McGRAW-HILL

ENCYCLOPEDIA OF SCIENCE & TECHNOLOGY

6 EBE-EYE

www.MHEST.com

Ebenales

An order of flowering plants, division Magnoliophyta (Angiospermae) in the subclass Dilleniidae of the class Magnoliopsida (dicotyledons). The order consists of 5 families and about 1750 species, the Sapotaceae (about 800 species) and Ebenaceae (about 450 species) being the largest and most familiar families. The Ebenales are woody, chiefly tropical, sympetalous plants (those with flowers have the petals joined by their margins, at least toward the base, forming a basal tube, cup, or saucer) with usually twice as many stamens (including staminodes) as corolla lobes. When there is a single set of stamens, these are generally opposite the corolla lobes. Ordinarily there are only one or two ovules in each locule and the placentation is axile (that is, the placenta forms a central axis in an ovary which is divided by longitudinal partitions into two or more chambers). In a cross section of such an ovary, the partitions radiate from the central axis (placenta) to the lateral wall of the ovary. Chicle (from *Manilkara sapota*, in the Sapotaceae) and ebony (*Diospyros ebeneum*) are obtained from members of the Ebenales. *See* CHICLE; DILLENIIDAE; EBONY; MAGNOLIOPHYTA; PLANT KINGDOM.

Arthur Cronquist

Ebola virus

Ebola viruses are a group of exotic viral agents that cause a severe hemorrhagic fever disease in humans and other primates. The four known subtypes or species of Ebola viruses are Zaire, Sudan, Reston, and Côte d'Ivoire (Ivory Coast), named for the geographic locations where these viruses were first determined to cause outbreaks of disease. Ebola viruses are very closely related to, but distinct from, Marburg viruses. Collectively, these pathogenic agents make up a family of viruses known as the Filoviridae.

Infectious agent. Filoviruses have an unusual morphology, with the virus particle, or virion, appearing as long thin rods. Virions have a diameter of 0.08 micrometer and a minimal length of 0.7–0.9 μm; this length is comparable to that of a small bacterium. However, when these viruses are grown in cell cultures, they can form long filamentous and branched forms that reach 14 μm or more. This is especially true for Ebola viruses.

A filovirus virion is composed of a single species of ribonucleic acid (RNA) molecule that is bound together with special viral proteins, and this RNA–protein complex is surrounded by a membrane derived from the outer membrane of infected cells. Infectious virions are formed when the virus buds from the surface of infected cells and is released. Spiked structures on the surface of virions are formed by three molecules of a single glycoprotein and are firmly embedded in the virion membrane. These structures project from the virion and serve to recognize and attach to specific receptor molecules on the surface of susceptible cells. This recognition and binding allows the virion to penetrate into the cell. Then, with the virion free to operate within the cytoplasm, the genetic information contained in the RNA molecule directs production of new virus particles by using the cellular machinery to drive synthesis of new viral proteins and RNA. *See* RIBONUCLEIC ACID (RNA); VIRUS.

Pathogenesis. Although much is known about the agents of Ebola hemorrhagic fever disease, the ecology of Ebola viruses remains a mystery. The natural hosts of filoviruses remain unknown, and there has been little progress at unraveling the events leading to outbreaks or identifying sources of filoviruses in the wild. Fortunately, the incidence of human disease is relatively rare and has been limited to persons living in equatorial Africa or working with the infectious viruses. However, in this time of rapid transportation over large distances, the threat of agents such as Ebola viruses spreading to remote areas is

taken very seriously by public health professionals. People infected with the virus are not as contagious as, say, persons suffering from a cold or measles, and the virus is spread primarily through close contact with the body of an infected individual, his or her body fluids, or some other source of infectious material.

Diagnosis. Ebola virus hemorrhagic fever disease in humans begins with an incubation period of 4–10 days, which is followed by abrupt onset of illness. Fever, headache, weakness, and other flulike symptoms lead to a rapid deterioration in the condition of the individual. In severe cases, bleeding and the appearance of small red spots or rashes over the body indicate that the disease has affected the integrity of the circulatory system. Contrary to popular belief, individuals with Ebola virus infections do not melt or have their organs dissolve, but die as a result of a shock syndrome that usually occurs 6–9 days after the onset of symptoms. This shock is due to the inability to control vascular functions and the massive injury to body tissues.

Ebola viruses can be found in high concentrations in tissues throughout the body and are especially evident in the liver, spleen, and skin. From studies of human cases and experimentally infected animals, it appears that the immune response is impaired and that a strong cellular immune response is key to surviving infections. This immunosuppression may also be a factor in death, especially if secondary infections by normal bacterial flora ensue. No severe human disease has been associated with Reston virus infections, and it appears that this virus may be much less pathogenic for humans than it is for monkeys. *See* IMMUNOSUPPRESSION.

Epidemiology. The first described cases of Ebola virus disease occurred in 1976, when simultaneous outbreaks of two distinct subtypes occurred in northern Zaire and southern Sudan. Many of the human infections occurred in local hospitals, where close contact with fatal cases, reuse of contaminated needles, and a low standard of medical care led to most fatalities. An Asian form of Ebola virus was discovered in a type of macaque, the cynomolgus, exported from the Philippines to the United States in late 1989. It was determined that a single monkey breeding facility in the Philippines was the source of the virus. Named Reston virus after the Virginia city in which infected monkeys were first identified, it represented a new form of Ebola virus. A repeat of this incident took place in early 1992, when monkeys from the same breeding facility were shipped to Siena, Italy.

In 1994, another new species of Ebola virus was identified, associated with chimpanzee deaths in the west African country of Côte d'Ivoire. This episode signaled the reemergence of Ebola virus in Africa, which had not been seen since 1979 when an outbreak of the Sudan subtype occurred. A large outbreak of disease was identified in early May 1995 in and around the city of Kikwit, Zaire. This outbreak of a Zaire subtype was a repeat of the 1976 episode, but this time 600 mi (1000 km) to the south. Another Ebola virus outbreak took place in 1995 in Gabon and resulted in human fatalities, but the details were slow to be revealed. This virus has been isolated, and preliminary reports indicate that it may be a variant of the Zaire subtype. This virus may have also been responsible for human fatalities in Gabon reported to have occurred in early 1996. Soon after this outbreak, yet another Reston virus outbreak occurred when infected monkeys were sent to a United States holding facility in Alice, Texas. That introduction was quickly detected and stopped by quarantine and testing measures implemented after the first Reston virus episode in 1989.

Control and prevention. Outbreaks of Ebola virus disease in humans are controlled by the identification and isolation of infected individuals, implementation of barrier nursing techniques, and rapid disinfection of contaminated material. Diagnosis of Ebola virus cases is made by detecting virus proteins or RNA in blood or tissue specimens, or by detecting antibodies to the virus in the blood. Such testing is important in tracking and controlling the movement of the virus during an outbreak.

Dilute hypochlorite solutions (bleach), 3% phenolic solutions, or simple detergents (laundry or dish soap) can be used to destroy infectious virions. No known drugs have been shown to be effective in treating Ebola virus (or Marburg virus) infections, and protective vaccines against filoviruses have not been developed. However, research is being directed at developing such treatment and should provide insights into the disease mechanisms used by filoviruses.

Evolution. Filoviruses have been shown to be genetically related to two other virus families, the Paramyxoviridae (including measles virus and mumps virus) and the Rhabdoviridae (including rabies virus and vesicular stomatitis virus). These viruses evolved from a common progenitor in the very distant past. These viruses evolved into very distinct lineages, yet have maintained a similar approach to reproducing.

The evolutionary profile, or phylogeny, for Ebola viruses and Marburg viruses has been determined. Ebola viruses have evolved into a separate and very distinct lineage within the filovirus family, and the individual species of Ebola viruses have also evolved into their own sublineages. Filoviruses show a great deal of diversity, indicating that they have likely coevolved with their natural hosts over a long period of time. Ebola viruses within a given subtype did not appear to show a great deal of change over periods of many years. This genetic stability or stasis in the wild suggests that these viruses have evolved to occupy very specific ecological niches. Anthony Sanchez

Bibliography. Centers for Disease Control, Ebola-Reston virus infection among quarantined nonhuman primates—Texas, 1996, *MMWR*, 45:314–316, 1996; B. N. Fields et al. (eds.), *Fields Virology*, 3d ed., 1996; L. G. Horowitz, *Emerging Viruses: AIDS and Ebola—Nature, Accident or Intentional*, 1996; A. Sanchez et al., Reemergence of Ebola virus in Africa, *Emerg. Infect. Dis.*, 1:96–97, 1995;

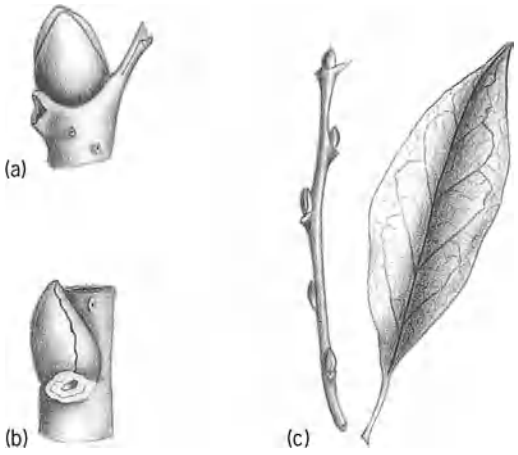
A. Sanchez et al., The virion glycoproteins of Ebola viruses are encoded in two reading frames and are expressed through transcriptional editing, *Proc. Nat. Acad. Sci. USA*, 93:3602–3607, 1996.

Ebony

A genus, *Diospyros*, of the ebony family, containing more than 250 species. Some species are important for succulent fruits, such as date plum, kaki plum, and persimmon, and several for timber, particularly the heartwood, the true ebony of commerce.

Although it is popularly supposed to be a black wood, most species have a heartwood only streaked and mottled with black. The heartwood is brittle, breaks with a conchoidal fracture, and is difficult to work, but it has long been in demand. The sapwood is white, becoming bluish or reddish when cut.

The use of ebony can be traced to the early Egyptians, who probably obtained it in Ethiopia. Ebony from India was known to the Greeks before 350 B.C. At present black ebony is used for knife handles, piano keys, finger boards of violins, hairbrush backs, inlays, and marquetry. Some of the woods called ebony, however, belong to different families, especially the pulse family, Leguminosae.



Persimmon (*Diospyros virginiana*). (*a*) Terminal bud. (*b*) Lateral bud. (*c*) Twig and leaf.

Persimmon (*D. virginiana*), of the southeastern United States, is one of numerous tropical or subtropical species (see **illus.**). Usually a medium-sized three with black chunky bark, it attains 100–125 ft (30–37 m) in height with a trunk 20–30 in. (50–75 cm) in diameter. The sapwood is in demand for the manufacture of weaving shuttles and heads of golf clubs. The fruit is sweet and edible when slightly overripe, but when immature it is extremely pungent. The species in tropical America are too small or rare to be of economic value, although several of them have black heartwood used locally for making walking sticks, inlays, and miscellaneous articles of turnery and carving. *See* FOREST AND FORESTRY; TREE.

Arthur H. Graves; Kenneth P. Davis

Ebriida

An order of flagellate Protozoa, subphylum Sarcomastigophora, class Phytamastigophorea. Two genera, *Ebria* and *Hermesinum*, remain of the once numerous order, as determined by fossil remains. They possess an internal solid siliceous skeleton forming a shallow flattened or slightly arched structure that is enclosed by clear cytoplasm. Skeleton form and structure are distinctive. These phytoflagellates do not have chromatophores. A conspicuous nucleus lies anteriorly and two long and fine flagella emerge near it. Reserve materials apparently are fat. *See* CILIA AND FLAGELLA.

The organisms are common in the inshore waters of the Atlantic Ocean and the Gulf of Mexico. However, numbers sufficient to influence the general ecology have not been observed, nor has any ingestion of particulate matter. Their only occurrence seems to be marine, although somewhat lowered salinities are certainly tolerated. *See* PHYTAMASTIGOPHOREA; PROTOZOA.

James B. Lackey

Ecdysone

The molting hormone of insects. It was isolated in crystalline form from the common silkworm (*Bombyx mori*) by A. Butenandt and P. Karlson in 1954, and its chemical structure was elucidated in 1965. It is a derivative of cholesterol and its structure can be described as $2\beta,3\beta,14\alpha$,22R,25-penta-hydroxy-Δ^7-5β-cholestene-6-one. Shortly thereafter the molting hormone of crustaceans (crustecdysone) was isolated and identified as 20-hydroxy-ecdysone (**Fig. 1**). The sample compound was identified as one of the active compounds in *B. mori* extracts by P. Hocks and coworkers, and H. Hoffmeister and coworkers proposed the name ecdysterone. A group of Japanese chemists, working on the isolation of steroids from plants, came across some similarities of the spectra of 20-hydroxy-ecdysone and their compounds inokosterone and isoinokosterone; the latter was identical with 20-hydroxy-ecdysone. Inokosterone carries the last hydroxy group in position 26 instead of 25, which reduces the biological activity by a factor of 10. Several other structurally related and biologically active steroids have been isolated from plant sources, including ecdysone, but nothing is known about a possible biological function of these steroids in plants. *See* STEROID.

Ecdysone is synthesized in insects from cholesterol; 7-dehydro-cholesterol may be an intermediate. It has been shown that ecdysone is also readily metabolized and inactivated by an enzyme system; the inactivation occurs in the living insect, as well as in homogenates. The activity of the inactivating system varies markedly during development. The most striking physiological activity of ecdysone is the induction of puffs (zones of gene activity) in giant chromosomes of the salivary glands and other organs of the midge *Chironomus*. It has been shown that 20-hydroxy-ecdysone also induces puffs. The

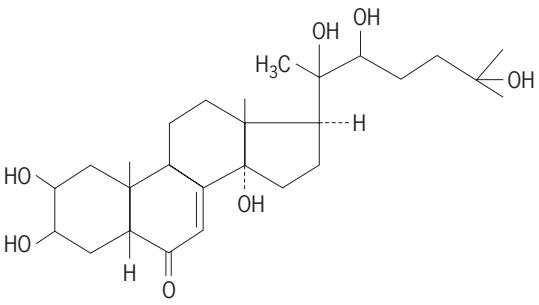


Fig. 1. Structure of 20-hydroxy-ecdysone.

induction of puffs has been visualized as the primary action of the hormone, indicating that ecdysone controls the activity of specific genes. The postulated mechanism for the action of ecdysone is shown in **Fig 2**; the primary effect is gene activation. (In the figure DNA is deoxyribonucleic acid, ATP is adenosine triphosphate, GTP is guanosine triphosphate, CTP is cytosine triphosphate, and UTP is uridine triphosphate.) The active gene produces messenger ribonucleic acid (mRNA), which in turn combines with ribosomes and becomes involved with protein synthesis. These proteins may be very specific and may be enzyme molecules. It has been shown that ecdysone stimulates the synthesis of mRNA, among which is the messenger for dopa decarboxylase. This enzyme is involved in the biosynthesis of the sclerotizing agent *N*-acetyl-dopamine. *See* CHROMOSOME; INSECT PHYSIOLOGY.

Peter Karlson

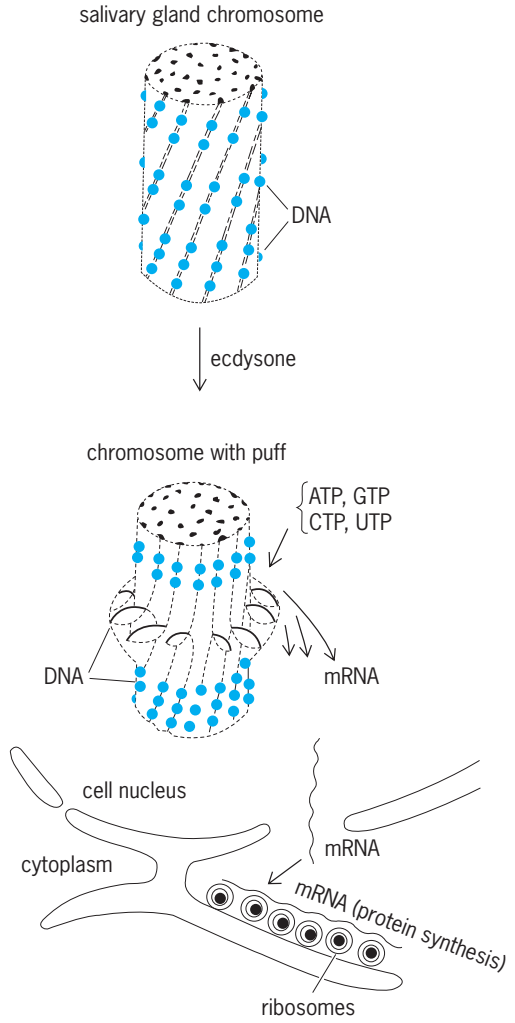


Fig. 2. Drawing of postulated mechanism of action of ecdysone in protein production in the midge *Chironomus*.

Bibliography. F. Schwalm, *Insect Morphogenesis*, 1987; F. Taylor and R. Karban (eds.), *The Evolution of Insect Life Cycles*, 1986; A. Zaslavsky, *Insect Development*, 1988.

Ecdysozoa

A major division within the animal kingdom containing the majority of animal species and comprising the phyla Priapulida, Kinorhyncha, and Loricifera (no common names), Arthropoda (insects, crustaceans, spiders, etc.), Onychophora (velvet worms), Tardigrada (water bears), and Nematoda (roundworms). These forms include animals with disparate morphologies, but their close relationship is suggested by molecular DNA sequence comparisons—especially of both large and small subunit RNA molecules—and is consistent with myosin heavy-chain sequence comparisons and with a variety of uniquely shared genomic and morphological features. Ecdysozoans were named after their habit of molting (ecdysis), unique among invertebrates, and presumably all are descended from a common molting ancestor. *See* ANIMAL KINGDOM; MOLTING (ARTHROPODA).

General characteristics. Ecdysozoa is the major subdivision of the Protostomia (see **illustration**). In most protostomes the adult mouth forms on the site of the blastopore, and the secondary body cavity (the coelom) forms by splitting between tissue layers of the mesoderm. None of the ecdysozoan phyla have a eucoelom—a secondary body cavity that contains the major body organs and acts as a fluid skeleton—but may have smaller coelomic spaces, particularly during development, that function as ducts or contain specialized organs. Fluid skeletons are furnished by primary body cavities that either are pseudocoels (in the absence of a blood circulatory system) or, in arthropods and onychophorans, are hemocoels (contain blood). In this feature they differ from most deuterostomes and lophotrochozoans (see illustration). The cuticle is largely flexible in some forms (priapulids, onychophorans) but is moderately to strongly sclerotized in others and even mineralized in some arthropods, and acts to stiffen the body wall and to serve as a site for muscle attachments, forming an exoskeleton in arthropods. Growth is accommodated by periodic molting and then secretion of a larger cuticle. None of the marine ecdysozoans has primary planktonic (free-floating) larvae, although in some arthropods several early molting stages are planktonic and serve for dispersal. *See* COELOM.

Priapulida, Kinorhyncha, and Loricifera. These phyla are pseudocoelomate worms, and all have minute

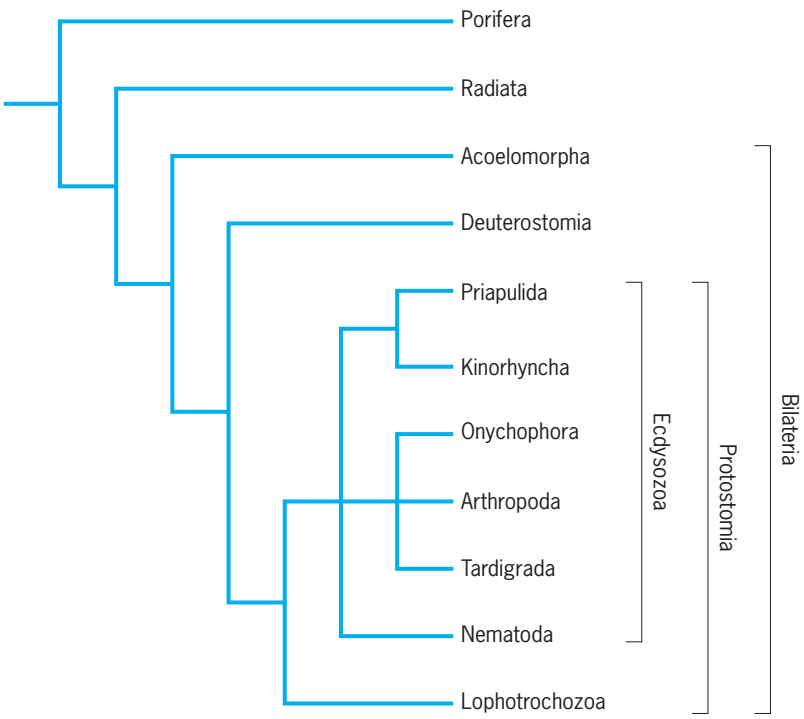
species that live interstitially in marine sediments; loriciferans are attached to sand grains, while the others are free-living. The cuticles bear secreted spines or plates that are sensory or aid in burrowing. Some priapulids also grow to much larger sizes [to 20 cm (8 in.)] and form prominent burrow structures. Kinorhynchs have a segmented cuticular exoskeleton and longitudinal muscle system, but the other two phyla are unsegmented. *See* KINORHYNCHA; LORICIFERA; PRIAPULIDA.

Arthropoda, Onychophora, and Tardigrada. These phyla are more closely related to each other than to other ecdysozoans on molecular evidence. All three are segmented to some extent and possess limbs. In onychophorans and tardigrades, the cuticle is rather flexible, and the limbs (lobopods) contain pseudocoelomic cavities that provide turgor. In arthropods, however, the cuticle is stiffened to form an exoskeleton, and the limbs act as levers operated by intrinsic muscles. All these phyla were originally marine but have their chief diversity in terrestrial habitats today; the onychophora are now entirely nonmarine. *See* ARTHROPODA; ONYCHOPHORA; TARDIGRADA.

Nematoda. This phylum consists of small, roundish worms that are unsegmented and pseudocoelomate, though minute species may essentially lack any body cavity. Many nematodes are parasitic, and one highly derived parasitic group, the Nematomorpha (horsehair worms), is sometimes considered as a phylum of its own. *See* NEMATA (NEMATODA); NEMATOMORPHA.

Evolution: diversity. The earliest fossils that are very likely to be ecdysozoans are Early Cambrian trails and feeding marks left by scratching limbs, probably dating 525–520 Ma (million years ago), although burrows from as early as about 540 Ma could have been formed by priapulids. Body fossils of ecdysozoans appear around 530–525 Ma, represented then by stem groups (extinct branches) of priapulids, arthropods, and onychophorans, and by extinct arthropod allies that are segmented but evidently lack jointed appendages; some of these are quite large by Cambrian standards, up to 50 cm (20 in.) long. There is also an extinct Early Cambrian vermiform group, the paleoscolecids, which may be pseudocoelomate ecdysozoans. Tardigrades are first known from about 510–500 Ma, while the other phyla do not have body fossil records.

Ecdysozoans dominate the early animal fossil record, with a great disparity of form (mostly extinct variants of arthropod architectures) and in terms of species richness. During the Cambrian (543–500 Ma), over 70% of known animal species are trilobites, an extinct arthropod class. Ecdysozoans continued to be prominent in the seas until today, where they are now challenged for dominance only by Mollusca among invertebrate phyla. The earliest indications of animal fossils from terrestrial environments are traces, burrows in soils deposited near 410 Ma; body fossil assemblages appearing near 400 Ma and slightly later are almost entirely of arthropods, including millipedes, springtails, and rarely insects; they are dominated by detritus feeders and predators. Today, arthropod species richness, especially of insects, eclipses that of all other animal phyla in terrestrial habitats. For insects, the small body sizes and modular body plans within a secreted, segmented exoskeleton and with jointed appendages that can be specialized individually—together with the habitat heterogeneity afforded by the rise of diverse global plant communities—have produced unique opportunities for evolutionary radiations, which seem to have been fully exploited. *See* FOSSIL; TRILOBITA.



Relationships among phylum-level taxa within the Ecdysozoa, and the relation of the Ecdysozoa to other major animal groups.

Evolution: genomics. The two best-known model organisms for evolutionary genetic and developmental studies, the fly *Drosophila* (Arthropoda) and the roundworm *Caenorhabditis* (Nematoda), are ecdysozoans. Understanding of the genetic processes underpinning body plan development has been worked out largely in these forms, which have both segmented and unsegmented architectures, respectively. As the key genes regulating these processes are widely distributed among other animal groups as well, including humans, comparative studies using the ecdysozoan models are helping to clarify the evolution of animal genomes during the branchings of the tree of life. *See* ANIMAL EVOLUTION; ORGANIC EVOLUTION.

James W. Valentine

Bibliography. R. C. Brusca and G. J. Brusca, *Invertebrates*, Sinauer Associates, Sunderland, MA, 2003; J. Cracraft and M. J. Donoghue (eds.), *Assembling the Tree of Life*, Oxford University Press, 2004; R. A. Fortey and R. H. Thomas (eds.), *Arthropod Relationships*, Chapman and Hall, 1998; M. W. Kirschner and J. C. Gerhart, *The Plausibility of Life*, Yale University Press, 2005; J. W. Valentine, *On the Origin of Phyla*, University of Chicago Press, 2004.

Echinacea

A superorder of Euechinoidea, having a rigid test, the periproct within the apical system, keeled teeth, a complete perignathic girdle, and branchial slits. J. Durham and R. Melville (1957) include five orders in this group. There were formerly distributed among the Stirodonta and Camarodonta in the classification of R. Jackson (1912). *See* ECHINODERMATA; ECHINOIDA; EUECHINOIDEA; HEMICIDAROIDA; PHYMOSOMATOIDA; SALENIOIDA; TEMNOPLEUROIDA.

Howard B. Fell

Echinodermata

A phylum of exclusively marine animals with a peculiar body architecture dominated by a five-part radial symmetry. Echinodermata [from the Latin *echinus* (spine) + *dermis* (skin: "spiny skins")] include the sea stars, sea urchins, and related animals. The body wall contains an endoskeleton of numerous plates (ossicles) composed of calcium carbonate in the form of calcite and frequently supporting spines. The plates may be tightly interlocked or loosely associated. The spines may protrude through the outer epithelium and are often used for defense. The skeletal plates of the body wall, together with their closely associated connective tissues and muscles, form a tough and sometimes rigid test (hard shell) which encloses the large coelom. A unique water-vascular system is involved in locomotion, respiration, food gathering, and sensory perception. This system is evident outside the body as five rows of fluid-filled tube feet. Within the body wall lie the ducts and fluid reservoirs necessary to protract and retract the tube feet by hydrostatic pressure. The nervous system of these headless animals arises from the embryonic ectoderm and consists of a ring around the mouth with connecting nerve cords associated with the rows of tube feet. There may also be diffuse nerve plexuses, with light-sensing organs, lying below the outer epithelium. The coelom houses the alimentary canal and associated organs and, in most groups, the reproductive organs. The body may be essentially star-shaped or globoid. The five rows of tube feet define areas known as ambulacra, ambs, or radii; areas of the body between the rows of tube feet are interambulacra, interambs, or interradii.

The larvae are usually planktonic (free-floating) with a bilateral symmetry, but the adults are usually sedentary and benthic (bottom-dwelling in a marine environment). They inhabit all oceans, ranging from the shores to the greatest ocean depths.

The phylum comprises about 7000 existing species. The Echinodermata have a good fossil record with about 13,000 fossil species. They first appeared in the Early Cambrian and have been evolving, since pre-Cambrian time, for well over 600 million years. During this time several differing body plans have arisen. The surviving groups show few resemblances to the original stock. The existing representatives fall into five, possibly six, classes: Crinoidea (sea lilies and feather stars); Asteroidea (sea stars); Ophiuroidea (brittle stars); Echinoidea (sea urchins, sand dollars, and heart urchins); and Holothuroidea (sea cucumbers). The recently described class Concentricycloidea (sea daisies) has been referred to the Asteroidea by some experts and retained as a distinct class by others.

The outline of classification for the phylum is shown below. Classes with living representatives are elaborated to the level of subclass or order.

Phylum Echinodermata
"Homalozoans"
Class: Ctenocystoidea
Stylophora
Homostelea
Homoiostelea
"Crinozoans"
Class: Eocrinoidea
Paracrinoidea
Rhombifera
Diploporita
Crinoidea
Blastoidea
Coronoidea
Parablastoidea
Edrioblastoidea
Subclass: Inadunata
Camerata
Flexibilia
Articulata
Order: Millericrinida
Cyrtocrinida
Bourgeticrinida
Isocrinida
Uintacrinida
Roveacrinida
Comatulida
"Asterozoans"
Class: Somasteroidea
Asteroidea
Order: Platyasterida
Trichasteropsida
Paxillosida
Notomyotida
Spinulosida
Valvatida
Velatida
Forcipulatida
Class: Concentricycloidea
Order: Peripodida
Class: Ophiuroidea
Order: Stenurida
Oegophiurida
Phrynophiurida
Ophiurida
"Echinozoans"
Class: Helicoplacoidea
Camptostromatoidea
Edrioasteroidea
Cyclocystoidea
Ophiocistioidea
Echinoidea

Subclass: Perischoechinoidea
Order: Bothriocidaroida
Echinocystitoida
Subclass: Cidaroidea
Order: Cidaroida
Subclass: Euechinoidea
Order: Diadematoida
Echinothurioida
Pedinoida
Arbacioida
Echinoida
Phymosomatoida
Salenioida
Temnopleuroida
Clypeasteroida
Holectypoida
Cassiduloida
Spatangoida
Class: Holothuroidea
Order: Dendrochirotida
Dactylochirotida
Aspidochirotida
Elasipodida
Molpadiida
Apodida

Morphology

Despite displaying an astonishing variety of body forms, existing echinoderms share common anatomical features: symmetry, body wall, skeleton, nervous system, coelom, alimentary system, reproductive organs, and water-vascular system.

Symmetry. Adult echinoderms usually show pentamerism, a five-part radial symmetry in which there are normally five axes radiating out from a central point in the body. The mouth may lie on the underside, and the anus may lie on top of the animal. There are numerous exceptions: in the crinoids the mouth and anus lie on the same side, and in the holothuroids the mouth and anus lie at opposite ends of a cylindrical body. The side of the animal bearing the mouth is termed the oral side, and the side away from the mouth is the aboral side. The relationship of the mouth with the substratum was at one time considered important so that those echinoderms which were essentially sessile, with a mouth facing away from the substratum, were classified as subphylum Pelmatozoa, while those mobile forms in which the mouth faced the substrate were classified as Eleutherozoa. These differing life habits have little bearing upon formal echinoderm classification, but the terms "pelmatozoan" and "eleutherozoan" are useful descriptors of lifestyles in the phylum.

It has been suggested that five-part symmetry in echinoderms confers strength upon the skeleton supporting the body wall. Pentamerism was not an original feature of the echinoderms. Certain early groups such as the carpoids did not show it, and modern echinoderm larvae have bilateral symmetry.

Body wall. The body wall is cloaked in a thin epidermis, a layer of epithelial cells. The external layer is often ciliated, and the electron microscope has shown that these cells have an outer membrane that bears a great number of minute tubular extensions like microvilli. This arrangement of the epidermis is similar in appearance to the absorptive epithelium of many intestinal tissues. Below the epidermis, the dermis contains the skeletal elements. Internal to the dermis are the body-wall muscles and connective tissue. In the crinoids, ophiuroids, some asteroids, and some echinoids, the skeletal elements make up a great part of the body-wall volume. In the more flexible asteroids, the spherical echinoids, and the holothuroids, the volume of these elements is reduced. The inner surface of the body wall is also lined with epithelium and borders the coelom.

