development of substrate-specific sensors or the recognition of structural features in biomolecules (for example, nucleic acid probes, affinity cleavage reagents, and enzyme inhibitors). *See* BIOPOLY-MER; CYCLOPHANE; ENZYME INHIBITION.

A major step in the development of supramolecular chemistry over the last 20 years involved the design of systems capable of spontaneously generating well-defined, supramolecular entities by self-assembly under a given set of conditions.

The information necessary for supramolecular self-assembly to take place is stored in the components, and the program that it follows operates via specific interactional algorithms based on binding patterns and molecular recognition events. Thus, rather than being preorganized, constructed entities, these systems may be considered as self-organizing, programmed supramolecular systems.

Self-assembly and self-organization have recently been implemented in numerous types of organic and inorganic systems. By clever use of metal coordination, hydrogen bonding, and donoracceptor interactions, researchers have achieved the spontaneous formation of a variety of novel and intriguing species such as inorganic double and triple helices termed helicates, catenates, threaded entities (rotaxanes), cage compounds, grids of metal ions, and so on.

Another major development concerns the design of molecular species displaying the ability to perform self-replication, based on components containing suitable recognition groups and reactive functions. Self-recognition processes involve the spontaneous selection of the correct partner(s) in a self-assembly event—for instance, the correct ligand strand in helicate formation.

A major area of interest is the design of supramolecular devices built on photoactive, electroactive, or ionoactive components, operating respectively with photons, electrons, or ions. Thus, a variety of photonic devices based on photoinduced energy and electron transfer may be imagined. Molecular wires, ion carriers, and channels facilitate the flow of electrons and ions through membranes. Such functional entities represent entries into molecular photonics, electronics, and ionics, which deal with the storage, processing, and transfer of materials, signals, and information at the molecular and supramolecular levels. Dynamic and mechanical devices exploit the control of motion within molecular and supramolecular entities. *See* INORGANIC PHOTO-CHEMISTRY; ION TRANSPORT; PHOTOCHEMISTRY.

The design of systems that are controlled, programmed, and functionally self-organized by means of molecular information contained in their components represents new horizons in supramolecular chemistry and provides an original approach to nanoscience and nanotechnology. In particular, the spontaneous but controlled generation of well-defined, functional supramolecular architectures of nanometric size through self-organization—supramolecular nanochemistry—represents a means of performing programmed engineering and processing of nanomaterials. It offers a powerful alternative to the demanding procedures of nanofabrication and nanomanipulation, bypassing the need for external intervention. A rich variety of architectures, properties, and processes should result from this blending of supramolecular chemistry with materials science. See NANOCHEMISTRY; NANOTECHNOLOGY. [J.M.Le.]

Surface (geometry) A two-dimensional geometric figure (a collection of points) in three-dimensional space. The simplest example is a plane—a flat surface. Some other common surfaces are spheres, cylinders, and cones, the names of which are also used to describe the three-dimensional geometric figures that are enclosed (or partially enclosed) by those surfaces. In a similar way, cubes, parallelepipeds, and other polyhedra are surfaces. *See* CUBE; POLYHEDRON; SOLID (GEOMETRY).

Any bounded plane region has a measure called the area. If a surface is approximated by polygonal regions joined at their edges, an approximation to the area of the surface is obtained by summing the areas of these regions. The area of a surface is the limit of this sum if the number of polygons increases while their areas all approach zero. See AREA; CALCULUS; INTEGRATION; PLANE GEOMETRY; POLYGON.

Methods of description. The shape of a surface can be described by several methods. The simplest is to use the commonly accepted name of the surface, such as sphere or cube. In mathematical discussions, surfaces are normally defined by one or more equations, each of which gives information about a relationship that exists between coordinates of points of the surface, using some suitable coordinate system. *See* COORDINATE SYSTEMS.

Some surfaces are conveniently described by explaining how they might be formed. If a curve, called the generator in threedimensional space, is allowed to move in some manner, then each position the generator occupies during this motion is a collection of points, and the set of all such points constitutes a surface that can be said to be swept out by the generator. In particular, if the generator is a straight line, a ruled surface is formed. If the generator is a straight line and the motion is such