Skeleton. The skeleton is unique and consistent in all groups. Almost every plate, spine, or ossicle is composed of a single calcite crystal. During development each plate grows as a result of calcite secretion by a group of cells. The plate grows like a three-dimensional lattice (stereom) to produce a reticulate (having or resembling a network of fibers or lines) ossicle, with the tissue that secretes and maintains it (stroma) occupying the spaces within (**Fig. 1**).

Sensory and neuromuscular system. Echinoderms have simple sensory systems. The nervous system is ectodermal in origin and comprises a radial nerve cord lying along each ambulacrum. Each radial nerve connects with the circumesophageal nerve ring, and it is believed that interradial coordination takes place via this ring. The absence of a head means that sensory receptors and large areas of integrative nervous elements are not aggregated in one place. The nervous system is essentially diffuse, although specialized sense organs, such as the recently discovered "eyes" of certain brittle stars, occur. It appears that the circumoral ring and radial nerve cords take

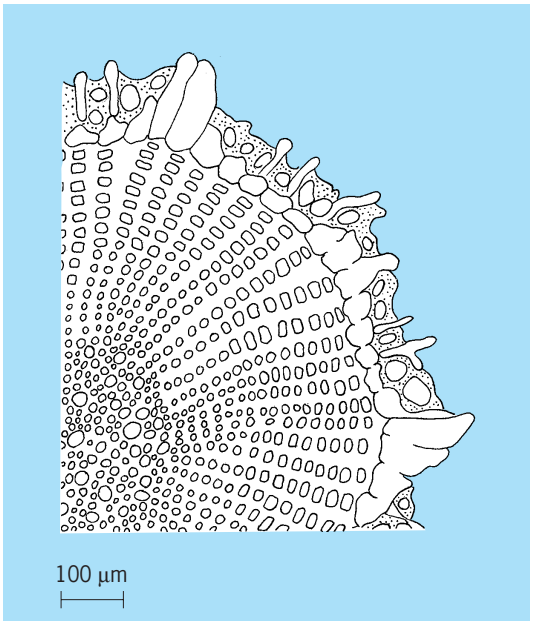


Fig. 1. Cross section through the spine of an echinoid showing the microscopic structure of the skeleton. The stereom forms a continuous mesh, and the stroma that secretes it lies in the interspaces which, in this case, are arranged in radial rows.

responsibility for gross activites such as posture, coordinated locomotion, and food collecting, but in the asteroids and the echinoids there are well-developed basiepithelial nerve networks that coordinate the numerous test appendages. Echinoderm muscles are principally smooth, but striated muscles have been detected in a few instances. *See* NERVOUS SYSTEM (INVERTEBRATE).

Coelom. The coelom is extensive and usually arises by enterocoely, or pouching, from the gut; but when there is direct development, with larval stages greatly reduced or absent, it may be schizocoelic, involving splitting in the mesoderm. The perivisceral coelom surrounds the viscera. During development, part of the coelom forms the water-vascular system, so the spaces within the tube feet and vessels are coelomic. *See* COELOM.

Alimentary system. In general the mouth is situated on the same side as the tube feet. From it arises the alimentary canal, which runs through the body to the anus (absent in the ophiuroids and some asteroids). The gut is lined by endoderm, and in some groups digestive caeca (blind pouches at the beginning of the large intestine) increase the area for absorption.

Particulate food is collected by the tube feet in many crinoids, ophiuroids, and holothuroids. Most asteroids are carnivores and engulf the prey, or evert part of the stomach over it, while some echinoids have well-developed chewing teeth for rasping at encrusting plants and animals. Many holothuroids, some echinoids, and a few asteroids ingest mud or sand, obtaining sustenance from associated organic material such as bacteria and diatoms. *See* DIGESTION (INVERTEBRATE).

Reproductive system. The reproductive organs lie within the coelom in the interradial position, except in the crinoids in which they arise on the arms. In most echinoderms there are five compact gonads, but the holothuroids have just one, and certain echinoids may have two, three, or four. The sexes are usually separate, The gonads discharge gametes by short ducts into the surrounding seawater. In some echinoderms, especially in polar regions, fertilized gametes are retained, and the young develop in special sacs in the body wall.

Water-vascular system. This is a multifunctional fluid-filled coelomic system that probably evolved first as a respiratory system but later took on increasingly important roles in food collection and locomotion (**Fig. 2**). It comprises a number of protrusible, hollow, tentacle-like tube feet, arranged externally in rows along each radial area of the body. These areas, composed of single to multiple rows of tube feet, are termed ambulacra. A small modified test ossicle, the madreporite, perforated by many irregularly shaped pores, connects with the water-vascular system via a calcified stone canal. It was thought that the water-vascular system opened directly to the surrounding seawater via the madreporite so that extra fluid could be drawn in to replenish the water-vascular fluid. The fluid volume of the water-vascular system may be more constant than was previously believed, and relatively little fluid passes in or out of the madreporite. Each tube foot is associated with a small bulbous reservoir inside the body. The feet are extended hydrostatically, and retracted by longitudinal muscles in their walls.

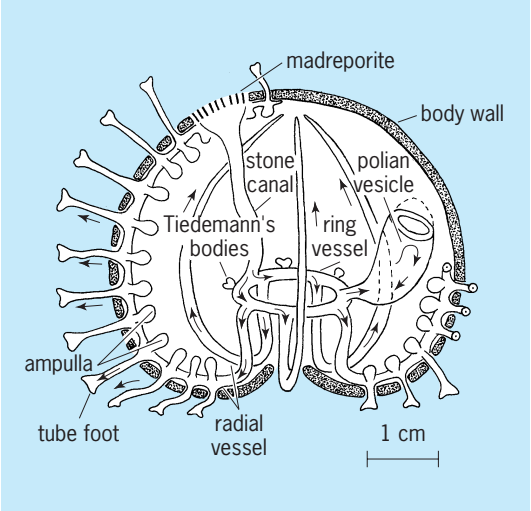


Fig. 2. Diagram of the water-vascular system in an echinoid. Arrows show direction of fluid flow.

The form and the activity of tube feet vary greatly between the groups of echinoderms. In the crinoids they are fine, tapering structures suited for food collection and having no locomotory role. In the brittle stars they lack terminal suckers and are more important for food gathering than locomotion. In some asteroids, particularly the burrowers, they are tapering and pointed, while in others they are suckered and capable of adhering to the substratum. This is also the case in regular echinoids (which have no front or back end and can move in any direction), but in irregular echinoids (which have a definite front and back and do move in a particular direction) the tube feet are modified to suit the burrowing way of life. In holothuroids the tube feet surrounding the mouth are modified as tentacles of varying form and complexity and are used for food collecting. The rest may be present as locomotor or respiratory organs. In a few groups, such as the apodous and molpadiid holothurians, tube feet are missing from the body wall. When tube feet serve an important locomotor function, postural muscles govern the stepping movements.

Each tube foot is under the control of the radial nerve cord and is connected by a short lateral branch to the radial water-vascular canal as well as to the individual ampullae. All the radial canals are linked by the circumoral ring canal so that water-vascular fluid may be withdrawn from one part of the system to be supplied to another.

Saclike structures are associated with the ring canal. These are polian vesicles, which act as reservoirs for water-vascular fluid, and Tiedemann's bodies, which act as glands in which wandering coelomocytes (coelom corpuscles) are formed.

Embryology

Most echinoderms have an indirect development with a prolonged, food-gathering larval stage (**Fig. 3**). The larva feeds, usually on microscopic phytoplankton, using a ciliary mechanism. Two well-marked larval types occur: the pluteus group, with long-armed, bilaterally symmetrical, easel-shaped forms, common to ophiuroids and echinoids; and the auricularia group, barrel-shaped forms with a winding ciliated band which may be produced into lobes. The latter group is common to asteroids and holothuroids. In most asteroids the auricularia stage is followed by a similar but more complex larva, the bipinnaria, or sometimes also by an anchored final larval stage, the brachiolaria. Surviving crinoids have essentially a direct development, sometimes with a simple yolky larva, a vitellaria, which does not feed. This is also the case for several members of other extant echinoderm groups. *See* INVERTEBRATE EMBRYOLOGY.

Phylogeny

The auricularia larva presents close and striking resemblances to the tornaria larva of some enteropneusts, and the enterocoelous development parallels that in primitive chordates. Hence echinoderms and chordates have long been regarded as related. However, the significance of similarities in the larvae of echinoderms and protochordates must be viewed in the context of molecular, morphological, and paleontological research. It seems likely that the similarities between the larvae of ophiuroids and echinoids, and asteroids and holothuroids, are due to convergent evolution and not to common evolutionary origins. The results of paleontology, molecular

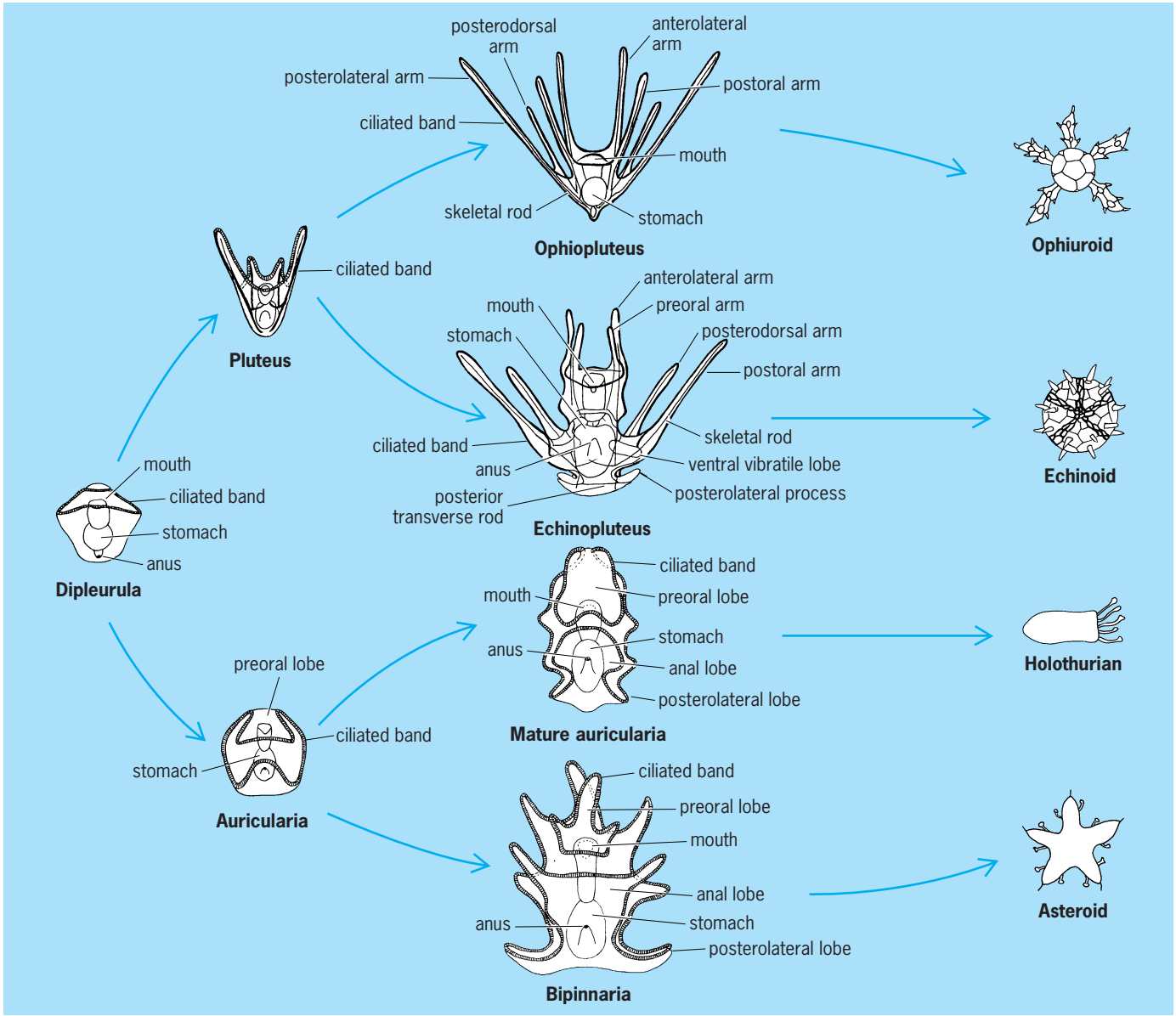


Fig. 3. General scheme indicating relationships of bilaterally symmetrical echinoderm larvae. (*After H. B. Fell*)

studies, and morphology indicate that ophiuroids and asteroids are closely related. Therefore it follows that within the phylum larval similarities do not indicate phylogenetic affinities. It is inadvisable to try to extrapolate beyond the phylum, so as to infer phylogenetic affinity between hemichordates and echinoderms solely on the ground that the auricularia resembles the tornaria. E. Marcus expressed the opinion that indirect development, with possession of pelagic larvae, must be prototypical for echinoderms and protochordates, and the asteroids and ophiuroids must be closely related, and therefore broad phylogenetic conclusions cannot be drawn on the basis of their larvae. L. Hyman grouped the extant echinoderms as their larval similarities suggest, and concluded that the arrangement adopted by paleontologists must be wrong. H. B. Fell and later authors have reaffirmed that the paleontological evidence overrules nebulous embryological considerations. David L. Pawson; Andrew C. Campbell

Ecology and Feeding

Echinoderms occur everywhere in the world's oceans, but they are usually rare in areas where the salinity of the water is greatly reduced or where pollution levels are significant. In many areas, echinoderms are dominant invertebrates in terms of numbers or biomass, and locally sea stars can be the top predators. On a rocky shore, predatory sea stars and vegetarian sea urchins may be common on and under rocks. Brittle stars are typically concealed under rocks; they may emerge at night to forage for small organisms. In deeper water, where conditions are suitable for suspension feeding, huge aggregations of brittle stars or feather stars may occur, with their arms extended into the water for feeding. Echinoderms flourish on and near coral reefs, and it is in these tropical areas that they can achieve their greatest diversity. In polar regions, especially around Antarctica, echinoderms, along with sponges, dominate in most habitats (**Fig. 4**). Sandy to muddy bottoms may be populated with large numbers of burrowing detritus-eating sea urchins, long-armed brittle stars, and sea cucumbers. In such habitats, burrowing sea cucumbers may extend their sticky tentacles into the surrounding water to capture small drifting organisms. In deeper water, below 100 m (330 ft), stalked sea lilies can be common on hard substrates, efficiently suspension-feeding with their arms deployed in the form of a parabolic bowl. In the deep sea, echinoderms can also be dominant; on abyssal (relating to great ocean depths) plains, sediment-swallowing sea cucumbers can make up more than 95% of the total biomass on the sediment surface. Most echinoderm groups occur in the deep sea to depths in excess of 9000 m (5.6 mi). David L. Pawson

Fossils

Echinodermata are an especially important group of fossil invertebrates. Except for reworked fragments found in some nonmarine deposits, echinoderm remains occur exclusively in strata laid down on sea bottoms—chiefly those of shallow seas. Many of these deposits, ranging in age from Cambrian to late Tertiary, contain abundant echinoderm fossils. In addition, fossil echinoderms undoubtedly are widely distributed beneath all oceans, although remains of these organisms in deep-water sediments, even of recent origin, are virtually unknown because of their inaccessibility.

Fig. 4. Echinoderms in the deep sea near Antarctica at a depth of 595 m (1950 ft). Three sea cucumbers (*Scotoplanes globosa*) are feeding on seafloor sediments. At bottom right is a brittle star, probably *Ophiomusium* species. Scattered on the seafloor but barely visible are several feather stars. (*Courtesy of the U.S. National Science Foundation*)

The echinoderms are well adapted to preservation as fossils owing to their abundant calcareous skeletal elements. Paleontological importance of the group is explained partly by this fitness but more by the diversity of their kinds, the generally short-lived existence and wide geographic distribution of most recognized taxonomic units, and the clearness with which evolutionary trends can be defined.

Classification and Description

For some years, fossil and living echinoderms have been arranged into four or five subphyla. As formal groupings, the subphyla have been largely abandoned, for it is recognized that the subphylum definitions and distinctions are artificial rather than truly reflecting phylogeny. However, the subphylum names are useful for grouping the various echinoderm classes, and they are used here in an informal sense: homalozoans, crinozoans, echinozoans, and asterozoans. Similarly, the terms eleutherozoan (free-moving) and pelmatozoan (sedentary or sessile), previously used in formal classifications, are useful today in categorizing the lifestyles of echinoderms. *See* ELEUTHEROZOA; PELMATOZOA.

Homalozoans. The distinctive features that separate the extinct homalozoans (or carpoids) from all other Echinodermata are a complete lack of radial

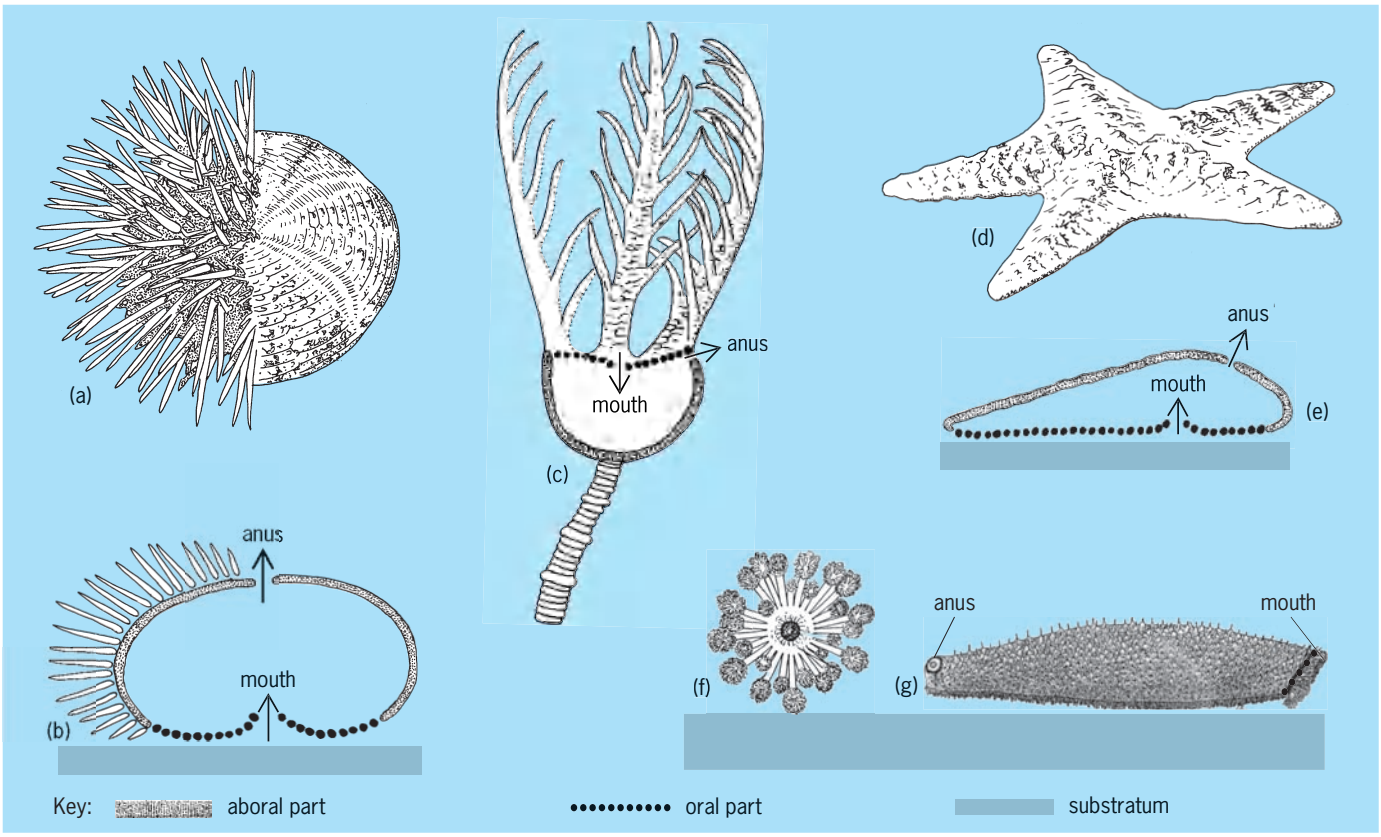


Fig. 5. Representative types of echinoderms. (*a*) Regular echinoid (*Lytechinus*, Recent), oblique aboral view, right half with spines removed, showing two ambulacra and three interambulacra. (*b*) Diagrammatic section through *a*, showing the stoutly built test. (*c*) Crinoid showing the stem and three arms, calyx sectioned. (*d*) Asteroid (*Dermasterias*, Recent), aboral oblique view. (*e*) Diagrammatic section through the ray at left in *d* and opposite the interambulacrum. (*f*) Holothuroid, oral view showing tentacles around the mouth. (*g*) Holothurian, lateral view.

symmetry in arrangement of skeletal parts and the flattened body form, from which one or more specialized slender appendages may extend. Plates enclosing the body are irregular in shape and size and vary in number between forms. There are four quite dissimilar classes of carpoids: Homostelea, Ctenocystoidea, Stylophora, and Homoiostela. Some experts, especially R. Jefferies, believe that some of the homalozoans, the Stylophora in particular, are not echinoderms at all, but are "calcichordates," groups forming the stem of the chordates. There has been considerable debate in the scientific literature about this matter, with persuasive evidence presented on both sides. Most paleontologists tend toward the idea that all homalozoans are true echinoderms. The homalozoans range in age from earliest Cambrian to Carboniferous. *See* CARPOIDS; HOMALOZOA.

Crinozoans. The crinozoans evidently include types of echinoderms least modified from the ancient progenitors of the phylum. Crinozoans range from Cambrian to Recent, but only the Crinoidea, abundant in Paleozoic formations, have survived to the present. In some classifications, the Crinoidea and Paracrinoidea are treated as the subphylum Crinozoa in the strict sense, and the other classes are grouped together in the subphylum Blastozoa. A characteristic crinozoan feature is orientation of the body with the oral (ventral) side directed upward and aboral (dorsal) side downward; thus, although commonly used, the designations ventral and dorsal are not suited to these particular echinoderms (**Figs. 5***c* and **6**). There are nine classes of crinozoans:

(1) Eocrinoidea (the crinozoan stem group) and (2) Paracrinoidea are crinozoan groups attached by a stalk. They have ovoid bodies enclosed by irregularly arranged plates that commonly lack well-developed radial symmetry and possess a combination of characters typical of the two following classes. (3) Diploporita and (4) Rhombifera (previously grouped in the polyphyletic Cystoidea, with saclike bodies, and each possibly polyphyletic) are a diverse group of primitive crinozoans distinguished especially by pores or tubular canals that penetrate irregularly arranged plates enclosing the body. Attachment to the substrate was by means of a stem composed of superposed discoid plates. (5) Crinoidea (sea lilies) include the most abundant crinozoans, most of the fossil forms anchored to the sea bottom by means of a short to very long stem and distinguished by upwardly directed arms which functioned for gathering food. Some ancient and most modern crinoids are stemless, known as feather stars, and adapted for a free-living existence. True radial symmetry and regularity of skeletal construction are attributes that distinguish crinoids. (6) Blastoidea (budlike forms) are small to

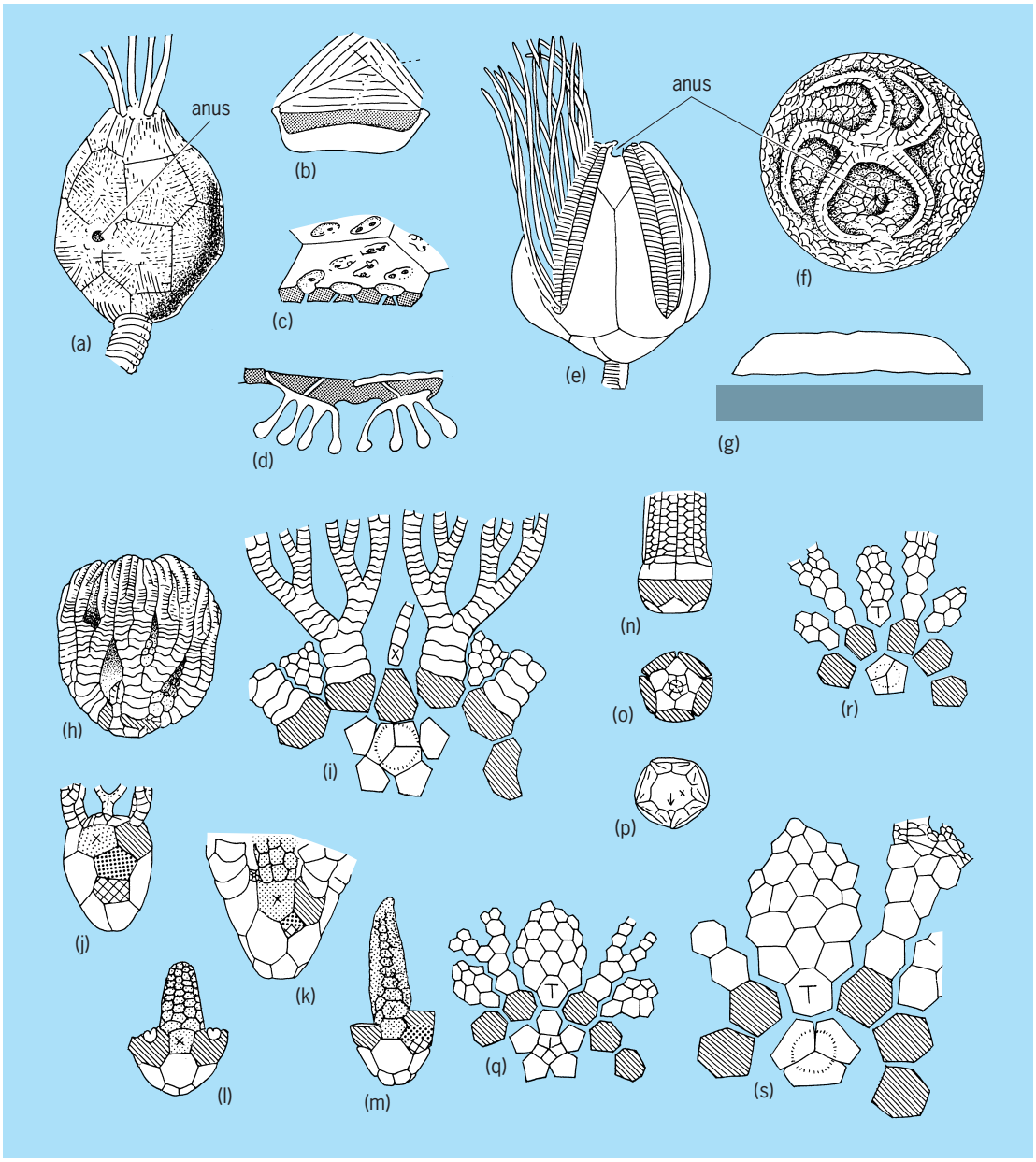


Fig. 6. Fossil pelmatozoan echinoderms. (*a*) Rhombiferan cystoid (*Echinoencrinus*, Ordovician). (*b*) Diagrammatic oblique view and section through a cystoid pore rhomb (half of one rhomb toward front bisected by a suture between plates). (*c*) Oblique view and section of a diploporitan cystoid plate. (*d*) Section through the ambulacrum of a blastoid (*Pentremites*, Mississippian) showing hydrospires beneath the lancet and side plates. (*e*) Blastoid (*Pentremites*) with brachioles restored along the side of one ambulacrum. (*f*) Edrioasteroid (*Carneyella*, Ordovician), oral view. (*g*) Section through a stalkless edrioasteroid. (*h*, *i*) Flexible crinoid (*Taxocrinus*, Mississippian), posterior side and partial plate diagram showing the right posterior plane of bilateral symmetry defined by the infrabasal circlet. (*j–p*) Inadunate crinoids, some showing anal sacs and parts of arms: (*j*) *Carabocrinus*, Ordovician; (*k*) *Botryocrinus*, Devonian; (*l*) *Cyathocrinites*, Mississippian; (*m*) *Cupulocrinus*, Ordivician; (*n–p*) *Delocrinus*, Pennsylvanian. (*q*) Diplobathrid camerate crinoid (*Ptychocrinus*, Ordovician), plate diagram. (*r*, *s*) Monobathrid camerates: (*r*) *Macrostylocrinus*, Devonian; (*s*) *Periechocrinites*, Mississippian, plate diagrams showing noteworthy distinctions in basal and radial circlets. (*Parts f, g after L. H. Hyman, The Invertebrates, vol. 4: Echinodermata, McGraw-Hill, 1955*)

medium-sized crinozoans having a highly developed radial symmetry and very regular arrangement of the few plates surrounding the body. The body was attached to the substrate by a very slender stem, which had a circular cross section. (7) Coronoidea is a small group of middle Paleozoic crinozoans, similar to blastoids and probably ancestral to them. Coronoids differ in the structure of the ambulacral plate and possess erect, as opposed to recumbent, ambulacra. (8) Parablastoidea is another small, principally Ordovician class of crinozoans with cystoid-like thecae but with uniserial ambulacra. (9) Edrioblastoidea comprises a single Ordovician genus with a theca that is blastoid-like in shape and symmetry but lacks brachioles or hydrospires (or pore rhombs). *See* CRINOZOA.

Eocrinoids. This small group of Cambrian to Silurian crinozoans combines characteristics of

cystoids and crinoids, yet differs significantly from both. They resemble cystoids in their mode of branching of the ambulacral grooves and ventrolateral location of the anus but lack thecal pores or distinct pore rhombs; they are like crinoids in plate structure and similarity of plate arrangement in the calyx (plated body). The eocrinoids have stems and unbranched or bifurcating arms. They were probably ancestral to all other crinozoans. *See* EOCRINOIDEA.

Paracrinoids. Paracrinoidea are stem-bearing echinoderms that also combine features of cystoids and crinoids, but in a manner quite unlike that of the eocrinoids. The paracrinoids, now known only from Middle Ordovician to Lower Silurian deposits, have numerous plates of the calyx that are not arranged in series and that lack a ventrally differentiated area corresponding to the tegmen of crinoids. The plates have a cystoid-like pore structure, but the arms are comparable to those of crinoids.

Cystoids. The Ordovician to Devonian cystoids, comprising the classes Diploporita and Rhombifera, are extinct crinozoans of globose, subcylindrical, or flattened ellipsoidal form that are characterized mostly by irregularity of the body plates (Fig. 6*a*) and a short, weak stem. A variable number of slender appendages (brachioles) on the upper side of the calyx, supplemented by ambulacral grooves, served for gathering food. The mouth was located centrally at the summit of the calyx, and an anus was fairly well down on one of the sides, which accordingly is defined as posterior. Other small orifices, interpreted as hydropore and gonopore, occur near the anus. In one class, Diploporita, the numerous calyx plates are perforated by numerous pairs of minute tubular openings that allow seawater to circulate through them (Fig. 6*c*). Remaining cystoids are named Rhombifera (rhomb-bearing forms) because some or all of their relatively few calyx plates are penetrated by rhomb-shaped groups of tubes running parallel to each other and to the plate surfaces (Fig. 6*a*). The two halves of any rhomb lie on adjoining plates, with the tubes crossing the plate boundaries at right angles (Fig. 6*b*). The pattern of calyx plates suggests that the blastoids and crinoids may be descendants of rhombiferan cystoids. *See* DIPLOPORITA; RHOMBIFERA.

Crinoids. In terms of abundance of fossil remains, the crinoids outrank all other echinoderms combined (Fig. 5*c*). There are about 5000 extinct species and at least 700 living kinds. Some Paleozoic deposits hundreds of feet thick in areas measured in hundreds of square miles are largely composed of fossil remains of trillions of crinoids. *See* CRINOIDEA.

Skeletal features. Although many modern and some ancient crinoids are stemless as adults, this group of pelmatozoans typically is attached to the sea bottom by a more or less elongate stem composed of superposed calcareous discs (columnals) perforated centrally by a circular, pentagonal, or pentastellate canal. At the opposite extremity of the stem, which exceptionally may be 15 m (50 ft) tall, is the crinoid body, encased in regularly arranged plates and surmounted by branched or unbranched free-moving arms. The conjoined plates below the free arms make up the so-called dorsal cup; this cup, along with the plates of the ventral surface composing the tegmen, makes up the crinoid calyx (Fig. 6*h–s*). The calyx and its attached arms are termed the crown. The mouth is located on the ventral surface. The anus lies on the tegmen, is raised above it on an anal sac (Fig. 6*k–m*), or may lie on the side of the dorsal cup. The posterior side of the crinoid is defined by the position of the anus in one of the interrays or by extra plates introduced in such a position on one side of the dorsal cup; the ray opposite to the posterior interray is defined as anterior. This establishes a plane of bilateral symmetry that more of less modifies the fundamental radial symmetry of the crinoid (**Fig. 7***a*).

The plates of the crinoid dorsal cup are arranged in a regular pattern of successive circlets, but with differences in various groups that furnish a basis for classification. Each circlet normally contains five plates. At the base of each ray is a radial plate. Beneath the circlet of radials are five basals that are interradial in position; some crinoids have a still lower circlet of infrabasals that alternate with the basals and hence occur in radial position. Crinoids with a single circlet of plates below the radials are termed monocyclic (Fig. 6*r* and *s*), and those with two circlets are dicyclic (Fig. 6*h–q*).

In some crinoids the tegmen is stoutly constructed of small, irregularly arranged plates, while in others it consists of a flexible, leathery integument that may be studded with calcareous ossicles. Five subequal oral plates larger than others of the tegmen may occur interradially around the mouth. The arms of crinoids are extremely variable in plan and construction. They may be unbranched or moderately to highly branched, with or without very numerous tiny branchlets called pinnules, and composed of a single or double series of arm plates (brachials). These characters, along with the mode of articulation among plates of the rays, are important also in classification.

Main types. Four subclasses of crinoids are recognized, three of them distributed from Ordovician to Permian; one of the three (Inadunata) persists to Middle Triassic. The fourth subclass (Articulata) ranges from the Mesozoic to present day.

The Inadunata ("not united" forms, referring to lack of incorporation of lower arm plates in dorsal cup) are crinoids with a relatively small dorsal cup containing one or two circlets of plates below the radials and having the arms entirely free above the cup (Fig. 6*j–p*). They include 1750 or more species that exhibit great variety in form and evolutionary trends. In a majority the anteroposterior plane of bilateral symmetry is well marked, without other deviation from a regular pentameral plan (Fig. 7*a*), but in one group (superfamily Homocrinacea) a surprising degree of bilateral symmetry was developed in the plane of the left anterior ray (Fig. 7*c*). Some of these crinoids have a strongly downbent crown that was hinged on the summit of the stem. In another group (superfamily Heterocrinacea) a subordinate plane of

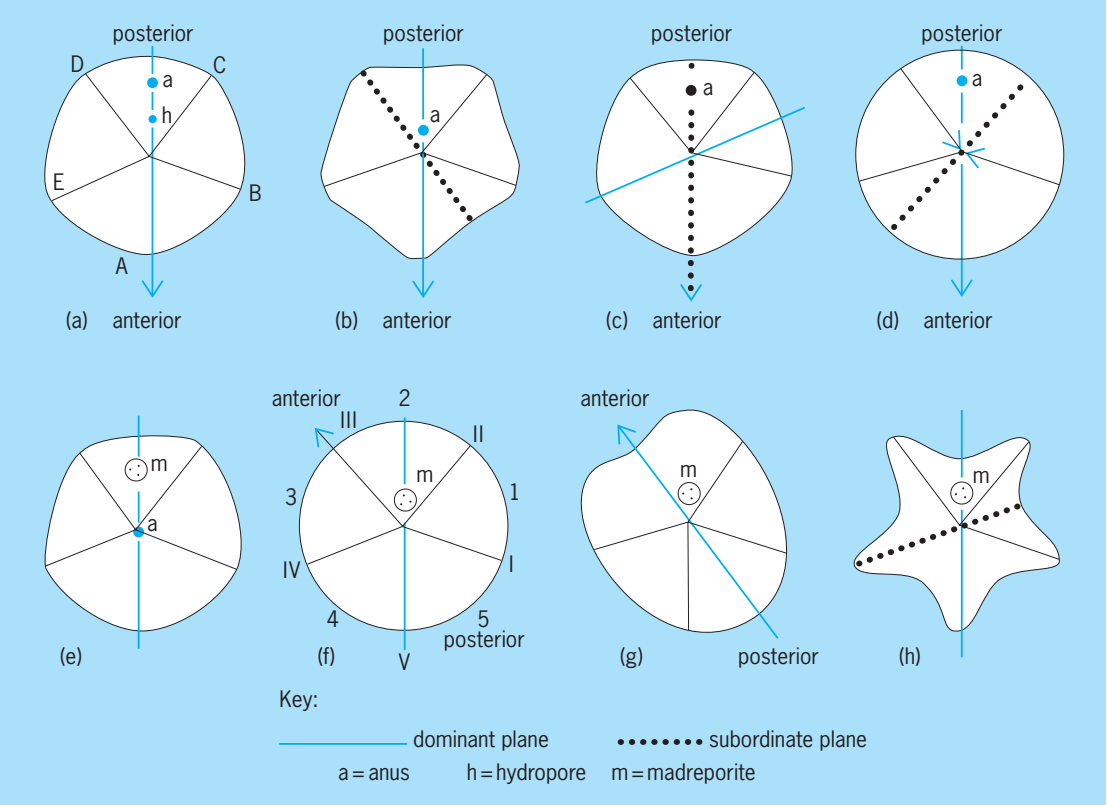


Fig. 7. Bilateral symmetry developed in various echinoderms. Anterior and posterior directions are indicated where distinguished. All diagrams represent aboral views. (*a*) Crinozoans (in general, most crinoids, cystoids, edrioasteroids); letters denote ray designations according to the Carpenter (P. H. Carpenter, 1884) system: A, anterior; B, right anterior; C, right posterior; D, left posterior; E, left anterior. (*b*) Blastoids, heterocrinoids, with subordinate bilateral symmetry in the left posterior plane. (*c*) Homocrinoids, with primary bilateral symmetry in the left anterior plane. (*d*) Glyptocrinoids, flexible crinoids, and rhombiferan cystoids, with subordinate bilateral symmetry in the right posterior plane. (*e*) Holothuroids. (*f*) Regular echinoids, rays marked according to the Lovén (S. Lovén, 1874) system, III being considered anterior. (*g*) Irregular echinoids, with prominent bilateral symmetry in the left posterior plane of crinozoans. (h) Asteroid, with subordinate bilateral symmetry in the left anterior plane.

bilateral symmetry is oriented in the left posterior plane (Fig. 7*b*). *See* INADUNATA.

The Camerata (chamber or "box" forms) are most numerous among ancient crinoids in both variety and quantity of individuals (Fig. 6*q–s*). More than 2500 species of these fossils have been described, two-thirds of which come from Mississippian rocks alone. The camerates are distinguished by the stout construction of their calyx, which incorporates lower ray plates and interradials in the dorsal cup, and subtegminal location of the mouth. Types with both one and two circlets of plates beneath the radials are common. Anteroposterior bilateral symmetry is developed almost universally in these crinoids as a modification of the dominant pentameral pattern (Fig. 7*a*); a secondary plane of bilateral symmetry directed through the right posterior ray prevails in the monobathrid suborder Glyptocrinina, as in all flexible crinoids (Fig. 7*d*). *See* CAMERATA.

The Flexibilia (flexibles) are a distinctive assemblage of exclusively dicyclic crinoids characterized by movable ligamentous union between most of the plates and several constant features in the organization of the calyx (Fig. 6*b* and *i*). They include approximately 300 described species, all of which exhibit a secondary bilateral symmetry in the right posterior plane in addition to their generally well-marked primary bilateral symmetry directed anteroposteriorly (Fig. 7*d*). *See* FLEXIBILIA.

The Articulata (forms divided into joints) comprise Mesozoic and Cenozoic crinoids, represented by about 500 fossil and about 700 living species. The stalked articulate crinoids have been classified into four orders: Millericrinida, Cyrtocrinida, Bourgeticrinida, and Isocrinida. All of these orders have fossil and extant representatives. Some experts have abandoned the order categories for the 100 species of living stalked crinoids and prefer to arrange them into 11 families. There are three orders of unstalked crinoids: Uintacrinida and Roveacrinida are extinct, while the Comatulida (feather stars) has about 600 living species. Extant sea lilies are confined to depths below about 100 m (330 ft). The stalk is usually attached to a hard substratum—a rock, or a piece of shell embedded in soft mud—by means of a terminal disc cemented to the substratum or by means of claw-like whorls of cirri. In a few species, the end of the stalk is branched and rootlike for anchoring the stalk in soft sediments. When detached from the substratum and lying on the seafloor, sea lilies are capable of slowly moving to a new site. In contrast, the feather stars are

highly mobile, often capable of active swimming by thrashing of the arms in an up-and-down motion. Feather stars range from shallow waters to great ocean depths. They are common on coral reefs, and in certain suitable habitats can reach population densities of hundreds of individuals per square meter. Sea lilies and feather stars are passive suspension feeders; they extend their arms into the water and capture small drifting organisms and particulate matter on their sticky tube feet. Captured items are conveyed to the mouth along food grooves by means of ciliary action. In extant crinoids, the gonads develop inside the pinnules near the bases of the arms, causing conspicuous swelling. At the appropriate time for breeding, the pinnules burst to release eggs or sperm into the surrounding seawater. *See* ARTICULATA (ECHINODERMATA).

Blastoids. Blastoidea, Ordovician to Permian, display a regular, fivefold radial symmetry (Fig. 6*e*), on which are superposed a first order of bilateral symmetry in the anteroposterior plane and a second order directed through the left posterior ray (Fig. 7*b*). The calyx of average-size specimens is small, with diameter of about 15 mm (0.6 in.) and height of 20 mm (0.8 in.). It is composed of 23 or 24 plates, of which five (lancet plates) are not visible externally on unweathered specimens. Within forks of the radials are a multitude of very diminutive so-called side plates that conceal the lancets; they form ambulacral areas bordered laterally by rows of threadlike free armlets (brachioles) that function as food gatherers. Particles of food are conducted to midlines of the ambulacra and thence upward to the mouth, which is at the center of the ventral surface. Beneath the ambulacra are extremely delicate, longitudinally folded, calcareous lamellae that enclose narrow troughs for circulation of water admitted through pores at the base of the brachioles; these structures, termed hydrospires, correspond to the slitlike parts of pore rhombs in the rhombiferan cystoids (Fig. 6*d*). More than 500 species have been described. *See* BLASTOIDEA.

Asterozoans. The asterozoans are star-shaped, free-moving echinoderms with well-marked radial symmetry in arrangement of their skeletal parts. They are easily recognized by their strongly developed arms and central disk (Fig. 5*d* and *e*). The arms of most asterozoans, with their tube feet functioning for locomotion and predation, are highly mobile. They range in age from Early Ordovician to Recent but are more abundant and varied in modern seas than in the fossil record. They are divided into four classes: Somasteroidea, Asteroidea, Concentricycloidea, and Ophiuroidea. The extinct Somasteroidea ranged from the Ordovician to the Devonian. Somasteroids resemble true sea stars in general shape, with broad arms, but the arms are unique in having a pinnate arrangement of rodlike plates. Asteroidea (sea stars) are distinguished by prominent flexible rays that join the central disk without clearly shown demarcation (Fig. 5*d* and *e*). Concentricycloidea are flattened and discoidal with no evident arms. The Ophiuroidea (brittle stars) have long, slender, snakelike arms which are sharply set off from the central discoid body. They are very active echinoderms, able to crawl rapidly in any direction.

Somasteroids. These rare, broad-armed sea stars are especially characterized by the featherlike arrangement of parallel, rodlike plates in their wide petaloid rays. The rods extend outward from the medially placed ambulacral plates. A food-carrying groove is contained in each ray. The aboral side of the skeleton has a rough, coarse meshwork.

Asteroids. The sea stars have a long fossil history, ranging back to the Ordovician. Sea stars are comparable to the ophiuroids in their general stellate form (Fig. 5*d* and *e*, Fig. 7*b*). They differ from the brittle stars in two main respects: lack of strongly marked separation between the body and its radially disposed hollow arms, and the open ambulacral grooves along the ventral side of the arms. The skeletal elements tend to be loosely joined; consequently, asteroids are not well suited for fossilization in a manner showing the skeletal arrangement of the entire animal. Sea stars may range in size from a few millimeters to over a meter (3.3 ft) in diameter. Ossicles along the ambulacra occur in two or four series differentiated as ambulacrals and adambulacrals. Some asteroids lack obvious arms and are essentially pentagonal with ambulacral grooves on their ventral (oral) side. Others, such as the Heliasteridae, develop as many as 44 arms as they grow, losing their fundamental pentameral symmetry but structurally resembling common types of asteroids. A single madreporite lies on the upper (aboral) surface, placed between the arms (interradially). An anus is present in the center of the aboral surface in most groups and absent in others (notably the Paxillosida).

Behavior. Asteroids are universal symbols of the ocean. They are conspicuous in a great variety of habitats, from mud to sand to rock to coral reefs. In some areas they can be present in great numbers, causing local habitat damage. The multiarmed, crown-of-thorns sea star, reaching 1 m (3.3 ft) in diameter, is common on Indo-Pacific coral reefs. It can move across reefs in large swarms, eating the soft coral tissues and causing short- to long-term devastation of reefs. Some groups of sea stars, notably the Paxillosida, have pointed rather than suckered tube feet, and they tend to swallow prey whole (such as small clams and sand dollars), ejecting the shells from the mouth when digestion is complete. The Forcipulatida are notorious predators; a forcipulatid may attach its suckered tube feet to the two shells of a clam or a mussel and then, for several hours if necessary, gently pull on the two shells, until the bivalve's muscles tire, the shells open a fraction of an inch, and the sea star drops its stomach into the bivalve, digesting the animal inside its own shell. Some deep-sea asteroids are mud-swallowers, digesting from the mud whatever organic material may be present.

Classification. The classification of the Asteroidea has been revised by many authors over the past 50 years or so. In recent years, the most commonly used classification of both fossil and extant forms is that proposed by D. B. Blake. Extinct

Paleozoic orders are Platyasterida, Pustulosida, Hemizonida, and Uractinida. Post-Paleozoic sea stars include the extinct order Trichasteropsida, and seven extant orders, all of which also have fossil representatives: Paxillosida, Notomyotida, Velatida, Valvatida, Spinulosida, Forcipulatida, and Brisingida. *See* ASTEROIDEA; FORCIPULATIDA; NOTOMYOTIDA; PAXILLOSIDA; VALVATIDA.

Concentricycloids. The sea daisies comprise three extant deep-sea species, two known from the Pacific Ocean and one from the Atlantic. Sea daisies are flattened, discoidal echinoderms up to 15 mm (0.6 in.) in diameter. The upper surface is covered with delicate overlapping plates; the lower surface has a mouth frame and a single peripheral ring of tube feet. These animals have been found only in association with pieces of wood on the deep-sea floor. Presumably, they attach to the wood substratum using their suckered tube feet. This class was discovered and diagnosed in 1986, and its formal status has been debated ever since. Several scientists have argued persuasively that the concentricycloids should be placed in the Asteroidea, while others maintain that they represent a separate class of echinoderms.

Ophiuroids. The brittle stars or serpent stars composing the Ophiuroidea are highly mobile echinoderms closely related to the sea stars. They are distinguished readily by their external form, since the small, rounded to pentagonal or scalloped disk is sharply set off from the symmetrically placed long, slender arms that extend radially outward from it. The arms may be smooth or spiny. Almost invariably they are five in number and unbranched, but in a few kinds, such as the Recent basket stars (Gorgonocephalidae), the arms are repeatedly bifurcated. All known fossil ophiuroids have simple arms. The skeleton of the central disk is composed of many regularly arranged plates, without any deviation on either the aboral or oral surfaces from perfect pentamerous symmetry. The mouth is at the center of the oral surface; an anus is lacking. A single interradial madreporite is present on a plate near the mouth; in some basket stars there are five interradial madreporites around the outer margin of the mouth skeleton. The arms are solid, not hollow as in asteroids. The arm skeleton is distinctive in being internal, consisting of "vertebrae" formed from fused ambulacral ossicles. Brittle stars are common in a great variety of habitats—living under rocks, concealed in coral reefs, or scattered in enormous numbers across great expanses of mud on the deep-sea floor. Many brittle stars are selective detritus feeders; others feed by extending their arms into the seawater and capturing small drifting organisms or particles on their sticky tube feet. Recent investigations have shown that certain brittle stars have well-developed eyes in their upper-arm skeletons, and they can respond rapidly to changing light regimes, seeking shelter in crevices as necessary.

Fossil ophiuroids are known from Lower Ordovician to Pleistocene (Recent), but they are not abundant. Fewer than 100 species have been described from Paleozoic, Mesozoic, and pre-Recent Cenozoic formations, as compared with approximately 2000 known living species. Ophiuroids are arranged in four orders: the extinct Stenurida (Ordovician to Devonian) and Oegophiurida (Ordovician to Carboniferous), and the extant Phrynophiurida and Ophiurida (Ordovician to Recent). *See* OPHIUROIDEA.

Echinozoans. These echinoderms generally have an ovoid or globose body, and they lack armlike appendages. Some echinozoans may have a low discoid body or an elongate cylindrical to almost wormlike form. All have conspicuous radial symmetry of their skeletal parts, and in almost all the symmetry is pentamerous. The skeleton of some echinozoans consists of a rigidly constructed test to which movable jaw parts and external spines are attached, whereas in others all skeletal parts either are joined together flexibly or are reduced in size and separated from one another by leathery tissue. The echinozoans comprise seven classes: Helicoplacoidea, Camptostromatoidea, Edrioasteroidea, Cyclocystoidea, Ophiocistioidea, Echinoidea, and Holothuroidea, the first five of which are long-extinct groups, while the others are represented by both fossil and living forms. (1) The Helicoplacoidea are Lower Cambrian forms with a flexible, expansible test composed of spirally arranged plates. (2) The Lower Cambrian Camptostromatoidea are conical or domal animals with plates of varying size overlapping on the lower theca. (3) Edrioasteroidea ("sessile sea stars") range from the Ordovician to the Permian. They have a many-plated test, ranging from rigid to somewhat flexible, of generally discoid shape with five distinct, straight or curved rays on the upper surface. Some were able to move about, but nearly all are found attached to some hard foreign object, such as a brachiopod or other invertebrate shell. (4) The Cyclocystoidea are Ordovician-to-Devonian echinoderms with a disk-shaped body, the upper surface of which is covered by plates arranged in concentric rings. (5) The Ophiocistioidea are Ordovician-to-Devonian echinoderms that differ from other classes in having large fine-plated tube feet, but share some skeletal features with the Holothuroidea, with which they have sometimes been aligned. (6) The Echinoidea (sea urchins) are characterized by a rigid to flexible skeleton of varied shapes, though mostly ovoid to subglobose, covered by numerous long or short movable spines (Fig. 5*a* and *b*). Pentameral symmetry is strongly developed, modified in all orders by more or less distinct bilateral symmetry and in some by differentiation of anterior and posterior regions. Very abundant in present-day seas, sea urchins range from Early Ordovician onward. (7) The Holothuroidea (sea cucumbers) are soft-bodied, leathery-skinned echinozoans with an elongate cylindrical body and skeletal parts usually consisting of discrete, generally microscopic, ossicles ("little bones") embedded in the body wall and internal tissues (Fig. 5*f* and *g*). *See* ECHINOZOA.

Edrioasteroids. A unique assemblage of mainly early Paleozoic attached echinoderms (Fig. 6*f* and *g*), Edrioasteroidea were distributed from Lower Cambrian to Upper Carboniferous. They had no stem for anchorage but were fixed by their entire base. The upper (ventral) side was covered by many

flexibly joined plates, which are sharply differentiated into rows of paired ambulacrals with adjoining adambulacrals and irregular interambulacrals. The five ambulacral tracts curve outward from the centrally located mouth, with their extremities often deflected rather strongly in a consistent way. The anus lies in one of the interrays that accordingly is regarded as posterior. *See* EDRIOASTEROIDEA.

Helicoplacoids. Numerous representatives of Helicoplacoidea have been found in the oldest fossil-bearing strata (*Olenellus Zone*) of the Lower Cambrian at localities in western Nevada and southern California. The body is ovoid to pyriform, enclosed in a test of loosely joined, spirally arranged rows of elongate plates, with columns, interpreted as ambulacral, and intervening ones, as interambulacral, originating at an apical pole and extending to the opposite oral pole. The test could be distended and retracted so as to change its shape and volume. Other echinoderms associated with these oldest echinozoans include representatives of the pelmatozoan class Eocrinoidea. *See* HELICOPLACOIDEA.

Echinoids. The class Echinoidea is the most important group of fossil eleutherozoan echinoderms (Fig. 5*a* and *b*; **Fig. 8**). They are distinguished from crinozoans by freedom from a sessile mode of life, and they share with the holothuroids a downward orientation of their oral surface. Modern species of echinoids total approximately 800, as shown by T. Mortensen's several-volume monograph (1928–1952) covering all known living kinds. In comparison, 2000 species of sea stars, 2000 species of brittle stars, and 1200 species of sea cucumbers are alive today. All eleutherozoan groups are known as far back as Ordovician time, and yet fossil echinoids are more numerous than all others of the assemblage combined. Such disparity reflects inequality in fitness of the different sorts of skeletons to be preserved. Not all parts of echinoid skeletons may be found intact. Multitudes of movable spines attached to the test during life tend to be separated and scattered, as does the echinoid masticatory apparatus, which is known as the Aristotle's lantern. Thus, knowledge of the entire skeletal structure of most fossil echinoids is incomplete. *See* ECHINOIDEA.

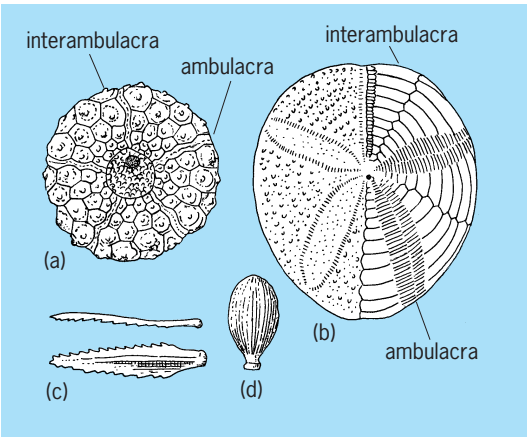


Fig. 8. Fossil echinoids. (*a*) Regular form (*Echinocrinus*, Mississippian), aboral view showing the anus and periproct, wide interambulacra and narrow ambulacra. (*b*) Irregular form (*Eupatagus*, Paleogene), aboral view, left side showing an unweathered appearance with many spine bases, right side showing the plate structure with ornamentation omitted; anterior ray directed upward. (*c*, *d*) Types of echinoid spines: (*c*) *Porocidaris*, Paleogene; (*d*) *Balanocidaris*, Jurassic.

Morphology. The echinoid test is typically globular, unevenly ovoid, or discoid in shape (Figs. 5 and 8). Mainly it is formed by 10 meridionally disposed bands of plates, of which five are ambulacral, distinguished by porelike openings between or through the plates for passage of the tube feet; the five bands of plates that alternate with the ambulacrals are interambulacral and lack perforations for tube feet. In most echinoids, both living and fossil, the plates are joined together rigidly, but in a minority the union is somewhat flexible and imbricating (having overlapping edges). Virtually all post-Paleozoic echinoids have a constant arrangement of the ambulacral and interambulacral groups of plates, each of the 10 bands comprising a double column, making 20 columns of plates in the test as a whole. The plates are widest at the equator (ambitus) of the test and increasingly narrow toward the oral and aboral poles.

The mouth, located on the underside, is surrounded by a peristome, an area of naked integument or flexibly united small plates. The anus generally is placed on the aboral side of the test opposite the mouth, but it may be shifted backward to a posterior position or even to the oral surface. The anus is surrounded by a bare or small-plated periproct. Another part of the test important for orientation and classification is the so-called apical system of 10 special plates at the aboral pole, five so-called ocular plates alternating with five genital plates. One of these genital plates is an interambulacrally placed sieve plate (madreporite) that is a landmark for identifying homologous parts, not only among diverse kinds of fossil and living echinoids but among echinoderms generally. Fossil remains of echinoids include numerous spines not associated with the test that bore them. Even so, the spines are commonly distinctive and useful as paleontological tools (Fig. 8*c* and *d*).

Orientation. Seemingly, there should be little reason for discussing the orientation of echinoids because the downward-directed oral surface clearly is ventral and, at least among irregular echinoids which are characterized by obvious bilateral symmetry, locomotion consistently is in a single direction along the path defined by the plane of this symmetry. If such direction of movement is considered to be forward, the ambulacral ray on this side of the test must be anterior and the opposite interambulacrum is posterior (Fig. 7*g*). In irregular echinoids, such as the spatangoids, the spines nearly all point backward. In many irregular genera, the mouth is found to have shifted forward and the anus rearward. These observations have led to adoption by specialists of a scheme of numerical designation of the ambulacral and interambulacral rays, introduced by S. L. Lovén (1874) and known as the Lovén system (Fig. 7*f*). It is applied to fossils as well as living echinoids and to both irregular and regular types. Orientation of

the regular echinoids seemingly should not be easy, since they move in any direction with equal ease and without detectable preference; also, except for the madreporite, the test has perfect radial symmetry. Inclusion of the regulars is possible only by use of the off-center location of the madreporite on the aboral surface as a reference point, assuming that relative position is the same in all echinoids. The interambulacrum with the madreporite is identified as number 2 of the Lovén system.

Among crinozoans, which almost certainly embrace the ancestors of echinoids, the posterior interradius is readily and positively identified. If this orientation is applied to echinoids, a discrepancy becomes evident at once, for the adopted anteroposterior plane of echinoids clearly is that coinciding with the left posterior ray of the crinozoans (Fig. 7*e* and *f*). Interambulacrum 2 (Lovén), considered as right anterior by echinoid students, is equivalent to the posterior interambulacrum of crinoids, for example. This does not mean that echinoid orientation is wrongly conceived but merely that evolution of these echinoderms has pursued a divergent path of its own, which incidentally duplicates the subordinate plane of bilateral symmetry in some crinozoans (blastoids, heterocrinoids).

Habitats and behavior. Echinoids can be extremely common from the rocky intertidal down to abyssal ocean depths. The irregular urchins, such as sand dollars, sea biscuits, and heart urchins, are always associated with soft substrates—sand, sandy mud, or mud. Sand dollars and sea biscuits live at the surface of sand or burrow just beneath the surface. Many of these echinoids ingest sand, obtaining sustenance from small organisms, such as bacteria, associated with the sand grains. Heart urchins burrow into the substratum and remain buried, feeding on particulate matter in the mud. The regular urchins are usually associated with hard substrates, typically rock, coral, or rubble, but numerous species may occur on sand or mud in quiet waters. Most of the regular urchins are vegetarians, feeding primarily on algae. Typically, echinoids reproduce by liberating eggs or sperm into the seawater from genital pores on top of the body. There are usually five pores, but in irregular urchins there may be four, three, or two pores. Fertilization of eggs takes place in the seawater, and the developing embryo becomes a planktonic larva. Within a few weeks, the urchin develops on the left side of the larva, breaks away, and sinks to the bottom as a juvenile. Some urchins have a nonfeeding yolky larva, and in others the larval stage is omitted altogether, the young being brooded by the female parent in special chambers.

Classification. Former taxonomic arrangement of the echinoids recognized two main subclasses, the regular echinoids (Regularia) and irregulars (Irregularia). Present knowledge, based on fossils as well as living forms, shows that this arrangement is artificial. The so-called regulars are a composite assemblage that on the one hand contains several extinct primitive kinds, exclusively Paleozoic, and on the other, advanced forms of modern type, all Mesozoic and Cenozoic, that include the stocks (Diadematacea) from which irregular echinoids were derived. Accordingly, classification was considerably revised (J. W. Durham and R. V. Melville, 1957; A. B. Smith, 1984; and others) so as to take account both of Mortensen's comprehensive work on living echinoids and of evidence from paleontology.

Ophiocistioids. Ophiocistioidea is an aberrant echinozoan group containing five known genera distributed from Lower Ordovician to Middle Devonian. The chief peculiarity of the group is the presence of numerous armlike appendages attached to the oral surface in an ambulacral position. They are interpreted as much-enlarged podia that are covered by abundant minute plates, a character seemingly shared with at least one early holothurian. Small ossicles in the ophiocistioid body wall closely resemble ossicles of holothurians. An early relationship with the holothuroids has been suggested.

Holothuroids. The sea cucumbers, or Holothuroidea, are a group of echinozoans that must have diverged very early from other echinoderm stocks. They clearly exhibit a fivefold radial symmetry characteristic of the phylum but otherwise differ radically from any other echinoderm assemblage. They are greatly elongated along the oral-aboral axis so as to have a subcylindrical form, and they lie on one of their sides identified as ventral because it is in contact with the substratum (Fig. 5*f* and *g*, Fig. 7*e*). The mouth, surrounded by feeding tentacles, lies at one extremity and the anus at the opposite end. The tentacles can assume a variety of forms, from richly branching (dendritic) to shield-shaped, to featherlike, to digitiform (fingerlike). Sea cucumbers are unique among living echinoderms in possessing an internal madreporite and a single gonad. In a small percentage of species, the body is enclosed in a test of conspicuous plates, but in most the body is flexible, the skeleton consisting of microscopic calcareous plates loosely distributed in the dermis. When the animal dies, decay of the uncalcified tissues liberates the skeletal parts, which almost invariably become scattered about. Less than a dozen whole-animal fossil holothuroids have been described. The fossil remains of the holothuroids usually consist only of discrete microscopic ossicles. Even so, about 200 species of these ossicles have been described from rocks ranging from Devonian to Pleistocene. Most of the fossil forms can be referred to one or the other of the orders of extant holothuroids. Sea cucumbers range from rocky shores to the greatest ocean depths. On the deep-sea floor they can be present in enormous numbers and may comprise 95% of the total weight of animals on the seafloor. Typically sea cucumbers ingest the soft sandy to muddy substrata on which they occur. Such feeding may be nonselective mud-swallowing or selective detritus ingestion. The sea cucumbers with branching tentacles are suspension feeders, extending their sticky tentacles into the water and capturing small drifting organisms or organic particles.

About 20 tropical species are prized as food, especially in Asia, and as a consequence these species are more or less fished out throughout the tropics. Six orders of living sea cucumbers are recognized,

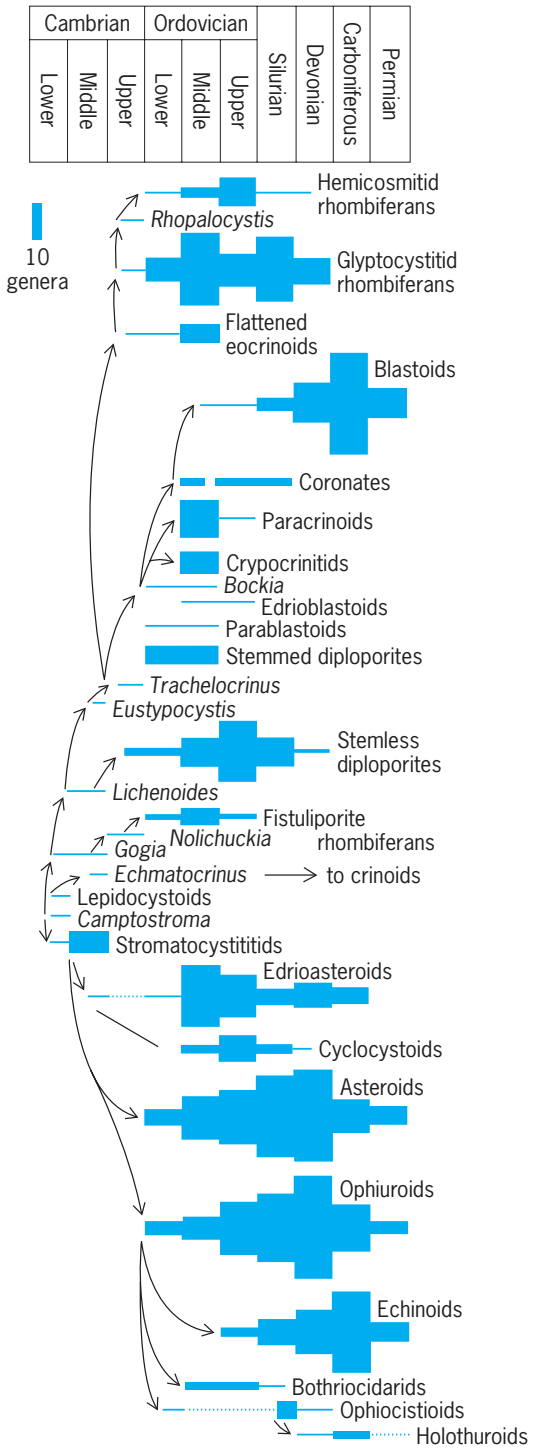


Fig. 9. Phylogeny and diversity range chart for the Paleozoic echinoderms. (*After C. R. C. Paul and A. B. Smith, The early radiation and phylogeny of echinoderms, Biol. Rev., 59:443–481, 1984*)

based upon characters of the feeding tentacles and skeleton. *See* HOLOTHUROIDEA.

Origin and Phylogeny

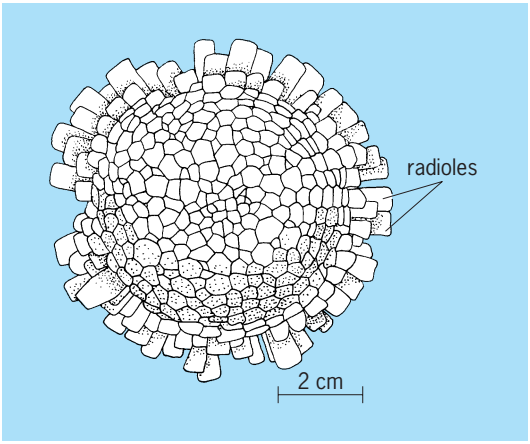
The evolution and relationships of the major groups of echinoderms are under active discussion, and several major issues have yet to be resolved. The extraxial/axial theory (EAT) of B. David and R. Mooi offers a new approach to discussion of skeletal homologies. Echinoderms evolved very rapidly near the beginning of the Paleozoic Era, and Lower Cambrian deposits contain such divergent branches of the phylum as homalozoans Helicoplacoidea, Edrioasteroidea, and Eocrinoidea (**Fig. 9**). These are primitive types of echinoderms. Cystoids, crinoids, and blastoids, as well as all recognized main groups of asterozoans and echinozoans (except holothurians), appear in Ordovician strata. During the Paleozoic, numerous well-marked evolutionary trends are discernible in nearly all echinoderm groups, including free-moving forms (especially echinoids) as well as crinozoans. Many small classes of echinoderms became extinct during the Paleozoic, and the surviving groups, especially the crinoids, lost many members at the Late Permian mass extinction. All groups of modern echinoderms have their origin in early Paleozoic stocks, and the lines of their phylogeny are mostly indicated by the fossil record. Echinoids predominate in Mesozoic and Cenozoic echinoderms.

David L. Pawson; Raymond C. Moore;
J. John Sepkoski, Jr.

Bibliography. B. David and R. Mooi, Embryology supports a new theory of skeletal homologies for the phylum Echinodermata, *C.R. Acad. Sci. Paris Ser. 3*, 319:577–584, 1996; L. H. Hyman, *The Invertebrates: Echinodermata*, vol. 4, 1955; D. T. J. Littlewood, Echinoderm class relationships revisited, in *Echinoderm Research*, ed. by H. Roland et al., A. A. Balkema, Rotterdam, 1995; C. G. Messing, Living comatulids, *Paleontol. Soc. Pap.*, 3:3–30, 1997; R. Mooi, Not all written in stone: Interdisciplinary syntheses in echinoderm paleontology, *Can. J. Zool.*, 79:1209–1231, 2001; C. R. C. Paul and A. B. Smith (eds.), *Echinoderm Phylogeny and Evolutionary Biology*, 1988; A. B. Smith, Echinoderm larvae and phylogeny, *Annu. Rev. Ecol. Sys.*, 28:219–241, 1997; A. B. Smith, *Echinoid Palaeobiology*, 1984.

Echinoida

An order of Echinacea with a camarodont lantern, smooth test, imperforate noncrenulate tubercles, ambulacral plates of echinoid type, and shallow



***Colobocentrotus atratus*, aboral aspect, a Pacific species adapted for life on wave-exposed coral reefs.**

branchial slits. There are numerous tropical and temperate species, some of them remarkably adapted to living on coral reefs (see **illus.**). *See* ECHINOIDEA.

The five included families are principally distinguished by characters of the pedicellariae. The Parasaleniidae are oblong forms with trigeminate ambulacral plates. They range from the Eocene to the present day. The Echinidae possess trigeminate or polyporous plates with the pores in a narrow vertical zone. Strongylocentrotidae are polyporous, with the pores in 2–4 vertical series. The Echinometridae show a variety of forms, which include polyporous types with an oblong test, and trigeminate or polyporous types with a spherical test. *See* ECHINACEA; ECHINODERMATA.

Howard B. Fell

Bibliography. A. Smith, *Echinoid Palaeobiology*, 1984.

Echinoidea

A class of Echinodermata known as the sea urchins, also including sand dollars, sea biscuits, and heart urchins. In echinoids the body is enclosed in a hard shell, or test, formed from regularly arranged plates that bear movable spines (**Fig. 1**). There are no arms, but radii are represented by five double rows of tube feet arranged as meridians between the upper and lower poles of the body.

There are about 800 living species, and some 5000 fossil species have been recorded, included in 225 genera. Echinoids are classified in 15 orders, which are grouped in three subclasses. Sea urchins range in size from approximately 5 mm (0.2 in.) across the test to 20 cm (8 in.). Although many species are dull or dark in color, some are brilliant shades of purple, red, green, or orange. Others species have particolored striped spines, and deep-sea forms may be white. The earliest echinoids occur in the Late Ordovician. Morphological and molecular analyses suggest that the echinoids are most closely related to the holothuroids (sea cucumbers) within the Echinodermata. *See* ECHINODERMATA; EUECHINOIDEA; SAND DOLLAR; SEA URCHIN.

Relation to humans. Most sea urchins are harmless, although some tropical species have hollow, brittle spines that cause septic wounds if they break off after penetrating the skin. The genus *Araeosoma* carries venomous spines that can inflict dangerous wounds, and a related genus, *Asthenosoma*, is said to be able to kill an adult human. The egg masses, or roe, of sea urchins are highly prized in Asian cuisine. As a consequence, active fisheries for sea urchins and sea urchin farms exist in the United States, Chile, and other countries. *See* ECHINOTHURIOIDA.

Ecology. Sea urchins occur in all seas from low-tide level downward. *Pourtalesia* reaches a depth of more than 4.5 mi (7250 m) in the Banda Trench. The rounded sea urchins feed mainly on algae, often being partially covered by day under stones, shells, or pieces of seaweed held over the test by tube feet, and emerging at night or at high tide. Heart urchins, sand dollars, and the like live buried in mud or sand, feeding on organic matter associated with ingested sand or selectively feeding on detritus.

Sea urchins move slowly, using muscles at the bases of the spines to swing the spines like stilts. The tube feet are used in concert with spines to aid in locomotion. The suctorial tube feet can be used to ascend steep surfaces and act as anchors.

Numerous sea urchin parasites are recorded. Among these are protozoans, nematodes, and gastropods, of which several genera burrow into the test. In a few species, crabs may live in the rectum or on the test, feeding on spines and tube feet. Other animals shelter among the spines or attach themselves to exposed hard parts.

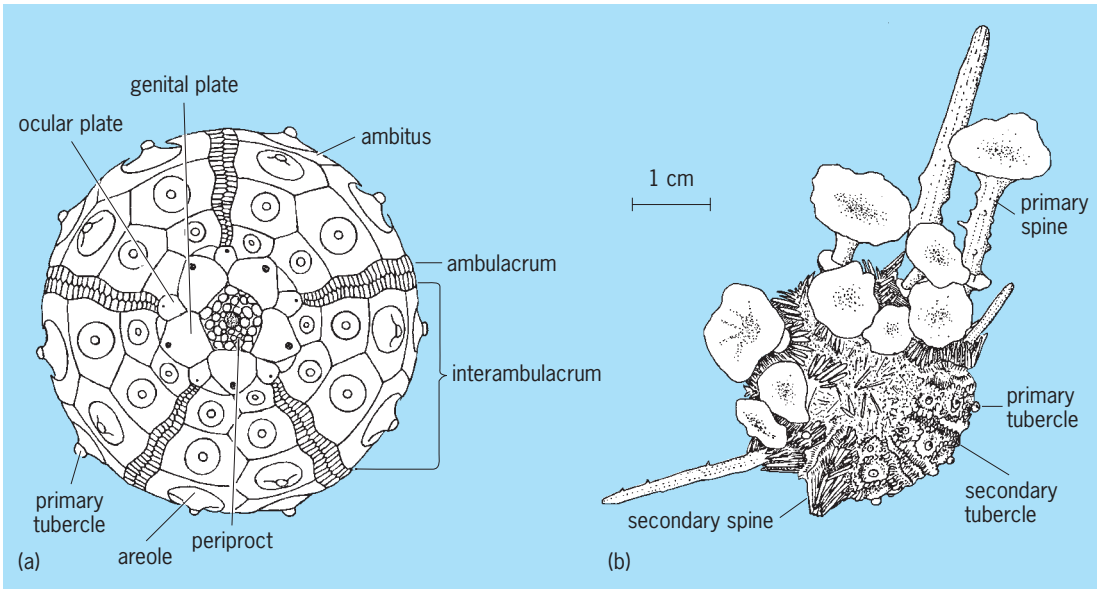


Fig. 1. Structure of the echinoid test of *Goniocidaris parasol*. (*a*) Naked test. (*b*) Test with spines.

Skeleton. The skeleton comprises the test and spines.

Test. The test is composed of regularly arranged plates that collectively form a rigid, or occasionally flexible, investing shell. In all echinoids since the Triassic, the test comprises 10 meridional areas, each composed of two vertical columns of plates. The meridians converge above and below at the upper and lower poles. The equatorial zone is termed the ambitus. Of the 10 meridional areas, five are ambulacra (or ambs); each amb consists of a double row of plates carrying tube feet. Alternating with the ambs are five meridional interambulacra (or interambs). The interambs are usually significantly wider than the ambs (**Fig. 2**).

Paleontology reveals that the test evolved only after much trial and error. The earlier Paleozoic forms showed extreme instability in the number of columns of plates. The cidaroids were the first group in which stability was achieved (in the Permian and Triassic), but a rigid, spherical shape was not adopted until the Jurassic. All of the bilaterally symmetrical irregular echinoids evolved from cidaroid ancestors. *See* PERISCHOECHINOIDEA.

Spines. Most echinoids carry spines that articulate with tubercles on the test plates and are moved by muscles at the spine bases. Large primary spines articulate with large primary tubercles, smaller secondary spines with secondary tubercles. If a ligament links the spine to the tubercle, the tubercle carries a small hole where the ligament is attached and is termed perforate. Tubercles without such a ligament are imperforate. The spine muscle may attach to an indented pattern on the edge of the boss (a protuberant part) on which the tubercle stands; if so, the tubercle is termed crenulate; if indentations are absent, the tubercle is noncrenulate. The muscle is attached to the test plate on a saucer-shaped depression around the tubercle, the areole. All these features are used in systematic diagnoses of the orders and families.

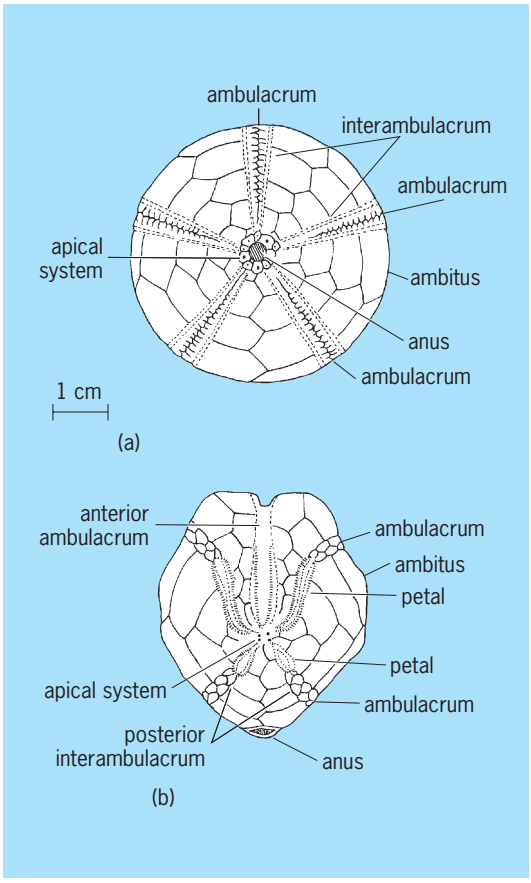


Fig. 2. Tests and related structures of (*a*) regular (endocyclic) echinoid and (*b*) irregular (exocyclic) echinoid.

At the upper pole lies a circlet of 10 plates, the apical system. Of these, five are at the top of the ambs and the five that alternate with them are at the top of the interambs. Each plate atop the interambs carries a gonopore, and is termed a genital plate. One of these also carries a madreporite (a delicately perforated sieve plate that takes in water to the vascular system). The other five are termed ocular plates, an old and misleading name for each ocular plate carries a tentacle, not an eye. The sea urchin adds new plates to its skeleton at the site of the ocular plates. Thus, the youngest plates in the test are nearer to the oculars, and plates more distant from the oculars, that is, toward the mouth, are the oldest. Individual plates grow by adding calcium carbonate to their margins. The resulting growth rings on the plates are seldom correlated with elapsed growth time.

In the spherical (regular) echinoids (Fig. 2*a*), the anus lies within the apical system on a membrane termed the periproct. These forms are sometimes termed endocyclic. In the sand dollars and heart urchins, the anus becomes displaced outside the apical system, it enters an interamb termed the posterior interamb, and the echinoid is said to have become irregular (exocyclic) [Fig. 2*b*]. Irregular echinoids tend to be secondarily bilaterally symmetrical, although a five-part radial symmetry is always evident. These features were formerly used in classification and are still valuable in determining trends of development in evolution. Irregular forms may exhibit modification of the apical system. The anus, for example, tends to obliterate the posterior genital plate (and its gonad) as it migrates backward, and the lost structures may not be replaced, or considerable distortion may ensue. *See* IRREGULARIA; REGULARIA.

Lantern. The mouth lies on the lower (oral) side of the test, surrounded by soft or plated skin, the peristome. In regular forms the mouth is central; in irregular forms it is central or may be displaced to a radius opposite the interamb that contains the anus. The radius containing the mouth is then termed anterior. In regular forms and in some irregular forms, the mouth is furnished with a ring of five powerful jaws, each with one large tooth. The jaws and teeth collectively make up the so-called lantern, first described by Aristotle. The lantern is moved by muscles that are attached to an inturned flange of the test, around the peristome, the perignathic girdle. Variations in the structure of the girdle, lantern, and teeth are used in classification. The teeth may be

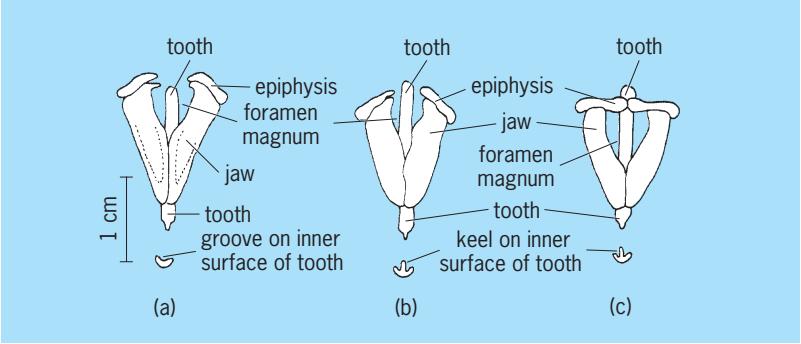


Fig. 3. Lantern structure and types of dentition for each main type. A single interradial jaw is illustrated, and below each is a cross section of the tooth. (*a*) Aulodont. (*b*) Stirodont. (*c*) Camarodont.

grooved or keeled, and the jaw may be partly or completely roofed over by epiphyses (**Fig. 3**). Three main types of dentition are distinguished: (1) aulodont, in which the teeth are grooved and epiphyses do not meet, so that there is an open foramen magnum in the jaw (Fig. 3*a*); (2) stirodont, in which the teeth are keeled within and the foramen magnum is open (Fig. 3*b*); and (3) camarodont, in which the teeth are keeled and the foramen magnum is closed by the epiphyses (Fig. 3*c*). These characters are useful in taxonomy. However, it was shown by T. Mortensen (during 1928–1951) that parallel dentitional evolution has occurred in various groups, and the three orders formerly based on dentition have been abandoned. As concise morphological descriptive terms, aulodont, stirodont, and camarodont remain valuable, and they are used in that sense in taxonomic diagnoses.

Water-vascular system. This system is highly developed and greatly influences the form of the ambulacral plates. The ring vessel rests upon the lantern and bears small reservoirs—polian vesicles. The stone canal passes upward through the coelom to the madreporite. The radial vessels and their ampullae lie on the inner surface of the ambs, within the test. The tube feet alone emerge to the exterior by way of paired pores.

Each tube foot traverses the test wall by means of two pores, termed a pore pair. One pore serves for the outward flow of hydrocoel fluid, the other for inward flow. The pore pairs lie on the amb plates (**Fig. 4**). In young stages and in the adults of Cidaroida and the irregular echinoids, each amb plate bears only one pore pair. Such amb plates are termed simple (Fig. 4*a*). In the older stages of noncidaroid regular echinoids, the amb plates tend to fuse into compound plates, which therefore carry more than one pore pair. The most common arrangement is that in which an arc of three pore pairs occurs on a plate. Such a plate is termed trigeminate or oligoporous (Fig. 4*b*, *d*, and *f*). Plates with four or more pore pairs are termed polyporous (Fig. 4*c*, *e*, and *g*). One component of a plate is usually larger than the others and is termed the primary. The position of the primary and the arrangement of the pore pairs provide characters used in taxonomy.

Diadematoid plates may be trigeminate or polyporous with the primary immediately over the lowest element (Fig. 4*b* and *c*). Arbacioid plates are similar, but the pore pairs lie in a vertical series, and the secondary elements (demiplates) are rectangular (Fig. 4*d* and *e*). Echinoid plates differ from diadematoid plates in having the primary as the lowest element (Fig. 4*f* and *g*). The ambulacra in some irregular echinoids tend to change from simple meridians into petal-shaped areas around the upper pole and around the mouth. Such ambs are described as petaloid, and the parts that surround the apical system are called the petals, whereas those around the mouth are termed phyllodes. A further

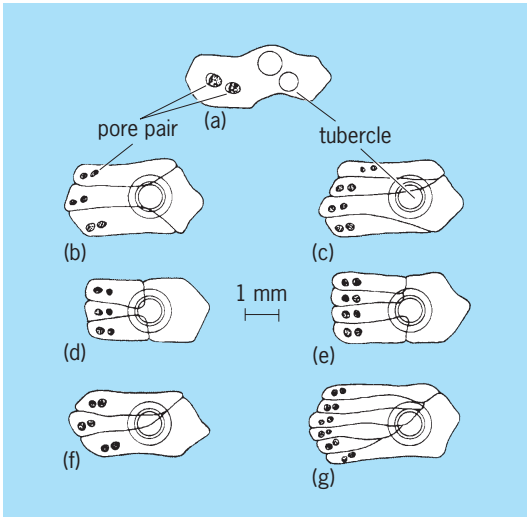


Fig. 4. Types of ambulacral plates as they are used in diagnoses of the orders. (*a*) Simple. (*b*) Trigeminate diadematoid. (*c*) Polyporous diadematoid. (*d*) Trigeminate arbacioid. (*e*) Polyporous arbacioid. (*f*) Trigeminate echinoid. (*g*) Polyporous echinoid.

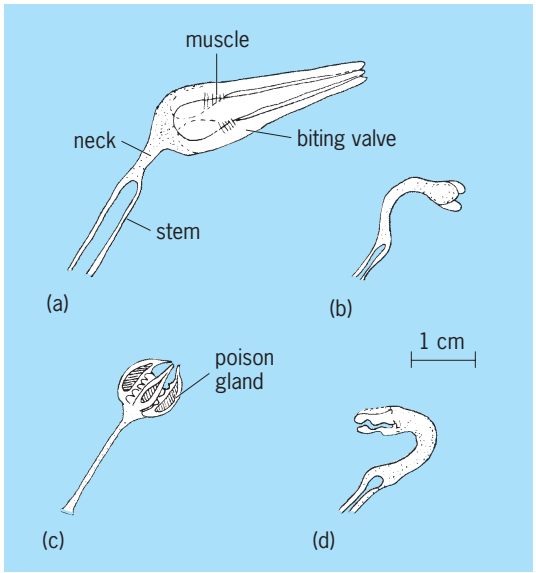


Fig. 5. Pedicellariae of echinoids. (*a*) Tridentate type. (*b*) Triphyllous or trifoliate. (*c*) Globiferous or gemmiform. (*d*) Ophicephalous.

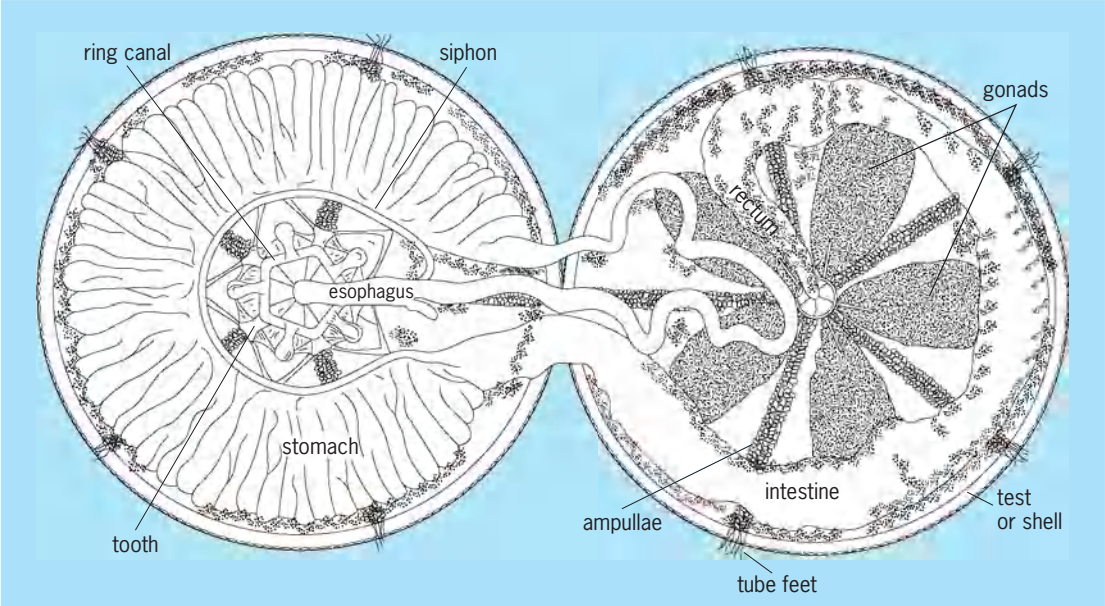


Fig. 6. Echinoid cut horizontally across the ambitus. Left is the adoral hemisphere, gonads removed. (*Adapted from W. R. Coe, Echinoderms of Connecticut, 1912; republished in T. I. Storer and R. L. Usinger, General Zoology, 3d ed., McGraw-Hill, 1957*)

development occurs in the Cassiduloida, where a flowerlike structure, the floscelle, is present.

The surface of the test plates is usually smooth but sometimes sculptured; that is, a raised pattern of ridges (epistoma) or of ridges and grooves ramifies among the tubercles. *See* TEMNOPLEUROIDA.

Some echinoids have special respiratory organs or gills attached to the peristome. The gills, if present, usually notch the margin of the peristome. The notches, termed gill slits, may vary in size and shape or may be absent. They provide characters used in taxonomy. Many irregular forms use the tube feet as respiratory organs.

Small grasping organs, pedicellariae, are well developed in echinoids, in which they take the form of a beak carried on a stalk. The beak is made up of three (rarely two or four) movable jaws, operated by muscles and sometimes provided with venom glands. They respond to tactile stimuli and seize any small organisms or particles that may touch the skin. The intrusive material is passed from one to the other until one ambital pedicellaria drops it off the urchin. The chief types are shown in **Fig. 5**.

Nervous system. The nervous system follows the same pattern as the water-vascular system. The tube feet, especially those near the mouth, may serve as tactile and taste organs. A few sea urchins have photosensitive eyespots on the upper surface of the test, scattered in the ectoderm. The so-called ocular pores are not sensitive to light. Minute spherical stalked bodies attached to the skin in some echinoids are believed to be organs of balance.

Alimentary system. The alimentary canal is tubular. It lies in the coelom, attached to the wall of the test by mesenteries. The stomach runs from the esophagus in a counterclockwise coil (as viewed from above), and the intestine retraces the route in reverse. The rectum passes upward to the anus (**Fig. 6**).

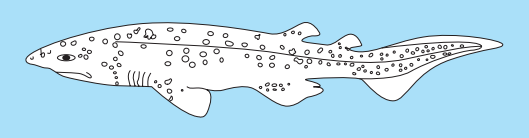
Reproduction. Like sea stars, sea urchins seem to become sexually mature after 1 year but continue to grow for several years. The life span is unknown but may average 5–6 years in medium-sized species, and some deep-sea and cold-water species may live much longer. The sexes are normally separate, although fertile hermaphrodites are known. The gonads may be reduced in number to only three or two. If a pelagic larva is present, it is an echinopluteus. In some species the young are brooded among the spines or in sunken petals, as in *Abatus*. Spines and some other organs are regenerated after injury. Autotomy (the process of self-amputation of appendages) and autoevisceration (expulsion of the digestive tract and associated organs through the anus) are unknown.

David L. Pawson; Howard B. Fell

Bibliography. J. W. Durham and R. V. Melville, A classification of echinoids, *J. Paleontol.*, 31(1):242–272, 1957; G. Hendler et al., *Sea Stars, Sea Urchins and Allies: Echinoderms of the Florida Keys and the Bahama Islands*, 1995; D. T. J. Littlewood, Echinoderm class relationships revisited, in *Echinoderm Research*, ed. by H. Roland et al., A. A. Balkema, Rotterdam, 1995; R. Mooi, Not all written in stone: Interdisciplinary syntheses in echinoderm paleontology, *Can. J. Zool.*, 79:1209–1231, 2001; T. Mortensen, *Monograph of the Echinoidea*, 5 vols., 1928–1951; D. Nichols, *Echinoderms*, 1962; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; J. S. Pearse and R. A. Cameron, Echinodermata: Echinoidea, in A. C. Giese, J. S. Pearse, and V. B. Pearse (eds.), *Reproduction of Marine Invertebrates*, vol. 6, pp. 514–662, Boxwood Press, Pacific Grove, 1991; A. B. Smith, Echinoderm larvae and phylogeny, *Annu. Rev. Ecol. Sys.*, 28:219–241, 1997; A. Smith, *Echinoid Palaeobiology*, 1984.

Echinorhiniformes

An order of sharks (bramble sharks) represented by one family, one genus, and two species. The family (Echinorhinidae) differs from all other sharks in possessing the following combination of characters: body and fins covered with thornlike denticles; two small spineless dorsal fins placed posteriorly to the pelvic fin origin; anal fin lacking; minute spiracles placed well behind the eyes; teeth similar in both jaws; and caudal fin without a subterminal notch. Maximum length is 2 m (6.6 ft).



Bramble shark, *Echinorhinus brucus*. (*Courtesy of J. S. Nelson, 2006*)

Echinorhinus brucus (see **illustration**) occurs in the western North Atlantic, Indian, and western Pacific oceans, whereas *E. cookei* is limited to parts of the Pacific Ocean. *Echinorhinus brucus* has relatively few large denticles, differing from *E. cookei*, with relatively numerous small denticles. Both species inhabit temperate waters of continental and insular shelves and slopes, usually to depths greater than 200 m (660 ft) and less than 900 m (2950 ft). Their food consists of fishes, including small sharks, and crabs. Bramble sharks are ovoviviparous (producing eggs that develop internally and hatch before or soon after extrusion).

The family was previously placed in the Squaliformes. *See* CHONDRICHTHYES; ELASMOBRANCHII; SELACHII.

Herbert Boschung

Bibliography. M. R. de Carvalho, Higher-level elasmobranch phylogeny, basal squaleans, and paraphyly, pp. 35–62, in M. L. J. Stiassny, L. R. Parenti, and G. D. Johnson (eds.), *Interrelationships of Fishes*, Academic Press, San Diego, 1996; J. S. Nelson, *Fishes of the World*, 4th ed., Wiley, 2006.

Echinothurioida

An order of Echinothuriacea with solid and hollow primary radioles, diademoid ambulacral plates, noncrenulate tubercles, and the anus within the apical system. The extant members of the family Echinothuriidae are all deep-water forms. The Echinothuriidae have a large flexible test which collapses into a disk at atmospheric pressure, and the middle element of the diademoid plates is much larger than the other two elements. Some species carry venomous spines. Echinothuriids range from the Late Jurassic to present day. *See* ECHINODERMATA; ECHINOIDEA; PYGASTEROIDA.

Howard B. Fell

Echinozoa

A subphylum of free-living echinoderms in which the body is essentially globoid with meriodional symmetry. They lack arms, brachioles, or other appendages, and do not at any time exhibit pinnate structure. The Echinozoa range from the Early Cambrian to the present day. There are four classes that can definitely be placed here: (1) Edrioasteroidea, Lower Cambrian to lower Carboniferous echinozoans in which the mouth and anus were both directed upward and ambulacra (three to five in number) served as food-collecting areas; (2) Echinoidea, the existing and fossil sea urchins, originating in the Middle Ordovician; (3) Ophiocistioidea of the Lower Ordovician through Middle Devonian, with a domed aboral surface cover with large polygonal plates and a flat adoral side with a mouth and five radiating ambulacra; and (4) Holothuroidea, the existing and fossil sea cucumbers, which apparently first appeared in the Devonian. Two other extinct echinoderm classes can be placed here for convenience: (5) Lower Cambrian Camptostromatoidea, conical or domal animals with plates of varying size that overlapped on the lower theca; and (6) Lower Cambrian Helicoplacoidea, cylindrical animals with a spirally plated test and three ambulacra on the surface. The latter class may be the sister group of both echinozoans and crinozoans. *See* CAMPTOSTROMATOIDEA; ECHINOIDEA; HELICOPLACOIDEA; HOLOTHUROIDEA.

The oldest definite echinozoans are stromatocystitids of the Lower and Middle Cambrian. This group, which may have camptostromatids as its sister group, may have been ancestral to other edrioasteroids and, perhaps, other echinozoans. The asterozoans (asteroids and ophiuroids) may have been derived from the echinozoans, making the subphylum paraphyletic. *See* ECHINODERMATA.

Howard B. Fell; J. John Sepkoski, Jr.

Bibliography. R. C. Moore and C. Teichert (eds.), *Treatise on Invertebrate Paleontology*, pt. T, 1978; C. R. C. Paul and A. B. Smith, The early radiation and phylogeny of echinoderms, *Biol. Rev.*, 59:443–481, 1984; C. R. C. Paul and A. B. Smith (eds.), *Echinoderm Phylogeny and Evolutionary Biology*, 1988.

Echiura

A small group of unsegmented, wormlike, marine animals once linked with the Sipuncula and Priapulida under the taxon Gephyrea. Since 1940, echiurans have been regarded as a separate phylum of the animal kingdom with affinities to the segmented annelid worms. They range from the tropical to the polar seas, burrowing in sea-floor sediments from the intertidal area to depths of 30,000 ft (9140 m). *See* ANNELIDA; PRIAPULIDA; SIPUNCULA.

Morphology. The echiuran body consists of a round or elongated trunk with an anterior prostomium (see **illustration**). The mouth is located at the anterior end of the trunk where the tonguelike prostomium

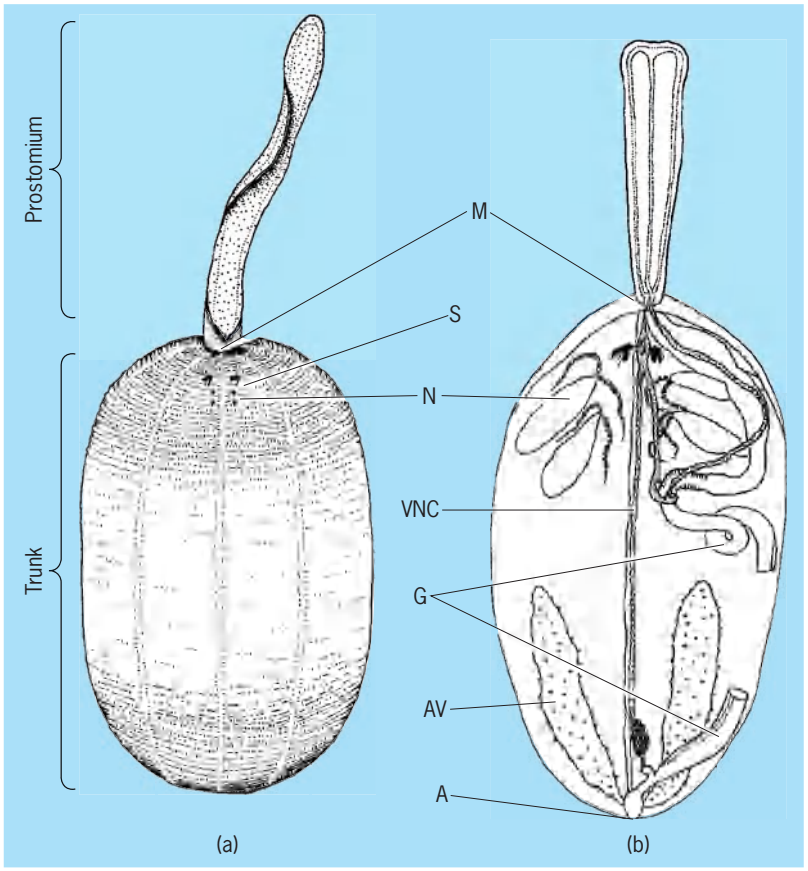
is attached. The anus is at the posterior end. The body wall of the trunk may be thin enough to see red blood and internal organs within it, or it may be green and opaque as in the Bonelliidae. The muscles of the body wall may be a uniform sheet or visibly gathered into varying numbers of bundles that run from anterior to posterior (front to back). There is usually a pair of setae (hooked bristles) situated ventrally a short distance behind the mouth, and in the genera *Echiurus* and *Urechis* there are one or more rings of setae that surround the anus. Within the burrow the trunk undergoes rhythmic contractions that move water through the burrow.

The prostomium is muscular and capable of considerable extension. It is important in the identification of species but often becomes accidentally detached and lost during collection. This can make identification difficult. In most echiurans the prostomium is elongated with a rounded or blunt anterior end. In the bonellid echiurans the tip of the prostomium is forked. Living in burrows in marine sediments, echiurans extend the prostomium out onto the surface where it collects and transports nutrient-rich sediments to the mouth to be consumed. Species in the genus *Urechis* have a short stubby prostomium. These species secrete a mucus net within their burrow and then pump water through it to filter out food.

The mouth, located at the base of the prostomium, leads into the gut, which is divided into distinct regions designed to systematically process the food. Two characteristic organs, the anal vesicles, are present at the posterior end of the gut near the anus. These organs may be tubular or branched and can extend for varying distances toward the anterior end of the body cavity. Water is intermittently drawn in through the anus to fill these sacs and then is expelled. Both respiratory and excretory functions have been proposed for these organs. Nephridia (organs typically known as having an excretory function in most invertebrates) are present on the ventral body wall behind the anterior setae. In echiurans the nephridia have been modified to gather ripe gametes (eggs or sperm) from the body cavity and store them prior to spawning. A ventral nerve cord runs from the anterior to the posterior end of the trunk. *See* DIGESTION (INVERTEBRATE); EXCRETION.

The sexes are separate and indistinguishable in most echiurans. In the Bonelliidae, however, the sexes are separate and very dissimilar, the male being minute and living within the female's egg storage organ. There they fertilize the eggs as they are spawned. The sex of bonellid offspring is determined largely by environmental factors.

Classification. Traditional morphological and embryological evidence has suggested a phylogenetic (evolutionary) relationship between these animals and the annelids, but the echiurans have remained a separate phylum. Embracing this taxonomic status, A. C. Stephen and S. J. Edmonds recognized three orders—Echiuroinea, Xenopneusta, and Heteromyota—the last order with one problematic species attached. Most species are included in the Echiuroinea, which is divided into families Bonelliidae and Echiuridae.



Echiuran anatomy. (*a*) External anatomy, ventral view. (*b*) Internal anatomy, dorsal view. A, anus; AV, anal vesicles; G, gut (middle section removed for clarity); M, mouth; N, nephridia (egg and sperm storage organs); S, setae; VNC, ventral nerve cord (*modified from A. C. Stephen and S. J. Edmonds, 1972*).

Contemporary evidence used to classify animals and help determine their relationships (systematics and phylogeny) now comes from molecular data (for example, DNA and RNA nucleotide sequences) and sophisticated microscopic techniques. The DNA sequence of one important gene in an echiuran has been found to be more similar to the same gene in annelid worms than it is to other closely related but nonannelid animals. This finding suggests that the echiurans may indeed be very closely related to the segmented annelid worms; however, their unsegmented body remained problematic. Recent fluorescent microscopic techniques focused on the structure of the ventral nerve cord have revealed structural repetition in the anterior to posterior axis of the organ that suggests the presence of a more subtle segmental body plan in echiurans. Taken together, these two findings have revived a hypothesis that considers echiurans to be members of a distinct class within the phylum Annelida. Classification below this taxonomic level has not been formally proposed. However, if one were, it would likely retain the orders currently recognized. John F. Pilger; Mary E. Rice

Bibliography. R. Hessling, The use of confocal laser-scanning microscopy (cLSM) and 3d reconstruction

for systematic zoology—on the phylogeny of the Echiura, *Microsc. Today*, 1(5):8–11, 2001; R. Hessling and W. Westheide, Are Echiura derived from a segmented ancestor?—Immunohistochemical analysis of the nervous system in developmental stages of *Bonellia viridis, J. Morphol.*, 252:100–113, 2002; D. McHugh, Molecular evidence that echiurans and pogonophorans are derived annelids, *Proc. Nat. Acad. Sci.*, 94:8006–8009, 1997; J. F. Pilger, Echiura, in F. W. Harrison and M. E. Rice (eds.), *Microscopic Anatomy of Invertebrates*, vol. 12, Wiley-Liss, New York, 1993; J. F. Pilger, Sipunculans and echiurans, in S. F. Gilbert and A. M. Raunio (eds.), *Embryology: Constructing the Organism*, Sinauer Associates, Sunderland, MA, 1997; A. C. Stephen and S. J. Edmonds, *The Phyla Sipuncula and Echiura*, British Museum (Natural History), 1972.

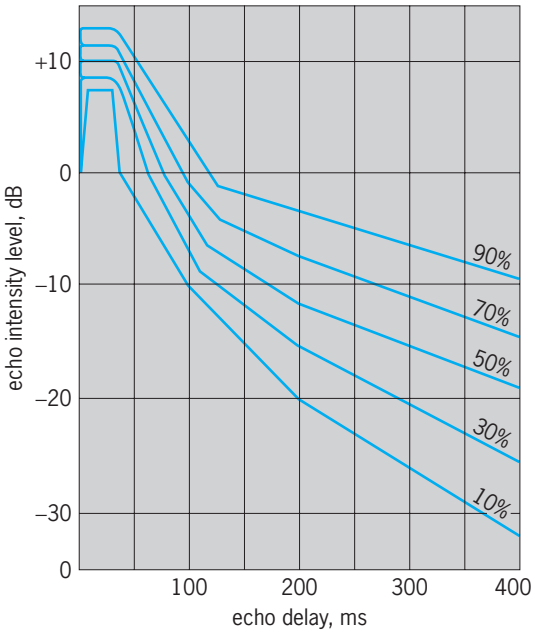
Echo

A sound wave which has been reflected or otherwise returned with sufficient magnitude and time delay to be perceived in some manner as a sound wave distinct from that directly transmitted. Multiple echo describes a succession of separately indistinguishable echoes arising from a single source. When the reflected waves occur in rapid succession, the phenomenon is often termed a flutter echo.

Echoes and flutter echoes are generally detrimental to the quality of the acoustics of rooms. They may be minimized through the proper selection of room dimensions, room shape, and distribution of sound-absorbing materials. Flutter echoes, for example, may be minimized by making the opposite walls of a room nonparallel or by making one of the walls highly sound absorptive. For a more complete discussion of the effect of room shape on echoes *see* ARCHITECTURAL ACOUSTICS

Individual echoes will generally be heard only if the time delay between the direct sound wave and a reflected sound wave is in excess of approximately 50 milliseconds. In a large auditorium where the reverberation time is of the order of seconds, many reflected waves will be present. These reflections will not be troublesome if their intensity is sufficiently below that of the initial sound. A relation between the approximate percent of listeners detecting some alteration in the acoustics of a room and the sound pressure level of the echo as a function of the delay time following the arrival of the direct sound is shown in the **illustration**. *See* REVERBERATION.

Echoes have been put to a variety of uses in measurement problems. For example, the distance between two points can be measured by timing the duration required for a direct sound originating at one location to strike an object at the other point and to return an echo to the location of the initial source. For the application of this principle to the detection of submerged objects *see* SONAR



Estimated percent of listeners disturbed by the echo related to the intensity of the echo (relative to initial sound intensity) and the time delay between the inital sound and perception of the echo. (*After R. H. Bolt and P. E. Doak, J. Acousi. Soc. Amer., 22:507–509, 1950*)

Ultrasonic echo techniques have achieved considerable success in nondestructive testing of materials. When an ultrasonic wave is propagated through a metal, the presence of a crack or other flaw will cause a sound wave, or echo, to be reflected back to the initial source location. Observing the time delay between the original sound and the perception of the echo permits the location of the flaw to be determined. This technique has been found particularly useful in such problems as examining metal casting for internal defects and determining the location of cracks in pipes or welded structures. It has also been employed to locate brain tumors in humans. For further discussion of the application of echo reflection techniques *see* REFLECTION OF SOUND; SOUND; ULTRASONICS.

William J. Galloway

Echo sounder

A marine instrument used primarily for determining the depth of water by means of an acoustic echo. A pulse of sound sent from the ship is reflected from the sea bottom back to the ship, the interval of time between transmission and reception being proportional to the depth of the water.

Echo sounders, sometimes called fathometers, are used by vessels for navigational purposes, not only to avoid shoal water, but as an aid in fixing position when a good bathymetric chart of the area is available. Some sensitive instruments are used by commercial fishers or marine biologists to detect schools of fish or scattering layers of minute marine life. Oceanographic survey ships use echo sounders for charting the ocean bottom. **Figure 1** shows an

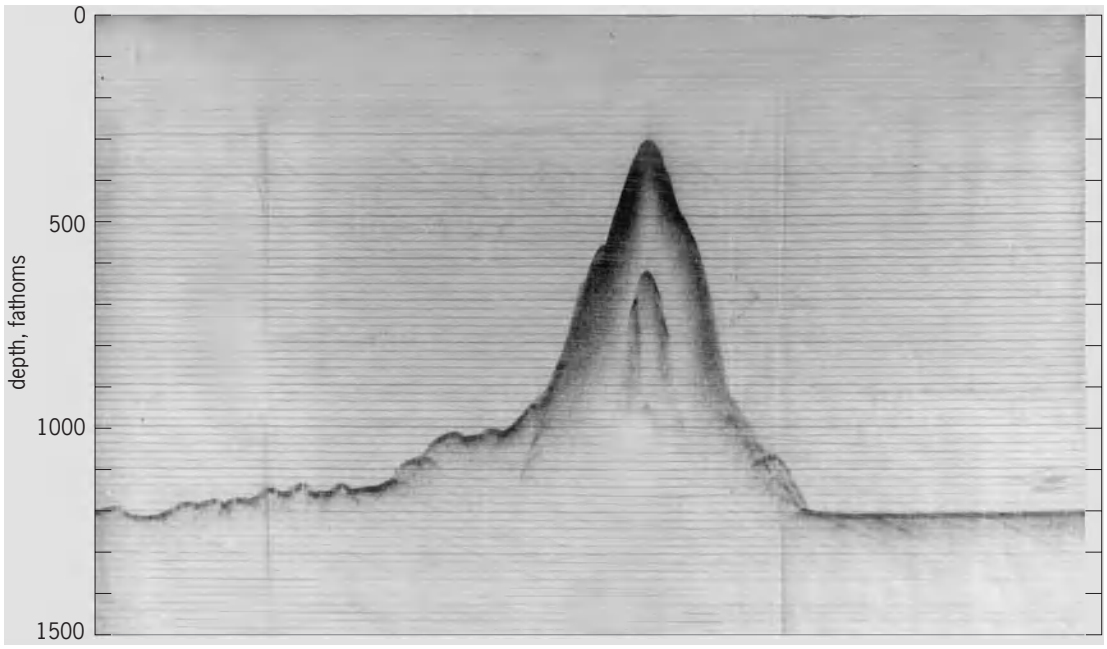


Fig. 1. Echo-sounder record of a seamount in the deep ocean. 1 fathom = 1.8 m. (*Woods Hole Oceanographic Institute*)

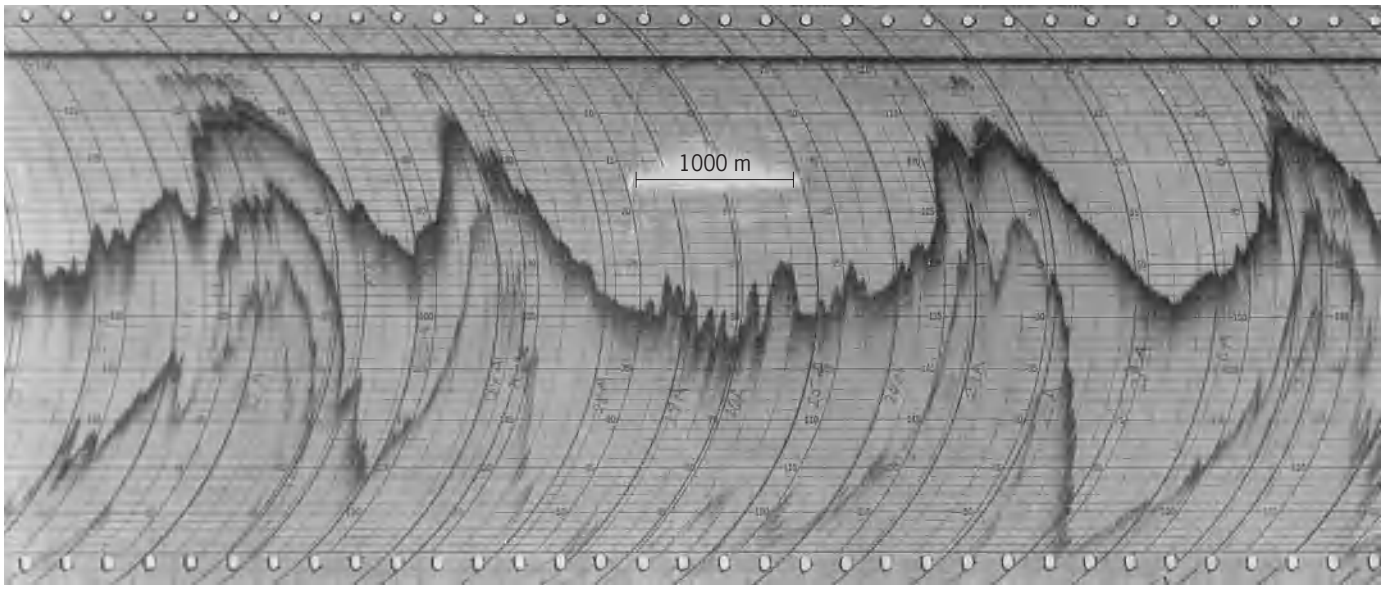


Fig. 2. Typical record of bottom profile obtained by an echo sounder. 1 m = 3.3 ft. (*USCGS*)

echo-sounder record obtained by oceanographers of a seamount (undersea mountain). *See* SCATTERING LAYER.

An echo sounder is really a type of active sonar. It consists of a transducer located near the keel of the ship which serves (in most models) as both the transmitter and receiver of the acoustic signal; the necessary oscillator, receiver, and amplifier which generate and receive the electrical impulses to and from the transducer; and a recorder or other indicator which is calibrated in terms of the depth of water. An echo sounder actually measures time differences, so some average velocity of sound must be assumed in order to determine the depth. The frequency generally employed is in the low ultrasonic range (20,000–30,000 Hz). The depth display may be given by a trace-type recorder which supplies a continuous permanent record (**Fig. 2**) or, as is the case with less expensive commercial instruments, it may be a dial-type or digital display, giving instantaneous depth of water. *See* MARINE GEOLOGY; SONAR; UNDERWATER SOUND.

Robert W. Morse

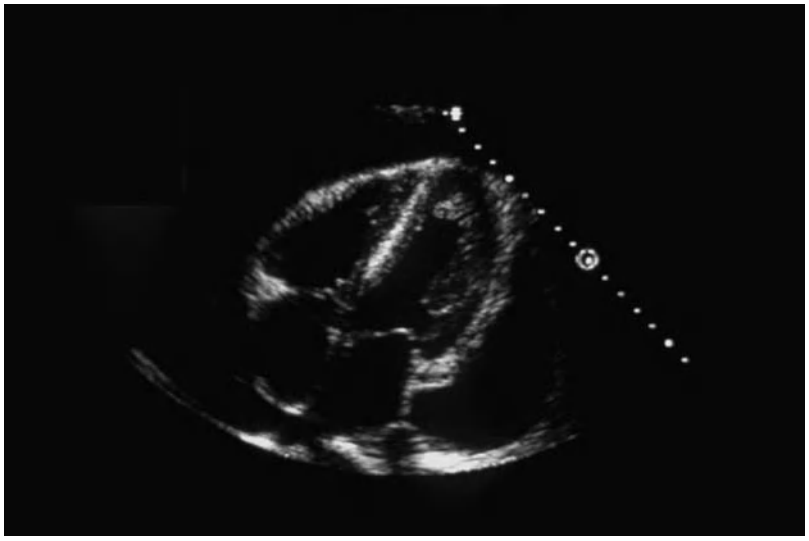
Echocardiography

A diagnostic procedure that uses ultrasound at a frequency of 2.5–10 MHz to provide an image of the heart. It is based on the principle that the interface between tissues of different acoustical impedance

causes the ultrasound to be reflected to the transducer, which spends a fraction of each second receiving these echoes. There are many interfaces between blood and the various structures in the heart that contact blood, such as the heart walls, valves, and great vessels. Also, the surface of the heart reflects ultrasound because it is surrounded by the lungs, which are filled with air.

The ultrasound apparatus can estimate the distance from the transducer to the tissue interface that is reflecting the sound, because the velocity of sound in tissue is relatively constant and the time required to send and receive sound can be measured (distance = velocity × time). A graphic representation of the locale of each interface can then be displayed on a video monitor or digital screen, and as the beam moves through the heart a composite two-dimensional image of cardiac structures can be constructed. The sampling rate of the ultrasound device is adequate to display heart motion in real time. *See* BIOMEDICAL ULTRASONICS; ULTRASONICS.

Procedure. An ultrasound image of the heart is generated on a monitor (see **illustration**). The image is generated by electronically moving the ultrasound beam repeatedly through an arc. The transducer is usually applied to the anterior chest by using a coupling gel devoid of air. This procedure is referred to as transthoracic echocardiography. Small transducers can also be attached to probes placed in the esophagus behind the heart—a procedure known as transesophageal echocardiography; during heart surgery the transducer can be placed directly on the heart—a procedure known as epicardial echocardiography; and a small transducer can be placed on a catheter to visualize the inside of vessels—a procedure known as intravascular ultrasound. The closer the transducer is to the heart, the higher the resolution of the resulting images. Axial resolution (parallel to the path of the sound beam) for all echocardiographic systems is quite high [approximately 1 mm (0.04 in.)], but lateral resolution (perpendicular to the path of the sound beam) is reduced and deteriorates the farther away the structure is from the transducer. The frame rate of the image is about 30 per second.



Two-dimensional echocardiogram from a position near the apex of the ventricles. The left ventricle is on the right, separated from the left atrium below by the mitral valve. The right ventricle is on the left, and the right atrium is below it with the tricuspid valve in between.

Two-dimensional echocardiography. Two-dimensional echocardiography is completely harmless and provides excellent real-time images of the heart. It is very useful for determining the anatomy and function of heart valves, detecting abnormal amounts of pericardial fluid, and defining the complex anatomy of congenital heart defects. It is effective for estimating size and function of heart chambers, thickness and mass of the heart wall, and size of the great vessels. The equipment is portable, and imaging can be accomplished during stresses, such as exercise. Thus, two-dimensional echocardiography is frequently employed to evaluate suspected or overt heart disease.

Blood flow. Echocardiography can also be used to assess blood flow in the heart by employing the Doppler principle. The movement of a structure during an examination performed by ultrasound results in a shift in the frequency of the returning sound wave (echo). The shift in frequency is proportional to the velocity of the moving object. Since the ultrasound frequency transmitted (fo), the speed of sound in tissue (C), and the angle of the sound beam to the direction of movement (θ) are known, velocity (V) can be calculated by measuring the frequency shift (Fd), using Eq. (1).

$$V = \frac{Fd \times C}{2fo \times \cos \theta} \tag{1}$$

Blood, the object usually interrogated by Doppler ultrasound of the heart, moves because of a pressure gradient (ΔP) in the heart. The pressure gradient can be calculated from the simplified Bernoulli equation (2) if velocity (V) is known.

$$\Delta P = 4V^2 \tag{2}$$

Doppler systems. Three basic types of Doppler echocardiographic systems are available on most ultrasound machines: continuous-wave, pulsed, and color-flow.

In a continuous-wave (CW) Doppler system, sound is continuously transmitted and received. The advantage of CW Doppler is that the very high velocities of severe structural heart defects are accurately quantitated. The disadvantage is range ambiguity, that is, the inability to locate the origin of the high-velocity blood flow along the sound beam.

Pulsed Doppler avoids range ambiguity by being able to range-gate the returning echoes so that only certain areas in the sound beam are sampled. However, pulsed Doppler cannot resolve very high velocities because the pulse repetition frequency limits Doppler frequency-shift sampling. Pulsed and CW Doppler are useful for detecting and localizing increased blood velocities associated with obstructions to flow. *See* DOPPLER EFFECT.

Pulsed Doppler can also be directed at the myocardium to measure the velocity of muscle contraction or relaxation. Currently, these tissue Doppler velocities are used to assess relaxation properties of the left ventricle.

The third system is color-flow Doppler. Small sample volumes [1–2 mm^3 (0.04–0.08 $in.^3$)] throughout the two-dimensional echocardiographic image are sampled for velocity. Only velocities fast enough to result from blood movement are displayed, superimposed on the two-dimensional echocardiographic image and color-encoded. Movement (velocity) toward the transducer is usually red, and movement away is usually blue. The relative velocity is indicated by the intensity of the color. Turbulent flow is a mosaic color (for example, green). The resultant images display blood moving through the heart in real time as a color on the black-and-white anatomy of the heart and its internal structures. Color-flow Doppler is excellent for displaying abnormal flow due to valve leakage or other structural defects.

M mode. The complete echocardiographic examination also includes selective sampling along certain radians in the arced two-dimensional image, and displaying of the movement of the structures crossed by the radian over time in an analog fashion. In this format, sampling rates can be increased to 1000 per second so that complex motion patterns of a structure can be elucidated. This format is called time-motion echocardiography or M-mode echocardiography. It was actually the first (1950s) commercially available form of ultrasound for examining the heart. It is adjunctive to two-dimensional and color-flow Doppler echocardiography for answering specific questions about heart structural dimensions and function. *See* HEART DISORDERS.

Michael H. Crawford

Bibliography. M. D. Cheitlin et al., ACC/AHA/ASE 2003 guideline update for the clinical application of echocardiography, *J. Amer. Coll. Cardiol.*, 42:954–970, 2003.

Echolocation

The biological sonar that bats, porpoises, and certain other animals use to navigate without the visual system. Several different groups of animals have evolved the ability to perceive objects by emitting sounds and hearing the echoes that the objects reflect to their ears. The locations and characteristics of the objects are represented by acoustic properties of the echoes, and the ears and auditory systems of these animals act as the sonar receiver. The sense of hearing is specialized for converting echo information into displays of objects, which are perceived as acoustic images that guide the animal's behavior. The best-known examples of echolocating animals are bats (Microchiroptera) and porpoises and toothed whales (Cetacea). However, several other kinds of mammals (some flying foxes, shrews, and rats) and birds (oilbirds and cave swiftlets) also can echolocate. *See* SONAR.

Echolocation sounds. Bats produce their ultrasonic sonar sounds from the larynx and broadcast them through the open mouth or through a specialized transmitting antenna formed around the nostrils. Many kinds of bats (leaf-nosed bats, horseshoe bats) have oddly shaped nasal structures for sound emission that appear bizarre unless it is realized that they guide the beam of the sonar transmissions. Porpoises produce their sonar sounds from structures located beneath the blowhole on the top of the head (through which they breathe) and project them into the water through the rounded, protuberant forehead, which contains acoustically specialized tissue. The sonar sounds of porpoises and whales are very brief impulses, or clicks, which contain a wide range of ultrasonic frequencies, all occurring at the same instant. The rate of emission of these sonar clicks depends in part upon the distance to objects that interest the animal, and can vary from several sounds per second to several hundred. Porpoises use their sonar to find fish and presumably to perceive objects beyond the relatively restricted range of vision under water.

The sonar sounds of bats vary from one family and species to another and consist of short (0.5 to about 10 ms) frequency-modulated (FM) signals that are often coupled with short or long (0.5 to over 100 ms) constant-frequency signals (**Fig. 1**). Their frequency span, or bandwidth, is important because the quality or sharpness of the acoustic images produced from FM echoes improves as the bandwidth increases. Under favorable conditions, bats that use wide-band FM signals can perceive the distance to a target with an acuity of a fraction of a millimeter, for example. Constant-frequency sonar signals are well suited for initial detection of targets because they concentrate their energy near a known frequency to which the sonar receiver can be turned, and long-duration constant-frequency signals are useful for determining the velocity of a target from the Doppler shift of echoes. Bats that emit constant-frequency signals perceive small changes in the frequency of echoes and use this ability to determine how fast they are moving relative to objects. Bats that emit only FM signals observe changes in a target's position from one echo to the next to determine relative velocity. *See* DOPPLER EFFECT; FREQUENCY MODULATION.

Most bats are insectivorous and hunt at night. They emit sonar sounds in a stereotyped pattern correlated with stages in the interception of a flying insect (Fig. 1). The entire interception, from initial detection to capture, lasts less than a second. Bats produce sonar sounds at a low emission rate while searching for prey; the rate increases progressively while approaching the target after it has been detected. During the approach, the bat tracks the target and decides whether to attempt interception. During the final stage of interception, which lasts only about a tenth of a second, the sonar sounds appear at a rate as high as 100 sounds per second. The constant-frequency or FM structure of the sounds (Fig. 1) represents the acoustic strategy that different species of

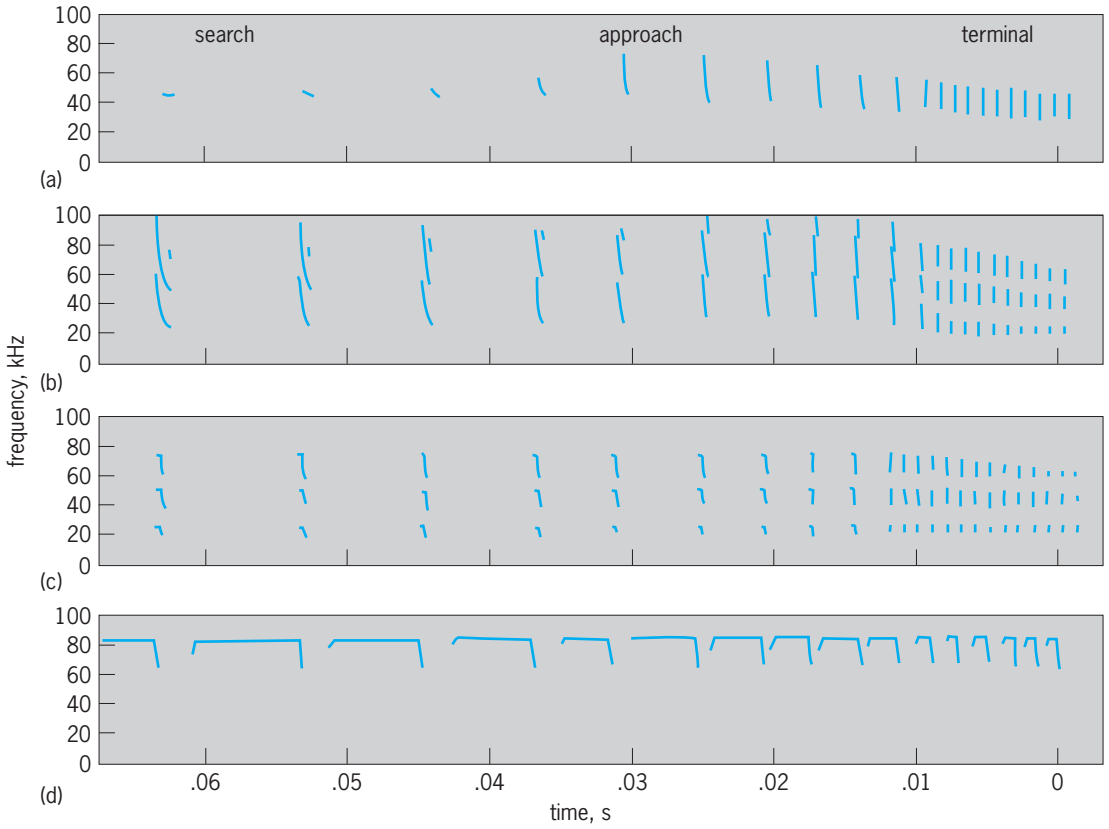


Fig. 1. Sound spectrograms representing sonar signals of (*a*) *Tadarid brasillensis*, (*b*) *Eptesicus fuscus*, (*c*) *Pteronotus personatus*, and (*d*) *Rhinolophus ferrumequinum* during search, approach, and terminal stages of pursuit of prey. Detection, identification, and tracking usually take place in less than 1 s. Spectrograms *b* and *c* use only FM signals, whereas *a* and *d* use constant-frequency signals during the search and change to FM signals during the approach and terminal buzz. (*After J. A. Simmons, F. B. Fenton, and M. J. O'Farrell, Echolocation and pursuit of prey of bats, Science, 203:16–21, 1979*)

bats have evolved to solve the problems of finding prey.

Auditory images. The images in echolocation are derived from either frequency-modulated or constant-frequency echoes.

Frequency-modulated images. The distance to a target, or target range, is represented acoustically by the delay of FM echoes. Bats first encode echo delay as the timing of nerve discharges but eventually represent delay by the location of neural responses within a discrete patch of the auditory cortex that functions as a biological map of target range. Some FM echo-locating bats combine their perception of target range with their perception of the shape of targets by perceiving the distance to different parts of the target along the same scale as the distance to the target as a whole.

The bat in **Fig. 2** emits an FM sonar sound and receives an echo from the target, here shown to be a flying moth. The target reflects echo components from parts located at different ranges (for simplicity, one echo component from its head and one from its wings). The target as a whole has a range, r, and the target's structure consists of the range separation, Δr. These target features are represented in the echo by the echo's delay, t, and the time separation of the two components within the echo, Δt, because the echo contains a separate component from each part of the target, at delay t and delay $t + \Delta t$. The bat's inner ear represents the FM emission and the echo as spectrograms analogous to those shown in Fig. 1. The spectrograms in Fig. 2 are patterns of mechanical excitation produced by the inner ear itself. Mechanical excitation then evokes neural discharges that convey the spectrograms along parallel neural channels into the auditory nervous system. The time separation of the echo components, Δt, creates notches in the spectrum of the echo as a whole separated by the frequency interval, Δf, which equals $1/\Delta t$. These spectral notches appear in the neural spectrogram of the echo as a "scalloping" pattern of the array of discharges (solid circles) caused by the lengthening of discharge latency at the lower amplitudes of the echo within the notch as compared to outside the notch. Some neural discharges near the center of the notch may even fail to occur at all (open circles) if the echo is very weak in the notch.

The delay of echoes is presented in the bat's brain by spectrogram delays, t_f, at each frequency in the emission and the echo (Fig. 2). Spectrogram delays are further processed to be encoded, frequency by frequency, in terms of the location of neural activity on the target-range map in the auditory cortex, which is the bat's auditory display of targets. Spectrogram delays are somehow integrated across frequency to determine the distance to the

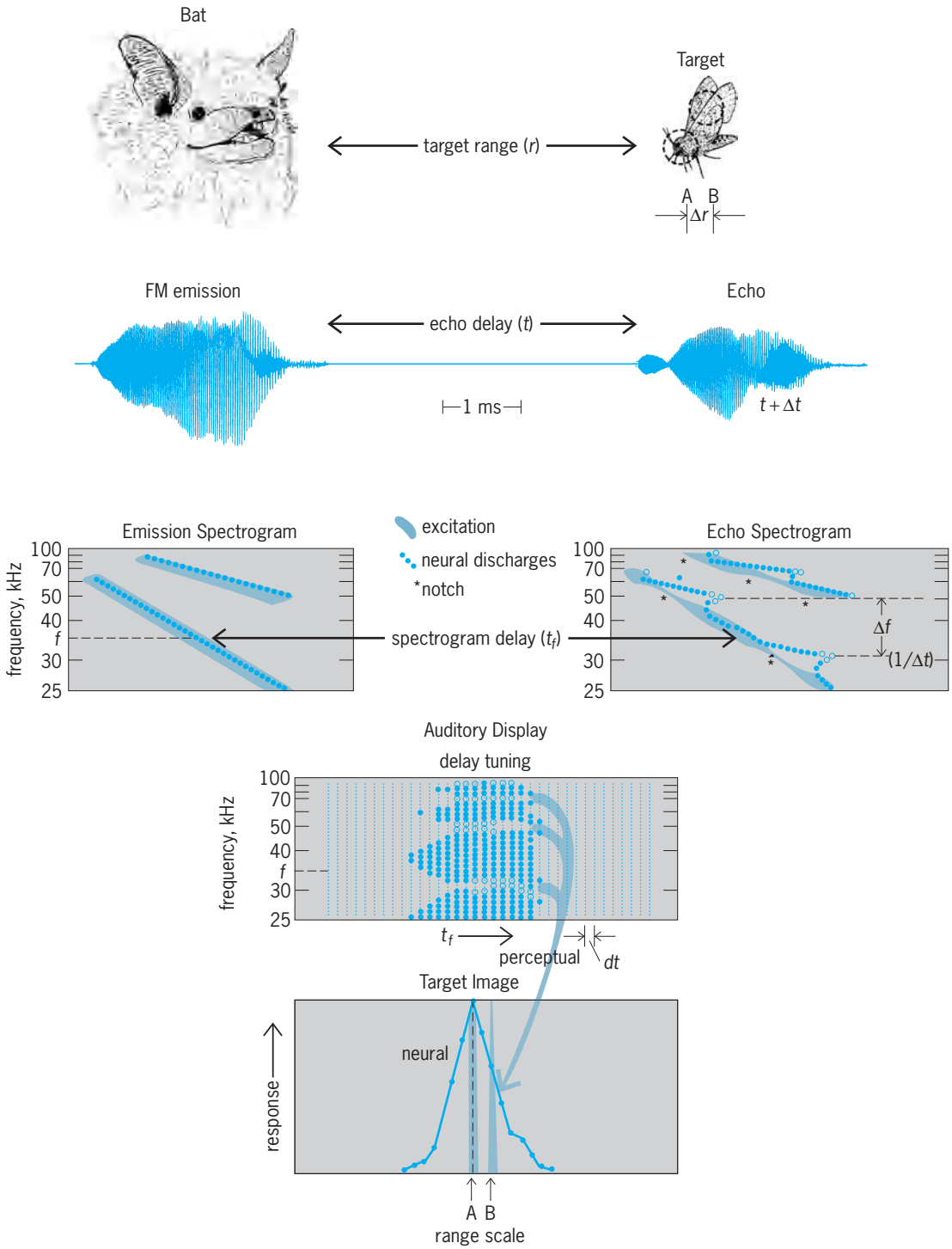


Fig. 2. Stages by which bats convert echoes of FM sonar sounds into images that incorporate the distance and the shape of the target. Details are discussed in the text. (*After J. A. Simons, C. F. Moss, and M. Ferragamo, Convergence of temporal and spectral information into acoustic images of complex sonar targets perceived by the echolocating bat, Eptesicus fuscus, J. Compar. Physiol. A, 1989*)

leading edge of the target from the earliest component of the echo, at delay t. The auditory display in Fig. 2 shows the distribution of neural activity along the range axis of the auditory cortex. The arrow leading from the auditory display to the target image shows the effect of the locations of the notches in the spectrum that produce a wavy pattern of activity in the cortex across frequencies. This wave pattern is converted into an estimate of the location of the echo's second component, at delay $t + \Delta t$. Somehow the single-peaked distribution of neural activity across the range map is converted into a two-peaked pattern to represent the target's image as two discrete components (shown as peaks labeled perceptual, *A* and *B*) located at ranges r and $r + \Delta r$. This conversion is an example of the brain creating a single, unitary psychological dimension out of neural representations that treat different features of stimuli

in different ways. It illustrates an important function of the brain that is not understood.

Constant-frequency images. The use by bats of constant-frequency echoes to determine target velocity from echo Doppler shifts is another example of neural information processing leading to perception. The bat's inner ear represents different frequencies of sound by the location of neural discharges within the population of neurons in the auditory nerve. The frequency of a constant-frequency echo activates only those neurons tuned to that particular frequency. This activity is passed along from the auditory nerve to higher centers of hearing in the brain. When the bat receives a particular constant-frequency echo frequency, neurons in a corresponding location on frequency maps in the auditory cortex respond, whereas other neurons do not. One of these cortical frequency maps becomes a map of target velocity for constant-frequency echoes. In this case, no neural computations are needed to create the frequency map because it already exists at the inner ear. However, neural computations are needed to adjust the relationship between each frequency and the size of the frequency change corresponding to each velocity, because the Doppler shifts are proportional to the frequency being shifted. Other frequency maps segregate neural responses to constant-frequency echoes to facilitate detection of the target's fluttering motions, such as the wing beats of insects, for identification of prey. *See* CHIROPTERA; PHONORECEPTION; ULTRASONICS.

James A. Simmons

Bibliography. D. R. Griffin, *Listening in the Dark*, 1986; P. Nachtigall and P. W. B. Moore, *Animal Sonar Systems: Processes and Performance*, 1988; J. A. Simmons, M. B. Fenton, and M. J. O'Farrell, Echolocation and pursuit of prey by bats, *Science*, 203:16–21, 1979; J. A. Simmons, C. F. Moss, and M. Ferragamo, Convergence of temporal and spectral information into acoustic images of complex sonar targets perceived by the echolocating bat, *Eptesicus fuscus, J. Compar. Physiol. A*, 1989.

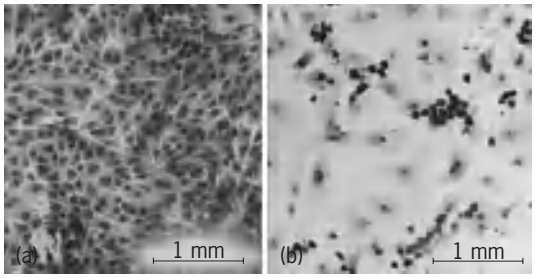
Echovirus

Members of the Picornaviridae family, genus *Enterovirus*. The name is derived from the term enteric cytopathogenic human orphan virus. More than 34 antigenic types exist. Only certain types have been associated with human illnesses, particularly with aseptic meningitis and febrile disease, with or without rash. Their epidemiology is similar to that of other enteroviruses. *See* ENTEROVIRUS; PICORNAVIRIDAE.

Echoviruses resemble polioviruses and coxsackieviruses in size (about 28 nanometers) and in many other properties. They are nonpathogenic for newborn mice, rabbits, or monkeys, but cytopathogenic for monkey kidney and other tissue cultures (see **illus.**). *See* COXSACKIEVIRUS; TISSUE CULTURE.

Diagnosis is made by isolation and typing of the viruses in tissue culture. Antibodies form during convalescence. *See* ANIMAL VIRUS.

Joseph L. Melnick



Monkey kidney tissue culture. (*a*) Normal culture. (*b*) Three days after infection with echovirus type 1. (*Courtesy of J. L. Melnick, Baylor College of Medicine*)

Bibliography. N. J. Schmidt and W. Emmons (eds.), *Diagnostic Procedures for Viral, Rickettsial, and Chlamydial Infections*, 7th ed., 1995; Z. Woldehiwet and R. Miodrag (eds.), *Rickettsial and Chlamydial Diseases of Domestic Animals*, 1993.

Eclipse

The occultation (obscuring) of one celestial body by another. Solar and lunar eclipses take place at syzygies of the Sun, Earth, and Moon, when the three bodies are in a line. At a solar eclipse, the Moon blocks the view of the Sun as seen from the Earth. At a lunar eclipse, the Earth's shadow falls on the Moon, darkening it, and can be seen from wherever on Earth the Moon is above the horizon. *See* SYZYGY.

Eclipses of the Sun could be seen from other planets as their moons are interposed between the planets and the Sun, though their superposition is not as coincidental in angular size as it is for the Earth-Moon system. Eclipses of the moons of Jupiter are well known, occurring whenever the moons pass into Jupiter's shadow. An eclipse of the Sun visible from Uranus that occurred in 2006 was imaged by the *Hubble Space Telescope*. Certain binary stars are known to eclipse each other, and the eclipses can be followed by measuring the total light from the system. *See* BINARY STAR; ECLIPSING VARIABLE STARS; JUPITER.

Related phenomena are transits, such as those of Mercury and Venus, which occur when these planets cross the face of the Sun as seen from Earth. They are much too small to hide the solar surface. Transits of the Earth will be seen from spacecraft. Occultations of stars by the Moon are commonly seen from Earth, and are studied to monitor the shape and path of the Moon; a solar eclipse is a special case of such an occultation. Occultations of stars by planets and by asteroids are now increasingly studied; the rings of Uranus were discovered from observations of such an occultation, and the atmospheres of Titan and Pluto are best known from occultation studies. *See* OCCULTATION; TRANSIT (ASTRONOMY); URANUS.

Solar Eclipses

A solar eclipse can be understood as an occultation of the Sun by the Moon or, equivalently, the Moon's shadow crossing the Earth's surface. The darkest

part of the shadow, from which the Sun is entirely hidden, is the umbra (**Fig. 1**). The outer part of the shadow, from which part of the Sun can be seen, is the penumbra.

Solar eclipses can be central, in which the Moon passes entirely onto the solar disk as seen from Earth, or partial, in which one part of the Sun always remains visible. Central eclipses can be total, in which case the Moon entirely covers the solar photosphere, making the corona visible for the period of totality, or annular, in which case the Moon's angular diameter is smaller than that of the Sun because of the positions of the Earth and Moon in their elliptical orbits. At an annular eclipse, a bright annulus of photospheric sunlight remains visible; it is normally thousands of times brighter than the corona, leaving the sky too blue for the corona to be seen.

The plane of the Moon's orbit is inclined by 5° to the plane of the Earth's orbit (the ecliptic), so the Moon's shadow commonly passes above or below the Earth each month at new moon. But two to five times each year, the Moon's shadow reaches the Earth, and a partial, annular, or total eclipse occurs. The Moon is approximately 400 times smaller than the Sun but is also approximately 400 times closer, so its angular diameter in the sky is about the same as the Sun's. Thus the Moon fits approximately exactly over the photosphere, making the phenomenon of a total eclipse especially beautiful.

Phenomena. The partial phases of a total eclipse visible from the path of totality last over an hour. In the minute or two before totality, shadow bands—low-contrast bands of light and dark caused by irregularities in the Earth's upper atmosphere—may be seen to race across the landscape. As the Moon barely covers the Sun, photospheric light shines through valleys on the edge of the Moon, making dots of light—Baily's beads—that are very bright in contrast to the background. The last Baily's bead gleams so brightly that it appears as a jewel on a ring, with the band made of the corona; this appearance is known as the diamond-ring effect (**Fig. 2**). It usually lasts for 5–10 s, and in the clearest skies for as long as 40 s.

During the diamond-ring effect, the solar chromosphere becomes visible around the limb of the Moon, glowing pinkish because most of its radiation is in the form of emission lines of hydrogen, mostly the red hydrogen-alpha line. Its emission-line spectrum apparently flashes into view for a few seconds, and is called the flash spectrum. As the advancing limb of the Moon covers the chromosphere, the corona becomes fully visible (**Colorplate** *a*). Its shape is governed by the solar magnetic field; common are equatorial streamers and polar tufts. At the maximum of the solar activity cycle, so many streamers exist that the corona appears round when it is seen in projection, as viewed from Earth. At the minimum of the solar activity cycle, only a few streamers exist so that the corona appears more elongated in projection. *See* SOLAR CORONA; SUN.

Totality (**Fig. 3**) lasts from an instant up through somewhat over 7 min, 30 s. At its end, the phenomena repeat in reverse order, including chromosphere, diamond ring, Baily's beads, shadow bands, and the partial phases.

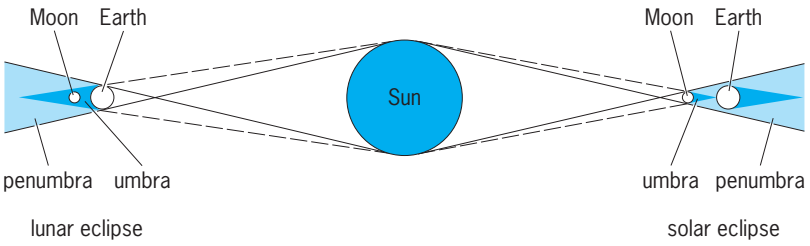


Fig. 1. Circumstances of solar and lunar eclipses (not to scale).

The corona is now monitored continuously in x-rays and in ultraviolet and visible light by orbiting spacecraft such as the *Solar and Heliospheric Observatory* (*SOHO*; ultraviolet and visible light), the *GOES Solar X-ray Images*, and the *Transition Region and Coronal Explorer* (*TRACE*; ultraviolet). The spacecraft *Solar-B*, a successor to the *Yohkoh* spacecraft, was launched in late September 2006 to continue such studies. Eclipse observations can provide ground truth for comparison with these missions on the days of eclipses. *See* SATELLITE (ASTRONOMY).

Positions and timing. The paths of the Sun and Moon in the sky intersect at two points, the ascending node and the descending node. Only when both the Sun and the Moon are near a node can an eclipse occur. Thus eclipse seasons take place each year, whenever the Sun is near enough to the node so that an eclipse is possible. Each eclipse season is 38 days long. Because the Sun's gravity causes the orientation of the Moon's elliptical orbit to change with an 18.6-year cycle, the nodes slide along the ecliptic and a cycle of two eclipse seasons—an eclipse year—has a period of 346.6 days, shorter than a solar year. *See* MOON.

Fig. 2. Diamond-ring effect at the beginning of totality during the solar eclipse of March 29, 2006, observed from Kastellorizo, Greece. (*Jay M. Pasachoff/Science Faction, Williams College Eclipse Expedition*)

Fig. 3. Solar eclipse of April 8; 2005, observed during totality from a ship in the mid-Pacific Ocean, west of the Galápagos Islands. (*Jay M. Pasachoff and Dava Sobel/Science Faction*)

There must be at least one solar eclipse each eclipse season, so there are at least two each year. There may be as many as five solar eclipses in a calendar year (technically, a year), though most of these will be partial. Some of the partials will occur near the ends of the eclipse seasons, and it is also possible for there to be three eclipse seasons during a solar year. Adding lunar eclipses (including penumbral lunar eclipses, which may not be noticeable), there may be seven eclipses in a year.

Saros. An important coincidence relates lunar months and eclipse years. A total of 223 lunar months (technically, synodic months, the period of the phases) takes 6585.32 days. A total of 19 eclipse years—passages of the Sun through the same node of the Moon's orbit—takes 6585.78 days, and 242 nodical months—passages of the Moon through the node—take 6585.36 days. (Nodical months are also called draconic months, after the ancient Chinese dragon once thought to have been devouring the Sun at a solar eclipse.) Thus eclipses appear with this period of 18 years $11^1/_3$ days (plus or minus a day, depending on leap years), a period known as the saros. Further, 239 periods of the variation of distance of the Moon from the Earth, the anomalistic month, is 6585.54 days, so the relative angular sizes of the Sun and Moon are about the same at this interval. (The anomalistic month differs from the nodical and sidereal months because the orientation of the Moon's elliptical orbit drifts around in its orbital plane.)

As a result of the saros, almost identical eclipses recur every 18 years $11^1/_3$ days. The significance of the $^1/_3$ day is that the Earth rotates one-third of the way around, and the eclipse path is shifted on the Earth's surface. Thus the June 30, 1973, 7-min eclipse in Africa was succeeded in a saros series by the July 11, 1991, eclipse in Hawaii, Mexico, and Central and South America, which reached maximum duration of 6 min 54 s in Mexico, and will lead to the 6 min 40 s eclipse of July 22, 2009. After a saros, the Sun is slightly farther west than its original position, and the Moon is slightly north or south, depending on whether it is near an ascending or a descending node, so the eclipses in a saros drift from north to south or from south to north, starting near one pole and departing from the other. A complete series takes 1244 to 1514 years.

Motion over Earth's surface. The Moon's shadow travels at approximately 2100 mi/h (3400 km/h) through space. The Earth rotates in the same direction that the shadow is traveling; at the Equator, the resulting motion of the Earth's surface makes up about 1040 mi/h (1670 km/h), making the speed of the eclipse across the Earth's surface 1060 mi/h (1730 km/h). When eclipses cross higher latitudes, the speed of motion associated with the Earth's rotation is not as high, so the eclipse speed is even higher. The supersonic Concorde took advantage of an equatorial eclipse in 1973 to keep up with totality for 74 min. Ordinary jet aircraft cannot keep up with eclipses, so the term eclipse chasing is almost never accurate.

Scientific value. Even with advances in space technology, total solar eclipses are the best way of seeing the lower corona. Coronagraphs, telescopes for which special shielding and internal occulting allow observation of the corona from a few sites in the world on many of the days of the year, are limited to the innermost corona or to use of special filters or polarization. *See* CORONAGRAPH; SOLAR CORONA.

Coronagraphs have been sent into orbit, notably aboard *Skylab*, the *Solar Maximum Mission*, and *SOHO*, but limitations in spacecraft control lead to the necessity of overocculting the photosphere, cutting out the inner corona. For example, the coronagraph on *Solar Maximum Mission* occulted 1.75 times the solar diameter. The innermost coronagraph, no longer in operation, on *SOHO*'s Large Angle Spectrographic Coronagraph (LASCO) overocculted to 1.1 solar radii, and LASCO's two remaining coronagraphs image the Sun from 1.5 to over 30 solar radii. Other types of observations from space also apply to the corona, such as imaging x-ray observations. Further, any space observation is extremely expensive, over 100 times more for *SOHO* than for a given extensive eclipse expedition. So it is useful and necessary to carry on ground-based eclipse observations, ground-based coronagraph observations, and space-based observations to get the most complete picture of the Sun. Further, independent derivations of certain basic quantities must be carried out to provide trustworthy results; there have been cases in which data reduced from space observations had to be restudied because the need for different calibrations had shown up in eclipse work.

Colorplate *b* is a merger of eclipse and *SOHO* spacecraft images, and shows how ground-based images during a total solar eclipse fill a doughnut-shaped region between the inner and outer spacecraft images that is not visible from Earth at other times. The central image from *SOHO* shows the Sun's disk at temperatures around 60,000–80,000 K through a filter showing helium gas. The outer *SOHO* image shows gas at the millions of degrees typical of the Sun's corona, with LASCO's coronagraphs hiding the bright inner corona. Filling in the doughnut ring allows coronal streamers to be traced from their roots on the Sun's surface, through the eclipse corona, and into the solar wind, the expanding outer corona. *See* SOLAR WIND.

Related eclipse studies include use of the advancing edge of the Moon to provide high spatial resolution for radio observations of the Sun and, historically, of celestial radio sources. The atmospheric effects of the removal of incident sunlight from the Earth's atmosphere have also been studied at eclipses.

Historically, the test of the deflection of starlight carried out at the eclipse of 1919 and repeated in 1922 was the first verification of Einstein's general theory of relativity. These results were restudied for the 1979 Einstein centennial, and their accuracy improved. The experiment is a very difficult one, and has been attempted at some eclipses, notably 1970 in Mexico and 1973 in Africa, without improving on the early results. But the effect has been verified to higher accuracy by studies in the radio part of the spectrum and by the observation of gravitational lenses, so optical eclipse tests are no longer necessary. *See* GRAVITATIONAL LENS; RELATIVITY.

Annular eclipses. Central eclipses in which the Moon is sufficiently far from the Earth that it does not cover the solar photosphere are annular. The Moon's umbra is about 232,000 ± 4000 mi (374,000 ± 6400 km) in length while the Moon's distance is 237,000 ± 16,000 mi (382,000 ± 25,000 km), so the umbra sometimes falls short of reaching the Earth's surface. Annular eclipses, like total eclipses, occur about every 18 months on average.

Since the corona is 1,000,000 times fainter than the photosphere, if even 1% of photosphere is showing the corona is overwhelmed by the blue sky and cannot be seen. So most annular eclipses are of limited scientific use.

The annular eclipse of 1984, visible from the southeastern United States, provided 99.8% coverage. The major scientific work carried out involved detailed timing of the Baily's beads in order to assess the size of the Sun. Such work has also been carried out at total eclipses. Some of the results, compared with solar diameters deduced from historical eclipse paths, seem to show a possible shrinking of the Sun by a measurable amount in a time of decades or centuries, which would lead to impossibly large effects on geological time scales; this has led to the suggestion that the Sun could be oscillating in size. But the question of whether any real effect is present has not been settled. The 1984 annular eclipse was so close to total that the corona could even be briefly seen and photographed, although no scientific studies of the corona were made. In the May 10, 1994, annular eclipse, widely viewed across the United States, only 94% of the Sun's diameter was covered, more typical of annular eclipses. The 1999 annular eclipse, whose path crossed Australia, provided 99% coverage, but since the Sun is about 1,000,000 times brighter than the full moon, that still left about 10,000 times more light than on a moonlit night.

TABLE 1. Total solar eclipses, 2005—2024†

Calendar date	Maximum duration of totality (min: s)	Geographic region of visibility
Apr. 8, 2005	0:42	Starts annular in western Pacific Ocean, becomes total in eastern Pacific, then becomes annular again for Costa Rica, Panama, Colombia, and Venezuela
Mar. 29, 2006	4:07	Eastern Brazil, Atlantic Ocean, Ghana, Togo, Benin, Nigeria, Niger, Chad, Libya, Egypt, Turkey, Russia
Aug. 1, 2008	2:27	Northern Canada, Greenland, Arctic Ocean, Russia, Mongolia, China
July 22, 2009	6:39	India, Nepal, Bhutan, Burma, China, Pacific Ocean
July 11, 2010	5:20	South Pacific Ocean, Easter Island, Chile, Argentina
Nov. 13, 2012	4:02	Northern Australia, south Pacific Ocean
Nov. 3, 2013*	1:40	Annular only at beginning and end of path: Atlantic Ocean, Gabon, Congo, Zaire, Uganda, Kenya, then annular in Ethiopia
Mar. 20, 2015	2:47	North Atlantic Ocean, Faeroe Islands, Arctic Ocean, Svalbard
Mar. 9, 2016	4:10	Indonesia (Sumatra, Borneo, Sulawesi, Halmahera), Pacific Ocean
Aug. 21, 2017	2:40	Pacific Ocean, United States (Oregon, Idaho, Wyoming, Nebraska, Missouri, Illinois, Kentucky, Tennessee, North Carolina, South Carolina), Atlantic Ocean
July 2, 2019	4:33	South Pacific Ocean, Chile, Argentina
Dec. 14, 2020	2:10	Pacific Ocean, Chile, Argentina, south Atlantic Ocean
Dec. 4, 2021	1:55	Antarctica
Apr. 20, 2023*	1:16	Total except at beginning and end of path: South Indian Ocean, western Australia, Indonesia, Pacific Ocean
Apr. 8, 2024	4:28	Pacific Ocean, Mexico, United States (Texas, Oklahoma, Arkansas, Missouri, Kentucky, Illinois, Indiana, Ohio, Pennsylvania, New York, Vermont, New Hampshire, Maine), southeastern Canada, Atlantic Ocean

*These are hybrid or annular-total eclipses. They are total along the middle portions of their paths but annular at the ends.
†Data courtesy of M. Littmann, K. Willcox, and K. Espenak, *Totality—Eclipses of the Sun*, Oxford University Press, 1999.

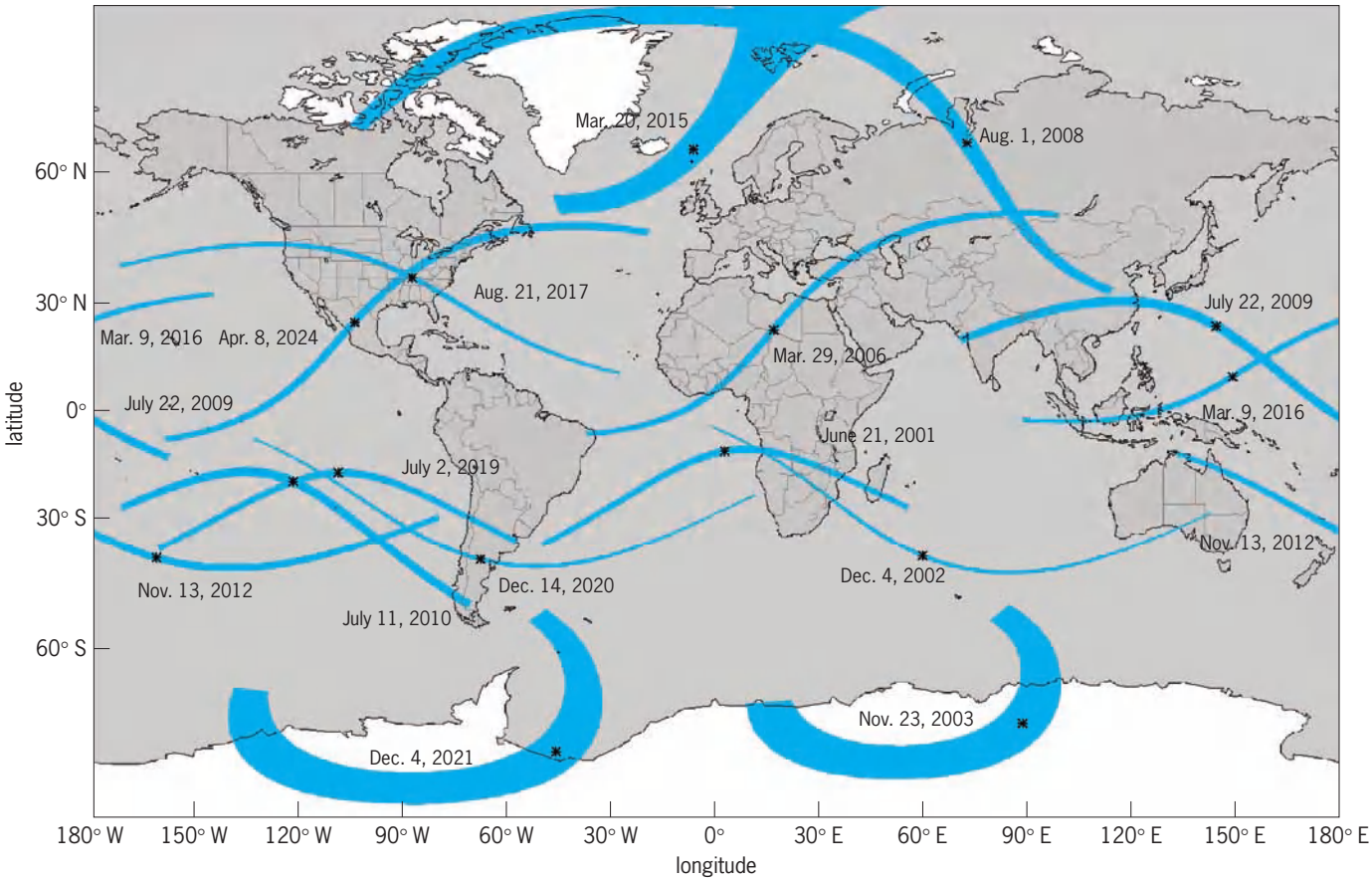


Fig. 4. Paths of all total solar eclipses from 2001 through 2025. (Paths of hybrid annular/total eclipses are not shown). The point of greatest eclipse is indicated by an asterisk on each eclipse path. (*F. Espenak, NASA's Goddard Space Flight Center, http://sunearth.gsfc.nasa.gov/eclipse/eclipse.html*)

The April 8, 2005, hybrid eclipse was annular at the ends of its path and total in the middle, over the Pacific Ocean, where the umbral cone reached the Earth's surface (Fig. 3)

Recent and future eclipses. Notable total eclipses in terms of duration and favorable weather were on June 21, 2001, in southern Africa; on December 4, 2002, in southern Africa and south Australia; and on March 29, 2006, in Africa, Turkey, and into Asia (**Table 1**; **Fig. 4**). The next total solar eclipses are on August 1, 2008, in Mongolia and China; on July 22, 2009, in China; and on July 11, 2010, on Easter Island and in southernmost South America. Major scientific expeditions carry out research on such occasions.

The next total eclipse to cross the continental United States will be in 2017, and the next eclipse to cross Canada will be in 2024.

Observing a solar eclipse. A solar eclipse is nature's most magnificent phenomenon. The false impression has grown that it is often hazardous to observe a solar eclipse, whereas, because of educational efforts on eye safety, actually few reports of eye injury exist and the incidence of lasting eye damage is even lower.

The total phase of an eclipse is completely safe to watch with the naked eye. The total brightness of the corona is only that of the full moon, so is equally safe to watch. The darkness at the disappearance of the diamond ring effect comes so abruptly that people have no trouble telling when the safe time begins, and the diamond ring at the end of totality is so relatively bright that it is clear when it is time to look away. A glance at the Sun just before or after totality is not harmful; it is only staring for an extended time (more than a few seconds with the naked eye) or looking at the not totally eclipsed Sun through binoculars or a telescope that can cause harm.

During the partial phases, it is possible to follow what is going on without any special aid by watching the ground under a tree. The spaces between the leaves make a pinhole camera, and project the solar crescent myriad times onto the ground. A pinhole camera can be made individually by punching a small hole (approximately 0.1–0.2 in. or 2–5 mm in diameter) in a piece of cardboard, and holding it up to the Sun. A crescent image is projected onto a second piece of cardboard held 8 to 40 in. (20 cm to 1 m) closer to the ground, or onto the ground itself. An observer looks at the second cardboard, facing away from the Sun.

For direct observation of the partial phases, a special solar filter must be used. Fogged and exposed black-and-white (not color) film that contains silver and is developed to full density provides suitable diminution of the solar intensity across the entire spectrum. Inexpensive commercial solar filters

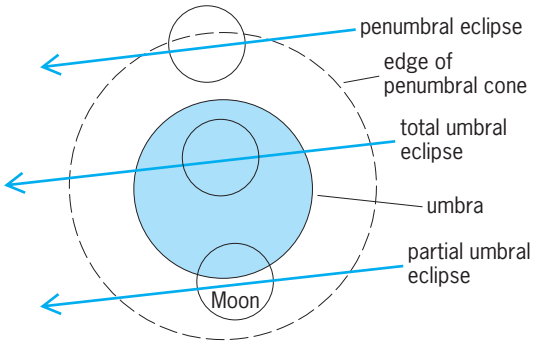


Fig. 5. Cross section through the Earth's shadow cone (umbra) at the Moon's distance, showing the motion of the Moon at three types of lunar eclipses: penumbral eclipse, partial umbral eclipse, and total umbral eclipse.

made of aluminized Mylar can also be used. Gelatin "neutral-density" filters are actually not neutral in the infrared, and so should not be used. Neutral-density filters made by depositing chromium or other metals on glass are safe if they are ND4 or ND5, and are commercially available, as is #14 welder's glass.

Jay M. Pasachoff

Lunar Eclipses

An eclipse of the Moon occurs when the Moon passes through some part of the Earth's two shadows. The inner shadow or umbra is that region where the Earth blocks all direct rays from the Sun. The outer shadow or penumbra is a zone where the Earth blocks part but not all of the Sun's rays. The two shadows result in three types of lunar eclipses (**Fig. 5**):

1. *Penumbral eclipse*. A part or all of the Moon passes through the Earth's penumbral shadow.
2. *Partial eclipse*. A part but not all of the Moon passes through the Earth's umbral shadow.
3. *Total eclipse*. The entire Moon passes through the Earth's umbral shadow.

A lunar eclipse can occur only during the full moon phase. Furthermore, the Moon must also be near one of the two nodes of its orbit. The nodes are the points where the plane of the Moon's orbit intersects the plane of the Earth's orbit. Since the Moon's orbit is tipped 5° to the Earth's orbit (the ecliptic), the Moon passes above or below the shadows during most full moons. But several times each year, full moon occurs near one of the nodes and some portion of the Moon passes through one or both of the Earth's shadows. *See* PHASE (ASTRONOMY).

All partial lunar eclipses begin and end with penumbral phases. Similarly, all total eclipses begin with a penumbral phase followed by a partial phase. Totality is then followed by a partial and a penumbral phase in that order (**Fig. 6**).

Lunar eclipses are visible from much larger geographic areas than are solar eclipses. This is due to the fact that each lunar eclipse can be viewed from the entire night hemisphere of the Earth. Unlike solar eclipses, which require special filters for eye protection, lunar eclipses are completely safe to view using the unaided eye or binoculars.

Visibility. For each type of lunar eclipse, the Moon's maximum depth into either the penumbra or umbra is characterized by a value called the eclipse magnitude. This is the fraction of the Moon's diameter immersed in the shadow. For total eclipses, the umbral eclipse magnitude is always equal to or greater than 1.0.

Penumbral eclipses can last up to 5 h. However, these events are of only academic interest since they are quite faint and difficult to see. Only the deepest penumbral eclipses (penumbral eclipse magnitude greater than 0.7) are visible to the unaided eye as a subtle shading along one edge of the Moon.

Partial eclipses are far easier to see and appear as an obvious bite or missing section along one edge of an otherwise full moon. They can last as long as $3\,{}^{1}/_{2}$ hours but are usually shorter. Partial eclipses are relatively colorless although the edge of the Earth's umbra may exhibit a reddish tinge when viewed through a telescope. The umbral eclipse magnitude of a partial eclipse is always greater than 0.0 and less than 1.0.

Total lunar eclipses have a maximum duration of nearly $2\,{}^{1}/_{2}$ hours. Their appearance is dramatically different from both penumbral and partial eclipses. When the Moon is completely immersed within the Earth's umbral shadow, it is illuminated only by indirect sunlight that has been refracted through and filtered by the Earth's atmosphere. In the process, most of the blue and green colors present in sunlight are removed. The remaining light is deep red or orange and is greatly dimmed. This gives total lunar eclipses their

Fig. 6. Stages of the total lunar eclipse of January 21, 2000. (*F. Espenak*)

Fig. 7. Total lunar eclipse of December 9, 1992. (*F. Espenak*)

characteristically blood red color, which frightened many early civilizations. If the Earth had no atmosphere, the Moon would be completely black during a total eclipse. *See* REFRACTION OF WAVES; SCATTERING OF ELECTROMAGANETIC RADIATION.

The actual appearance of the Moon during totality can vary from one eclipse to the next. The colors range from dark gray or brown, through various shades of red, to bright orange and yellow. These visual differences are directly related to the quality and quantity of dust suspended in Earth's atmosphere. Smoke from forest fires, airborne particles from dust storms, and ash from volcanic eruptions contribute to the color and darkness of the Moon during totality. When large amounts of these particulates are present, they have a strong filtering effect and produce a correspondingly darker eclipse. For example, the 1991 eruption of Mount Pinatubo in the Philippines pumped large amounts of volcanic ash 115,000 ft (35 km) high into the atmosphere where it was transported globally via the jet stream. Red sunsets were seen around the world, and the total lunar eclipse of December 1992 was so dark it rendered the Moon nearly invisible during totality (**Fig. 7**).

Frequency. The geometry and recurrence of lunar eclipses have much in common with solar eclipses. Lunar eclipses occur during eclipse seasons lasting about 39 days with a cycle of two seasons every 346.6 days. Lunar eclipses with similar characteristics repeat every 6585.32 days (about 18 years 11 days), which is the same Saros cycle described above for solar eclipses.

Every lunar eclipse is either preceded or followed by a solar eclipse within a time span of 15 days. Of course, the lunar eclipse will occur at full moon, while the solar eclipse will occur 2 weeks earlier or later at new moon. The link between lunar and solar eclipses means that there are a maximum of seven eclipses within a calendar year. They can occur in combinations of five and two, or four and three, with either solar or lunar eclipses in the majority.

In an average century, there are 243 lunar eclipses consisting of 89 penumbral, 84 partial, and 70 total eclipses. Every year has a minimum of two and a maximum of five lunar eclipses.

Future lunar eclipses. The twenty-first century will have 230 eclipses of the Moon, consisting of 87 penumbral, 58 partial, and 85 total eclipses. During this epoch, 76 years will have two lunar eclipses each, 18 years will have three eclipses, and 6 years will have four eclipses. No year contains more than four eclipses. The last time five lunar eclipses occurred during one calendar year was 1879, and the next will be 2132.

During the 22-year period 2004 through 2025, there will be 20 total eclipses of the Moon (**Table 2**). Deep eclipses have large magnitudes and long

TABLE 2. Total lunar eclipses, 2004–2025*

Date†	Umbral magnitude	Duration of totality	Geographic region of eclipse visibility
May 4, 2004	1.309	1 h 16 min	South America, Europe, Africa, Asia, Australia
Oct. 28, 2004	1.313	1 h 21 min	Americas, Europe, Africa, Central Asia
Mar. 3, 2007	1.238	1 h 14 min	Americas, Europe, Africa, Asia
Aug. 28, 2007	1.481	1 h 31 min	Eastern Asia, Australia, Pacific, Americas
Feb. 21, 2008	1.111	0 h 51 min	Central Pacific, Americas, Europe, Africa
Dec. 21, 2010	1.262	1 h 13 min	Eastern Asia, Australia, Pacific, Americas, Europe
June 15, 2011	1.705	1 h 41 min	South America, Europe, Africa, Asia, Australia
Dec. 10, 2011	1.110	0 h 52 min	Europe, Eastern Africa, Asia, Australia, Pacific, North America
Apr. 15, 2014	1.296	1 h 19 min	Australia, Pacific, Americas
Oct. 8, 2014	1.172	1 h 00 min	Asia, Australia, Pacific, Americas
Apr. 4, 2015	1.006	0 h 12 min	Asia, Australia, Pacific, Americas
Sept. 28, 2015	1.282	1 h 13 min	Eastern Pacific, Americas, Europe, Africa, western Asia
Jan. 31, 2018	1.321	1 h17 min	Asia, Australia, Pacific, western North America
July 27, 2018	1.614	1 h 44 min	South America, Europe, Africa, Asia, Australia
Jan. 21, 2019	1.201	1 h 03 min	Central Pacific, Americas, Europe, Africa
May 26, 2021	1.016	0 h 19 min	Eastern Asia, Australia, Pacific, Americas
May 16, 2022	1.419	1 h 26 min	Americas, Europe, Africa
Nov. 8, 2022	1.364	1 h 26 min	Asia, Australia, Pacific, Americas
Mar. 14, 2025	1.183	1 h 06 min	Pacific, Americas, western Europe, western Africa
Sept. 7, 2025	1.367	1 h 23 min	Europe, Africa, Asia, Australia

*Data courtesy of Fred Espenak, http://sunearth.gsfc.nasa.gov/eclipse/eclipse.html.
†Date of mid-eclipse (Greenwich Mean Time).

durations (for example, June 15, 2011, and July 27, 2018). Fred Espenak

Bibliography. F. Espenak, *Fifty Year Canon of Lunar Eclipses*: *1986–2035*, NASA RP-1178, Sky Publishing, 1989; F. Espenak, *Fifty Year Canon of Solar Eclipses*: *1986–2035*, NASA RP-1216, Sky Publishing, 1987; L. Golub and J. M. Pasachoff, *Nearest Star: The Surprising Science of Our Sun*, Harvard University Press, 2001; P. Guillermier and S. Koutchmy, *Total Eclipes: Science, Observations, Myths and Legends*, Springer-Praxis, 1999; M. Littman, K. Willcox, and F. Espenak, *Totality: Eclipses of the Sun*, 2d ed., Oxford University Press, 1999; B.-L. Liu and A. D. Fiala, *Canon of Lunar Eclipses 1500 B.C.—A.D. 3000*, Willmann-Bell, 1992; J. Meeus and H. Mucke, *Canon of Lunar Eclipses: −2002 to +2526*, 3d ed., Astronomisches Büro, Vienna, 1992; H. Mucke and J. Meeus, *Canon of Solar Eclipses: −2003 to +2526*, 2d ed., Astronomisches Büro, Vienna, 1992; J. M. Pasachoff, *A Field Guide to the Stars and Planets*, 4th ed., updated, Houghton Mifflin, 2006; J. M. Pasachoff, *The Complete Idiot's Guide to the Sun*, Alpha Books, 2003; J. B. Zirker, *Total Eclipses of the Sun*, expanded ed., Princeton University Press, 1995.

Eclipsing variable stars

Double star systems in which the two components are too close to be seen separately but which reveal their duplicity by periodic changes in brightness as each star successively passes between the other and the Earth, that is, eclipses the other. Studies of the light changes and the radial velocity changes of each component permit the computation of the radii, masses, and densities of the components—important quantities that cannot be measured directly in single stars. In addition, these close double stars are useful in studies of mass loss and of stellar evolution. Since eclipsing stars are variable in light, they are included in general variable star catalogs under the same system of nomenclature. *See* BINARY STAR; VARIABLE STAR.

Periods. The periods of light variation range from less than 3 h for very close systems to over 27 years for the peculiar system Epsilon Aurigae. However, the majority of the periods lie between 0.5 and 10 days. In many cases the periods are not constant but change with time. In a few cases the variation is caused by a slow change in the orientation of the major axis of an elliptical orbit; in such cases the rate of change combined with other quantities gives information concerning the manner in which the density of the star increases from the outer layers to the center. In most cases, however, the changes are unpredictable and are probably connected with ejections of matter from one of the stars.

Velocity curves. The radial velocity (velocity of approach or recession) of each star can be determined at any time by the displacement of the spectral lines. A plot of velocities against time over one period of orbital revolution is known as a velocity curve. The maximum radial velocity of approach or recession depends on the true orbital velocity of the star and the "inclination," or the amount by which the plane of the orbit is tilted relative to the line of sight to the Earth. (Technically, the inclination is the angle between a perpendicular to the orbit plane and the line of sight; when the inclination is 90°, the eclipses are central.) If the inclination is known, the orbital velocity of each star can be calculated from the radial velocity. The orbital velocity multiplied by the period will give the circumference of the orbit and from this the radius of each star about the center of mass. From the size of the orbit and the period, by using the law of gravitation, the mass of each star can be calculated in terms of the Sun's mass. However, the inclination cannot be determined from the velocity curve alone; the quantities finally determined are $m_1 \sin^3 i$, $m_2 \sin^3 i$, and $a \sin i$, where i is the inclination, $m_1$and m_2 the stellar masses, and a the radius of a circular orbit or half the major axis of an elliptical one.

Light curve. The light curve shows the changes in brightness of the system throughout one orbital revolution. The manner in which the light changes during the eclipse of each star by the other depends very strongly on three factors: the size of each star relative to the radius of the orbit (r_1/a andr_2/a) and the inclination. The determination of precisely what relative sizes and inclination give a computed curve which will approximate satisfactorily the observations is one of the more difficult problems of modern astronomy. However, by use of complex sets of tables the problem can be solved in many cases. Numerical values for these three quantities are thus found. Several alternate methods have been developed which make use of the capabilities of electronic computers. These take into account simultaneously the various interaction effects such as distortion of the shape of the components in the earlier methods. These had to be removed from the light curve before a solution could be carried out. *See* ECLIPSE; LIGHT CURVES.

Absolute dimensions. The masses, radii, and densities of the stars, usually expressed in comparison with those of the Sun, are called the absolute dimensions of the system. Combining the results from the light and velocity curves yields this fundamental information which cannot be obtained from either approach alone. The inclination, determined from the shape of the light curve, can be substituted in the quantities $m_1 \sin^3 i$, $m_2 \sin^3 i$, and $a \sin i$ to find m_1, m_2, and a. Thus the masses of the stars in terms of the Sun's mass and the size of the orbit in terms of the Sun's radius can be found. Since the radii of the stars in fractions of the size of the orbit have been determined from the light curve, the sizes of the stars relative to the size of the Sun (either in miles or kilometers) can now be computed. **Figure 1** shows the relative sizes and separation of the stars in a typical eclipsing system.

Complications. Before the light curve can be "solved" to give the above quantities, corrections must be made for other factors which influence

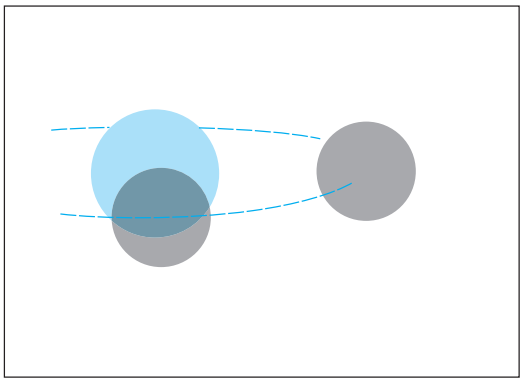


Fig. 1. Relative sizes and separation of components of R Canis Majoris. Shaded star represents cooler component. It is shown in two positions: at the middle of an eclipse and when separation would be greatest as viewed from Earth. Even at the greatest separation, all such systems appear as single stars.

the light changes. This is done by studying the light between eclipses where, were it not for these complications, the brightness of the system would not change. One of these effects is called ellipticity. The tidal attraction of each star for the other has caused distortions until the stars in extreme cases resemble footballs more than baseballs in shape. The technical term is prolate ellipsoids, although when the stars differ in size and mass they will also differ in shape. Further, the radiation of each star falling on the side of the other nearest it will cause this side to be brighter than the side turned away. The difference will be most marked for the cooler component where the effect of the intense radiation from the nearby hotter star is most strongly evident.

There are other effects, some very poorly understood, which cause light changes between eclipses, and all of these must be carefully studied before the analysis of the eclipse begins.

Evolutionary changes. Studies of single stars indicate that, when the hydrogen in the center of the star has been converted into helium, the star undergoes a relatively rapid expansion in size. The presence of a nearby companion complicates the picture considerably, but it does seem clear that much of the mass of the star must be lost to the system or possibly transferred to the other component.

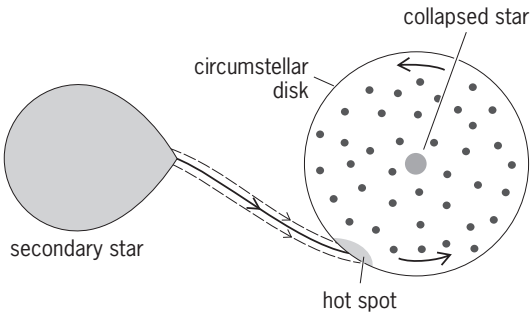


Fig. 2. Model generally accepted for certain types of eclipsing binary stars, in particular those known as dwarf novae.

The mass-losing star eventually becomes a collapsed object—a white dwarf, a neutron star, or a black hole. Each of these types has been identified in at least one binary system. Then when the secondary, originally less massive star begins its expansion, matter from it is transferred to the collapsed object, often with dramatic results. *See* BLACK HOLE; NEUTRON STAR; WHITE DWARF STAR.

It is now believed that all explosive variables (novae, recurrent novae, and so forth), with the exception of supernovae, are members of close binary systems. At least some of the x-ray sources are close binaries in this state, although the detection (by instruments carried aboard satellites) of x-radiation from Algol and other systems which are not yet in this state indicates that other physical mechanisms may also be responsible. Some systems show intermittent bursts of radiation at radio frequencies. Evidence indicates the presence of clouds of circumstellar material.

In a few of the eruptive variables, particularly those known as dwarf novae, rapid scintillation is found, presumably from a hot spot where the transferring mass collides violently with a circumstellar disk of relatively low-density material revolving around the collapsed star (**Fig. 2**); the scintillation stops periodically when the spot is eclipsed by the other component. Instruments on satellites have extended observations to the far ultraviolet, as well as the x-ray, regions of the spectrum. Thus, in addition to the classical reasons for studying eclipsing variable stars, observation of them leads into many branches of astrophysics. *See* ASTROPHYSICS, HIGH-ENERGY; CATACLYSMIC VARIABLE; NOVA; STAR; STELLAR EVOLUTION; X-RAY ASTRONOMY.

Frank Bradshaw Wood

Bibliography. A. H. Batten (ed.), Algols, *Space Sci. Rev.*, 50 (1–2):384, 1989; A. H. Batten, *Binary and Multiple Systems of Stars*, 1973; E. F. Milone, *Light Curve Modeling of Eclipsing Binary Stars*, 1995; J. Sahade and F. B. Wood,*Interacting Binary Stars*, 1978.

Ecliptic

The apparent path of the Sun across the sky over the course of a year. The Earth's mean orbit defines the ecliptic plane, and an observer on Earth sees the Sun, traveling in this plane over the course of a year, follow the ecliptic. The orbits of the Moon and planets are at slight angles to the ecliptic, but these objects never appear too far in the sky from the ecliptic, which is often drawn on star maps. The ecliptic was identified by Greek and Babylonian scientists in about the fifth century B.C.

The ecliptic intersects the celestial equator at an angle of approximately $23\frac{1}{2}^\circ$, an angle known as the obliquity of the ecliptic. The points of intersection are the vernal equinox and the autumnal equinox. As a result of this obliquity, the Sun can be overhead at different times of year for locations on Earth between latitudes $+23\frac{1}{2}^\circ$ and $-23\frac{1}{2}^\circ$, that is, between the Tropic of Cancer and the Tropic of

Capricorn, respectively. *See* ASTRONOMICAL COORDINATE SYSTEMS; EARTH ROTATION AND ORBITAL MOTION. Jay M. Pasachoff

Bibliography. J. Mitton, *Cambridge Dictionary of Astronomy*, Cambridge University Press, 2001; J. M. Pasachoff and A. Filippenko, *The Cosmos: Astronomy in the New Millennium*, 3d ed., Brooks/Cole, 2007; P. K. Seidelmann (ed.), *Explanatory Supplement to the Astronomical Almanac*, University Science Books, 1992.

Eclogite

A very dense rock composed of red-brown garnet and the grape-green pyroxene omphacite. Eclogites possess basaltic bulk chemistry, and their garnets are rich in the components pyrope ($Mg_3Al_2Si_3O_{12}$), almandine ($Fe_3Al_2Si_3O_{12}$), and grossular ($Ca_3Al_2Si_3O_{12}$), while the pyroxenes are rich in jadeite ($NaAlSi_2O_6$) and diopside ($CaMgSi_2O_6$). Accessory phases depend on the environment of crystallization, but may include quartz, kyanite, amphibole, olivine, orthopyroxene, zoisite, mica, and rutile, but never plagioclase.

Occurrences. Eclogite occurrences may be subdivided into three broad categories:

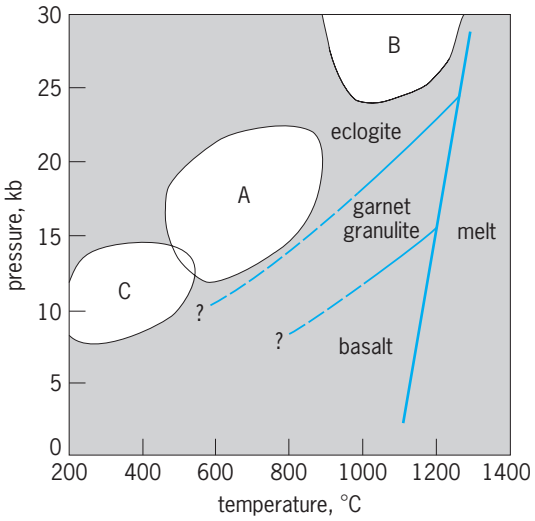
Group A eclogites are found as layers, lenses, or boudins in schists and gneisses seemingly of the amphibolite facies. Quartz, together with zoisite or kyanite, commonly occurs in these rocks. Amphibole of barroisitic composition may also be present. These rocks originate as metamorphosed basic bodies (dykes, sills, flows) or layers of dolomitic marl intercalated with metasedimentary schists. *See* QUARTZ.

Group B eclogites are found as inclusions in kimberlites and basalts. They are frequently accompanied by xenoliths of garnet peridotite; both eclogites and peridotites can be diamond-bearing, suggesting that these rocks come from deep in the mantle. *See* BASALT.

Group C eclogites are found as blocks and lenses in schists of the glaucophane-schist facies. Such eclogites do not contain kyanite, rarely contain quartz, but bear amphibole, epidote, rutile, or sphene. The presence of pillow lavas converted to eclogite indicates that these rocks had a crustal origin.

Conditions of formation. Eclogites can form over a wide range of pressure and temperature, but are clear indicators of rather high pressure, as their very high density might suggest. Experimental work has shown that rocks of basaltic composition are converted to eclogite at high pressures (see **illus.**).

The distribution of Fe^{2+} and Mg^{2+} between garnet and omphacite has been determined as a function of temperature, and allows approximate temperature regimes to be assigned to the three categories of eclogite. Eclogites of group C formed at relatively low temperatures and probably crystallized isofacially with rocks of the glaucophaneschist facies. Group A eclogites probably formed over a wider range of temperature and pressure, but at somewhat higher temperatures than group C. The very high temperatures estimated for group B eclogites are consistent with their origin in the upper mantle. *See* GARNET; OMPHACITE.



Possible pressure-temperature fields for eclogites from gneiss areas: A from kimberlites; B from blueschist terrains; C basalt-eclogite transition. 1 kilobar = 10^5 kilopascals. $°F = (°C \times 1.8) + 32$.

Eclogite facies. Eclogite is the name given to the highest-pressure facies of metamorphism; the critical mineral assemblage defining this facies is garnet + omphacite, together with kyanite or quartz in rocks of basaltic composition. Where sedimentary and granitic rocks have been metamorphosed under eclogite facies conditions, they result in spectacular omphacite + garnet + quartz-bearing mica schists and metagranitic gneisses such as are found in the Sezia-Lanzo zone of the western Alps. *See* METAMORPHIC ROCKS.

The relationship between the eclogite and other facies is often unclear. Eclogites, for instance, can form in relatively anhydrous parts of blueschist and amphibolite facies terrains where normally their hydrated equivalents would crystallize. Some such eclogites are inferred, on textural grounds, to have formed in a dry environment. In other cases the coarse textures, the presence of veins of diagnostic eclogite facies minerals such as omphacite + kyanite + quartz, and fluid inclusions suggest that their formation was accompanied by copious fluid. Although very high pressures are required to stabilize eclogite under hydrous conditions, the presence of diluent species (for example, CO_2 or halides) in the fluid lowers the required pressures.

Eclogites and the mantle. The high density of eclogite, together with its elastic properties, makes it a candidate for upper mantle material. Large quantities of basaltic oceanic crust are returned to the mantle through the process of subduction, where prevailing high pressures convert it to eclogite. The quantity and distribution of eclogite within the mantle is not known; that it occurs is known from the nodules brought up in kimberlite pipes and in basalts. *See* HIGH-PRESSURE PROCESSES. Timothy J. B. Holland

Bibliography. M. G. Best, *Igneous and Metamorphic Petrology*, 1982; B. W. Evans and E. H. Brown (eds.), *Blueschists and Eclogites*, 1986; L. A. Raymond, *Petrology: The Study of Igneous, Sedimentary, and Metamorphic Rocks*, 1994; D. C. Smith (ed.), *Eclogites and Eclogite-Facies Rocks*, 1989; M. Suk, *Petrology of Metamorphic Rocks*, 1983.

Ecological communities

Assemblages of living organisms that occur together in an area. The nature of the forces that knit these assemblages into organized systems and those properties of assemblages that manifest this organization have been topics of intense debate among ecologists since the early years of this century. On the one hand, there are those who view a community as simply consisting of species with similar physical requirements, such as temperature, soil type, or light regime. The similarity of requirements dictates that these species be found together, but interactions between the species are of secondary importance and the level of organization is low. On the other hand, there are those who conceive of the community as a highly organized, holistic entity, with species inextricably and complexly linked to one another and to the physical environment, so that characteristic patterns recur, and properties arise that one can neither understand nor predict from a knowledge of the component species. In this view, the ecosystem (physical environment plus its community) is as well organized as a living organism, and constitutes a superorganism. Between these extremes are those who perceive some community organization but not nearly enough to invoke images of holistic superorganisms. *See* ECOSYSTEM.

Emergent and collective properties. The crux of this debate is whether communities have emergent properties, a concept that itself is sufficiently confused, so that part of the ecological debate may be semantic. Some denote by emergent property any property of a group that is not a property of the component individuals. Others call any group trait a collective property, and reserve as emergent only those collective properties that do not derive trivially from properties of component individuals or from the very definition of the group. An emergent property, in this conception, represents a new level of organization, and cannot be predicted even with complete knowledge of all properties of individuals in the group. Thus, the number of species in a community is a property that is not defined for any particular species, but once one knows all the species in a community one can easily measure the property "number of species" for that community. In the restricted view of emergent property, the number of species in a community is a collective property, but not an emergent one.

Whether any properties of communities are emergent is debatable, but all ecologists agree that communities have collective properties that are significant both biologically and practically. For example, it is not a trivial matter to understand why the number of species in a given community is S and not $2S$, and what the consequences are of having S species. Also, human exploitation of ecological communities often rests on a thorough knowledge of certain collective properties, such as biomass produced per unit time, or rate of cycling of certain nutrients. *See* BIOMASS.

Size. Every community comprises a given group of species, and their number and identities are distinguishing traits. Most communities are so large that it is not possible to enumerate all species; microorganisms and small invertebrates are especially difficult to census. However, particularly in small, well-bounded sites such as lakes or islands, one can find all the most common species and estimate their relative abundances. The number of species is known as species richness, while species diversity refers to various statistics based on the relative numbers of individuals of each species in addition to the number of species. The rationale for such a diversity measure is that some communities have many species, but most species are rare and almost all the individuals (or biomass) in such a community can be attributed to just a few species. Such a community is not diverse in the usual sense of the word. Patterns of species diversity abound in the ecological literature; for example, pollution often effects a decrease in species diversity. The most popular species diversity statistic is given in the equation below, where

$$H' = \sum_{i=1}^{s} p_i \log (p_i)$$

S is the number of species and p_i is the fraction of all the individuals (or all the biomass) in the community contributed by species i. For many sets of data on diversities of groups of communities (or parts of communities, such as all birds or all plants), values of H' are highly correlated with species richness, suggesting that H' may not be the ideal expression of biotic diversity.

The main patterns of species richness that have been detected are area and isolation effects, successional gradients, and latitudinal gradients. Larger sites tend to have more species than do small ones, and isolated communities (such as those on oceanic islands) tend to have fewer species than do less-isolated ones of equal size. Later communities in a temporal succession tend to have more species than do earlier ones, except that the last (climax) community often has fewer species than the immediately preceding one. Tropical communities tend to be very species-rich, while those in arctic climates tend to be species-poor. This observation conforms to a larger but less precise rule that communities in particularly stressful environments tend to have few species.

Composition. Species composition is important, largely in assisting in the classification of communities. If the complete species list for each community and the proportional representation of each species

in it were known, no two communities would be found to be identical. The criteria are arbitrary for how similar these lists must be before two communities are viewed as the same.

Communities are usually denoted by the presence of species, known as dominants, that contain a large fraction of the community's biomass, or account for a large fraction of a community's productivity. Dominants are usually plants. Determining whether communities at two sites are truly representatives of the "same" community requires knowledge of more than just the dominants, however. "Characteristic" species, which are always found in combination with certain other species, are useful in deciding whether two communities are of the same type, though the designation of "same" is arbitrary, just as is the designation of "dominant" or "characteristic."

Although there is no objective answer to the question of whether two communities are the same, the question bears heavily on the matter of whether communities are in fact integrated superorganisms with emergent properties. One would not expect a superorganismic entity to have an infinite variety of representations, for the same cohesive forces that bind it into a holistic unit should constrain it to a limited number of forms. Organisms for the most part do not form a continuous gradient of all conceivable phenotypes, but rather are constrained by a number of physiological and morphological relationships (such as allometry). Thus there is generally no trouble in saying that individual X is a human, individual Y is a longleaf pine, and so on. Although there is phenotypic variation in all species, it is limited and centered on recognizable modes.

An active area of research among animal ecologists has been the search for limits to how similar two species can be and still occur in the same community. That there exists some limiting similarity beyond which coexistence is impossible is a theoretical consequence of conceptions of the ecological niche; but clear evidence, in any instance, that such limiting similarity is the reason why a given species is absent from a given community is exceedingly difficult to amass. There are two main difficulties. First, it must be demonstrated that some resource is limiting so that species that are too similar will find it in short supply. Second, an appropriate index of a species' use of the resource (its niche) must be determined. Much attention has been focused on the size of an animal or of its trophic appendages as such an index, though there is much evidence that many other forces affect these sizes. For few, if any, animal communities has it been shown that size similarity restricts species composition or that patterns of size differences among species are highly predictable or regular.

Spatial distribution patterns. A related matter bearing on the aptness of the superorganismic community metaphor is that most organisms have clear spatial boundaries; communities often do not. There is no difficulty discerning where one human ends and another begins, but there are not always clear boundaries to communities. Occasionally, very sharp limits to a physical environmental condition impose similarly sharp limits on a community. For example, serpentine soils are found sharply delimited from adjacent soils in many areas, and have mineral concentrations strikingly different from those of the neighboring soils. Thus they support plant species that are very different from those found in nearby nonserpentine areas, and these different plant species support animal species partially different from those of adjacent areas. Here two different communities are sharply bounded from each other.

Usually, however, communities grade into one another more gradually, through a broad intermediate region (an ecotone) that includes elements of both of the adjacent communities, and sometimes other species as well that are not found in either adjacent community. One has little difficulty telling when one is in the center of either of the adjacent communities A and B, but exactly when one passes from A to B is not easily discerned. The reason is that, though each species in a community is adapted to its physical environment and to some extent to other species in the community, the adaptations to the physical environment are usually not identical, and most of the adaptations to one another are not obligatory.

The environment created by the dominant species, by their effects on temperature, light, humidity, and other physical factors, and by their biotic effects, such as allelopathy and competition, may entrain some other species so that these other species' spatial boundaries coincide with those of the dominants. The mangrove skipper, *Phocides pygmalion*, can feed only on red mangrove, *Rhizophora mangle*, so whatever aspects of the physical environment limit the mangrove (especially temperature), the skipper will also be limited to the same sites. However, many other species that feed on the mangrove (for example, the io moth, *Automeris io*) also feed on other plants, and their spatial boundaries do not coincide with those of the mangrove. Nor do most of the species in a community share identical physical requirements. Black mangrove (*Avicennia germinans*) co-occurs with red mangrove in most sites, and the two are normal constituents of a community often termed mangrove swamp. But *Avicennia* can tolerate much colder weather than can *Rhizophora*, and so it is found much farther north in the Northern Hemisphere. Eventually it too ceases, and the mangrove community is replaced in more northerly areas by salt marsh communities. In some intermediate regions, salt marsh grasses and black mangrove are found together. There is no clear boundary between the two communities.

This continuous intergradation of most communities argues against the superorganism concept, but there are aspects of the spatial arrangement within communities that suggest that the component species are far from independent, and indicate, if not complete holism, at least that some community properties are not easily predicted from exhaustive knowledge of the component species. One example

is stratification, the vertical arrangement of canopy layers in most forests. Individuals of the different species do not have heights that are independently and continuously distributed from the ground to the top of the tallest tree. Instead, there are a few rather distinct strata, with each species at maturity characteristically occupying one of these. Tropical forests from all parts of the world, even though they may have completely different species compositions, usually contain five fairly clear strata: a topmost layer composed of the tallest tree species, two lower layers of smaller trees, a shrub layer, and a ground layer. There are doubtless good physical reasons why the diffusion of light can explain this characteristic structure given a knowledge of evolution and plant physiology (though no completely compelling explanation has yet surfaced), so it may be that this is an elaborate collective property rather than an emergent one. In either case, there is clearly a high degree of multispecies organization in this spatial arrangement. *See* PHYSIOLOGICAL ECOLOGY (PLANT).

In addition to vertical arrangement, horizontal locations of individuals of different species are usually not random. Usually, individuals of a given species are clumped; they are found on average closer to one another than one would have predicted. Probably the major reason for this is response to habitat heterogeneity: conspecific individuals tend to favor more similar habitats than do heterospecific individuals. Individuals of different species may also be nonrandomly arranged with respect to one another. Competitive interactions may cause two species typically to be found in different microsites, while mutualistic interactions or preference for a similar physical habitat may cause two species to be associated spatially. *See* POPULATION ECOLOGY.

Succession. More or less distinct communities tend to follow one another in rather stylized order. As with recognition of spatial boundaries, recognition of temporal boundaries of adjacent communities within a sere is partly a function of the expectations that an observer brings to the endeavor. Those who view communities as superorganisms are inclined to see sharp temporal and spatial boundaries, and the perception that one community does not gradually become another community over an extended period of time confirms the impression that communities are highly organized entities, not random collections of species that happen to share physical requirements. In this superorganismic view, ecological succession of communities in a sere is analogous to the life cycle of an organism, and follows a quite deterministic course. That secondary succession following a disturbance often leads to a community resembling the original one is the analog in the superorganism to wound repair in an organism. The driving force for succession, in this conception, is that the community or its dominant species modify the environment so that these dominant species are no longer favored, and when the dominant species are replaced, the bulk of the community, complexly linked to the dominant species and to one another, disappears as well, to be replaced by the next community in the sere.

This superorganismic conception of succession has been replaced by an individualistic succession. Data on which species are present at different times during a succession show that there is not abrupt wholesale extinction of most members of a community and concurrent simultaneous colonization by most species of the next community. Rather, most species within a community colonize at different times, and as the community is replaced most species drop out at different times. Thus, though one can usually state with assurance that the extant community is of type A or type B, there are extended periods when it is difficult to assign the assemblage of species at a site to any recognizable community.

That succession is primarily an individualistic process does not mean that there are not characteristic changes in community properties as most successions proceed. Species richness usually increases through most of the succession, for example, and stratification becomes more highly organized and well defined. A number of patterns are manifest in aspects of energy flow and nutrient cycling. *See* ECOLOGICAL SUCCESSION.

Functional organization. Living organisms are characterized not only by spatial and temporal structure but by an apparent purpose or activity. In short, they are "doing something," and this activity has been termed teleonomy. In humans, for example, various physiological functions are continuously under way until death intervenes. Communities have functions analogous to physiology.

In the first place, the various species within a community have different trophic relationships with one another. One species may eat another, or be eaten by another. A species may be a decomposer, living on dead tissue of one or more other species. Some species are omnivores, eating many kinds of food; others are more specialized, eating only plants or only animals, or even just one other species. These trophic relationships certainly unite the species in a community into a common endeavor, the transmission of energy through the community. This energy flow is patently analogous to an organism's mobilization and transmission of energy from the food it eats.

One aspect of energy flow is a candidate for an emergent property: the topology of the food web. Examination of this topology for a number of webs suggests that they are highly constrained in structure. For example, the maximum number of trophic levels in a web rarely exceeds five. One reason may be that the total amount of energy that has not already been degraded by the time the energy has passed through three or four levels may not be enough to sustain a viable population of a species that would feed at still higher levels. An alternative explanation is that the population dynamics of a web with so many levels would probably confer mathematical instability on the web, so that one or more species would be eliminated. Other properties of food webs that have been discerned include a low number of

omnivore species (those feeding on more than one level), and a tendency for the number of predator species and the number of prey species to be in the ratio of 4:3. No explanation is forthcoming for the latter observation. The former observation is held to reflect mathematical instability that arises from the presence of omnivores in a web. Other workers contend that neither pattern will be maintained when much more comprehensive data are available on what paths calories actually follow as they flow through a community. If these patterns do not turn out to be artifacts of incomplete knowledge, they would appear to be emergent properties reflecting a high degree of organization. *See* FOOD WEB.

Just as energy flows through the communities, so do nutrients move. A calorie of energy does not move in the abstract from organism to organism; rather, the calorie is bound up in a molecule that moves when one organism eats (or decomposes) another. Or the calorie may be respired away as a result of the metabolism of the organism that ingests it. A calorie of energy, once respired by some member of the community, is no longer available to the community. But the molecule associated with that calorie, or a new molecule produced from it, is still present and can go through the food web again. Thus nutrients can cycle within a community, while energy flow, once the energy is transformed to heat, is one-way. Nutrient cycling is analogous to circulation in an organism, and combines with energy flow to support the superorganism metaphor. Different nutrients cycle at different rates, and for several nutrients the cycle within the community is linked to cycles in other communities. Nutrients exist in abiotic entities, as well as biotic organisms, so nutrient cycling is as much an ecosystem trait as a community one. A number of properties of nutrient cycling (such as rate, turnover times, and sizes of different pools or compartments) have been measured. Whether any of these are emergent as opposed to collective properties has yet to be determined.

Productivity. By virtue of differing rates of photosynthesis by the dominant plants, different communities have different primary productivities. Tropical forests are generally most productive, while extreme environments such as desert or alpine conditions harbor rather unproductive communities. Agricultural communities are intermediate. Algal communities in estuaries are the most productive marine communities, while open ocean communities are usually far less productive. The efficiency with which various animals ingest and assimilate the plants and the structure of the trophic web determine the secondary productivity (production of organic matter by animals) of a community. Marine secondary productivity generally exceeds that of terrestrial communities. *See* AGROECOSYSTEM; BIOLOGICAL PRODUCTIVITY.

Reproduction. A final property that any organism must have is the ability to reproduce itself. Communities may be seen as possessing this property, though the sense in which they do so does not support the superorganism metaphor. A climax community reproduces itself through time simply by virtue of the reproduction of its constituent species, and may also be seen as reproducing itself in space by virtue of the propagules that its species transmit to less mature communities. For example, when a climax forest abuts a cutover field, if no disturbance ensues, the field undergoes succession and eventually becomes a replica of the adjacent forest. Both temporally and spatially, then, community reproduction is a collective rather than an emergent property, deriving directly from the reproductive activities of the component species. *See* ALTITUDINAL VEGETATION ZONES; BOG; CHAPARRAL; DESERT; ECOLOGY; GRASSLAND ECOSYSTEM; MANGROVE; MUSKEG; PARAMO; PUNA.

Daniel Simberloff

Bibliography. T. F. H. Allen and T. B. Starr, *Hierarchy: Perspectives for Ecological Complexity*, 1988; M. Begon, M. Mortimer and D. J. Thompson, *Population Ecology: A Unified Study of Animals and Plants*, 3d rev. ed., 1996; P. G. Digby and R. A. Kempton, *Multivariate Analysis of Ecological Communities*, 1987; J. R. Trabalka and D. E. Reichle (eds.), *The Changing Carbon Cycle*, 1986.

Ecological competition

The interaction of two (or more) organisms (or species) such that, for each, the birth or growth rates are depressed and the death rate increased by the presence of the other organisms (or species). Competition is recognized as one of the more important forces structuring ecological communities, and interest in competition led to one of the first axioms of modern ecology, the competitive exclusion principle. The principle suggests that in situations where the growth and reproduction of two species are resource-limited, only one species can survive per resource.

Lotka-Volterra equations. The competitive exclusion principle was originally derived by mathematicians using the Lotka-Volterra competition equations. This model of competition predicts that if species differ substantially in competitive ability, the weaker competitor will be eliminated by the stronger competitor. However, a competitive equilibrium can occur if the negative effect of each species on itself (intraspecific competition) is greater than the negative effect of each species on the other species (interspecific competition). Because the competitive exclusion principle implies that competing species cannot coexist, it follows that high species diversity depends upon mechanisms through which species avoid competition. *See* MATHEMATICAL ECOLOGY.

Species coexistence. One way that the problem of competitive coexistence has been addressed is to extend the competitive exclusion principle to a continuum of resources through the theory of limiting similarity. Limiting similarity is defined as the maximum level of similarity in the use of a set of resources that will allow competing species to coexist. However, this approach still assumes that several species

cannot coexist on the same resources unless they use those resources differently.

A more realistic approach to understanding the coexistence of competing species is to view community structure as dynamic, in which species densities do not remain constant over time at each spatial location. Fluctuations in densities may be a result of floods, fires, and other abiotic disturbances, or of biotic interactions, such as predation, herbivory, and disease. They key to coexistence within a homogeneous patch in such a dynamic world is the prevention of competitive equilibrium and hence competitive exclusion.

In general, competitive exclusion can be prevented if the relative competitive abilities of species vary through time and space. Such variation occurs in two ways. First, dispersal rates into particular patches may fluctuate, causing fluctuations in the numerical advantage of a species in a particular patch. This may occur, for example, through differences among species in dispersal ability. Second, competitive abilities of species may be environmentally dependent and, therefore, fluctuate with local environmental changes. Different species will be favored under different sets of environmental conditions, allowing each species to have periods of strong recruitment.

Competitive exclusion can also be avoided if fluctuations in environmental factors reduce the densities of potentially competing species to levels where competition is weak and population growth is for a time insensitive to density. Therefore, intensities of competition within and between species fluctuate with time, and competing species are able to coexist.

In patchy environments, local extinctions or population reductions resulting from predators or environmental fluctuations can allow species to colonize and grow for some time in patches in the absence of competition. In this way, even inferior competitors can persist if they are, for example, relatively less susceptible to predation or are relatively better dispersers.

Coexistence is not merely a result of environmental harshness or fluctuations but also involves the critical element of niche differentiation (that is, species must differ from one another if they are to coexist). However, the focus is not how species coexist by partitioning resources, but how species can coexist on the same resources by differing sufficiently in their responses to environmental conditions and fluctuations. *See* ECOLOGICAL COMMUNITIES; ECOLOGICAL SUCCESSION.

Applications. Competition theory has been applied to human-manipulated ecosystems used to produce food, fiber, and forage crops as well as in forestry and rangeland management. Although many characteristics of agricultural systems are similar to those of natural ecosystems, agricultural communities are unique because they are often managed for single-species (sometimes multispecies) production and they are usually characterized by frequent and intense disturbance. Studies of competition in agriculture have primarily examined crop loss from weed abundance under current cropping practices, and have evaluated various weed control tactics and intercropping systems. Because both total density and species proportion influence the outcome and interpretation of competition experiments, the ability to quantify the intensity of competition and to separate the effects of competition between individuals of the same species versus competition between individuals of different species (such as weed versus crop) depends on experimental design. Experimental approaches that systematically vary total and relative plant densities provide better quantification of competition in agroecosystems. Factors that influence competition in agroecosystems include the timing of plant emergence, growth rates, spatial arrangements among neighbors, plant-plant-environment interactions, and herbivory. *See* ECOLOGY; ECOLOGY, APPLIED.

Paul C. Marino

Bibliography. E. F. Connor and S. Simberloff, Competition, scientific method, and null models in ecology, *Amer. Sci.*, 74:155–162, 1986; S. R. Reice, Nonequliibrium determinants of biological community diversity, *Amer. Sci.*, 82:424–435, 1994; J. Weins, On competition and variable environments, *Amer. Sci.*, 65:590–597, 1977.

Ecological energetics

The study of the flow of energy within an ecological system from the time the energy enters the living system until it is ultimately degraded to heat and irretrievably lost from the system. It is also referred to as production ecology, because ecologists use the word production to describe the process of energy input and storage in ecosystems.

Ecological energetics provides information on the energetic interdependence of organisms within ecological systems and the efficiency of energy transfer within and between organisms and trophic levels. Nearly all energy enters the biota by green plants' transformation of light energy into chemical energy through photosynthesis; this is referred to as primary production. This accumulation of potential energy is used by plants, and by the animals which eat them, for growth, reproduction, and the work necessary to sustain life. The energy put into growth and reproduction is termed secondary production. As energy passes along the food chain to higher trophic levels (from plants to herbivores to carnivores), the potential energy is used to do work and in the process is degraded to heat. The laws of thermodynamics require the light energy fixed by plants to equal the energy degraded to heat, assuming the system is closed with respect to matter. An energy budget quantifies the energy pools, the directions of energy flow, and the rates of energy transformations within ecological systems. *See* BIOLOGICAL PRODUCTIVITY; FOOD WEB; PHOTOSYNTHESIS.

The peak of studies in ecological energetics occurred in the 1960s and early 1970s largely because a major concern of the International Biological

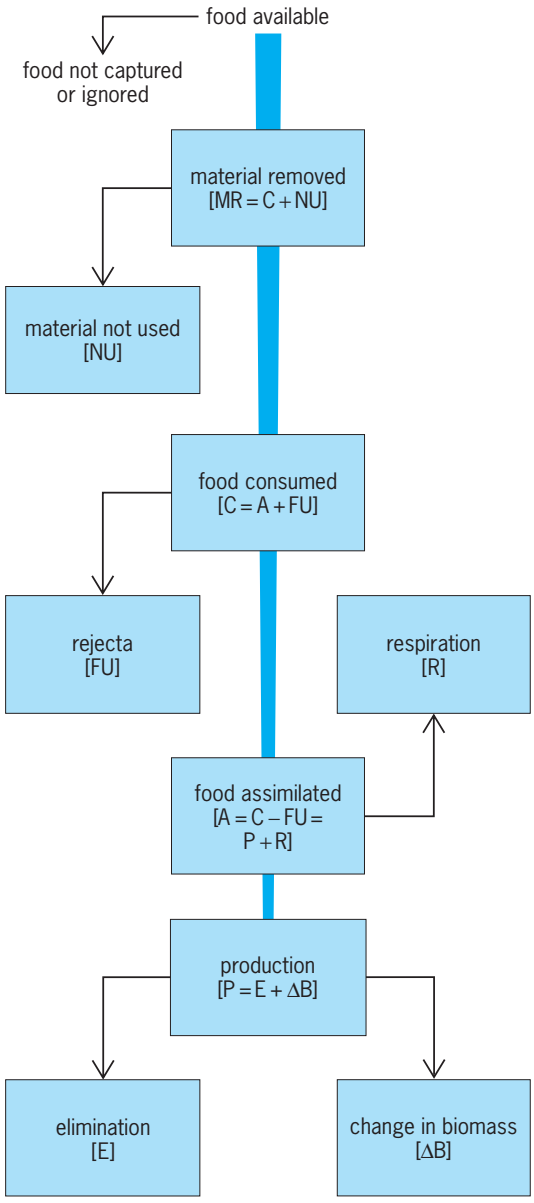


Fig. 1. Diagrammatic representation of energy flow through an ecological unit; abbreviations are explained in the text.

Program was an appraisal of the biological productivity of terrestrial and aquatic communities in relation to human welfare. Initially considered to have the potential of becoming a unifying language in ecology—an ecological Rosetta Stone—the subject has yielded little in the way of general theory.

The essentials of ecological energetics can be most readily appreciated by considering the schema (**Fig. 1**) of energy flowing through an individual; it is equally applicable to populations, communities, and ecosystems. Of the food energy available, only part is harvested (MR) in the process of foraging. Some is wasted (NU), for example, by messy eaters, and the rest consumed (C). Part of the consumed food is transformed but is not utilized by the body, leaving as fecal material (F) or as nitrogenous waste (U), the by-product of protein metabolism. The remaining energy is assimilated (A) into the body, part of which is used to sustain the life functions and to do work—this is manifest as oxygen consumption. The remainder of the assimilated energy is used to produce new tissue, either as growth of the individual or as development of offspring. Hence production is also the potential energy (proteins, fats, and carbohydrates) on which other organisms feed. Production (P) leads to an increase in biomass (B) or is eliminated (E) through death, migration, predation, or the shedding of, for example, hair, skin, and antlers.

Pathways. Energy flows through the consumer food chain (from plants to herbivores to carnivores) or through the detritus food chain. The latter is fueled by the waste products of the consumer food chain, such as feces, shed skin, cadavers, and

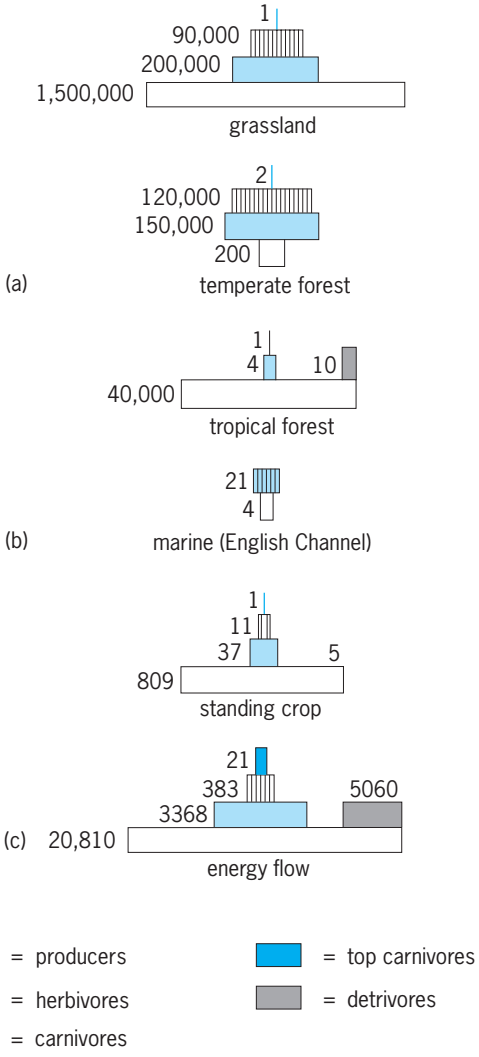


Fig. 2. Trophic levels of a number of ecosystems represented in different units. (*a*) As numbers of individuals per 1000 m^2 of grassland and temperate forest community in summer; microorganisms and soil animals excluded. (*b*) The standing crop or biomass (grams dry weight per meter squared) of terrestrial (Panamanian tropical rainforest) and marine (English Channel) communities; note the inversion of the marine pyramid. (*c*) The aquatic community of Silver Springs, Florida, represented as standing crop (kilocalorie per meter) and energy flow (kilocalories per meter per year). (*After E. P. Odum, Fundamentals of Ecology, 3d ed., W. B. Saunders, 1971*)

nitrogenous waste. Most detritus is consumed by microorganisms such as bacteria and fungi, although this food chain includes conspicuous carrion feeders like beetles and vultures. In terrestrial systems, more than 90% of all primary production may be consumed by detritus feeders. In aquatic systems, where the plants do not require tough supporting tissues, harvesting by herbivores may be efficient with little of the primary production passing to the detrivores.

Pyramids of biomass are used to depict the amount of living material, or its energetic equivalent, present at one time in the different trophic levels (**Fig. 2**). Although the energy flow cannot increase at higher trophic levels, pyramids of biomass may be inverted, especially in aquatic systems. This occurs because the index P/B is inversely related to the size of the organisms. Hence a small biomass may support a high level of production if the biomass is composed of small individuals (**Fig. 3**).

Units. Traditionally the calorie, a unit of heat energy, has been used, but this has been largely replaced by the joule. Confusion is possible, especially in the field of nutrition, because with an initial capital, Calorie may denote kilocalories. Biomass or standing crop is expressed as potential energy per unit area, but the other compartments in Fig. 1, for example *P* and *R*, are expressed in terms of energy flux or rates. The efficiency values such as P/A are dimensionless, but the ratio P/B is a rate—the inverse of the turnover time.

Measurement of energy flow. For illustrative purposes some general methods for assessing biological productivity are described here in the context of energy flow through a population. Production is measured from individual growth rates and the reproductive rate of the population to determine the turnover time. The energy equivalent of food consumed, feces, and production can be determined by measuring the heat evolved on burning a sample in an oxygen bomb calorimeter, or by chemical analysis—determining the amount of carbon or of protein, carbohydrate, and lipid and applying empirically determined caloric equivalents to the values. The latter three contain, respectively, 16.3, 23.7, and 39.2 kilojoules per gram of dry weight. Maintenance costs are usually measured indirectly as respiration (normally the oxygen consumed) in the laboratory and extrapolated to the field conditions. Error is introduced by the fact that animals have different levels of activity in the field and are subject to different temperatures, and so uncertainty has surrounded these extrapolations. Oxygen consumption has been measured in animals living in the wild by using the turnover rates of doubly labeled water (D_2O).

Levels of inquiry. Ecological energetics is concerned with several levels of inquiry: the partitioning of energy between the compartments denoted in Fig. 1; the pathways traced by the energy as it passes through the trophic levels; and the efficiency of energy transfer between trophic levels. The ratio of energy flux through one compartment in Fig. 1 to any previous compartment is referred to as an efficiency. Numerous efficiencies can be calculated both within and between trophic levels. The most common are the assimilation efficiency (A/C), namely the proportion of energy assimilated by the body from the food consumed, and the production efficiency (P/A), which denotes the proportion of energy assimilated which ends up as new tissue. These various efficiencies combine to limit the energy available to the higher trophic levels. The ratio of food consumed or ingested at one trophic level to that ingested by the next lower level is termed ecological efficiency. A value of 10% for this efficiency is often cited; consideration of the A/C and P/A efficiencies

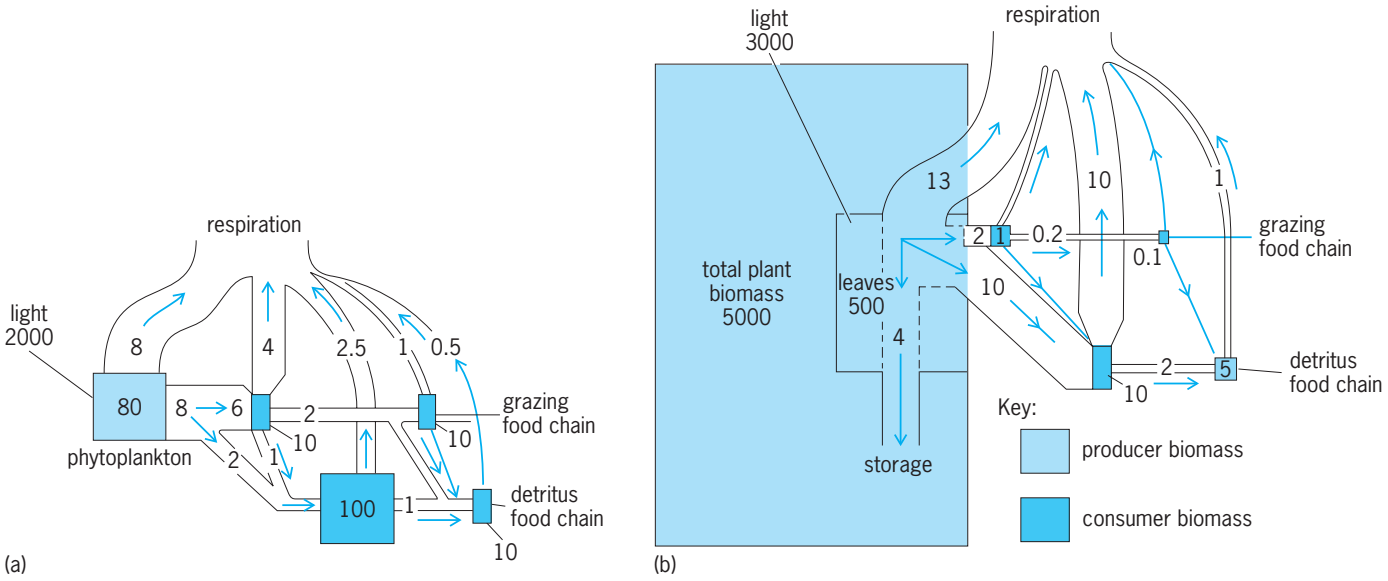


Fig. 3. Models of energy flow through two contrasting ecosystems: (*a*) a marine bay and (*b*) a forest. The boxes represent the biomass in kilocalories per meter and the flow lines show the energy flux (kilocalories per meter per day) along the various pathways. The boxes and the flux lines are scaled to indicate their relative magnitudes. (*After E. P. Odum, Relationships between structure and function in the ecosystem, Jap. J. Ecol., 12:108–118, 1962*)

Production efficiency of populations of various classes of animals living in the wild*

Animal group	Production efficiency (P/A), %
Shrews	0.9
Birds	1.3
Other mammals	3.1
Fish, ants, and termites	9.8
Invertebrates other than insects	25.0
Herbivores	20.8
Carnivores	27.6
Detrivores	36.2
Insects except ants and termites	40.7
Herbivores	38.8
Detrivores	47.0
Carnivores	55.6

*After W. F. Humphreys, Production and respiration in animal populations, *J. Anim. Ecol.*, 48:427–454, 1979.

of most organisms shows that it could seldom exceed 15–20%. However, the effect of heat losses at each trophic level in limiting the length of food chains in nature remains controversial.

Factors affecting efficiency. Respiration rate of organisms is scaled as the three-quarters power of body weight. Hence larger organisms have proportionately slower rates of respiration. This scaling factor seems to affect many rate processes in the body so that size does not influence those efficiencies which are the focus of ecological energetics. However, different types of organisms of the same size have different metabolic rates. For example, warm-blooded animals have much higher weight-specific respiration rates than cold-blooded ones. Analysis of energy budgets derived for wild-living animals shows that a number of taxonomic and trophic groups can be distinguished according to characteristic production efficiencies (see **table**). Production efficiency appears to be related to the general level of metabolic activity—animals with high rates of metabolism generally having low production efficiency.

Due to the loss of usable energy with each transformation, in an area more energy can be diverted into production by plants than by consumer populations. For humans this means that utilizing plants for food directly is energetically much more efficient than converting them to eggs or meat. *See* BIOMASS; ECOLOGICAL COMMUNITIES; ECOSYSTEM.

W. F. Humphreys

Bibliography. S. Brody, *Bioenergetics and Growth*, 1945; R. M. May, Production and respiration in animal communities, *Nature*, 282:443–444, 1979; E. P. Odum, *Fundamentals of Ecology*, 3d ed., 1971; K. Petrusewicz and A. Macfadyen, *Productivity of Terrestrial Animals*: *Principles and Methods*, 1970; J. Phillipson, *Ecological Energetics*, 1966; S. L. Pimm, *Food Webs*, 1982.

Ecological methods

The methods required for the measurement of features of plant and animal populations or for quantifying energy flow or other aspects of ecological communities. Ecologists are concerned with such a wide range of variables that they may use many other techniques, including in particular those for the measurement of weather conditions or for determining features and composition of air, water, and soil. *See* ECOLOGICAL COMMUNITIES.

Patterns of distribution. Ecologists often attempt to describe the way that individual plants or animals are distributed in the habitat. One pattern is the clumping of organisms. This may occur because only part of the habitat is suitable for them (as when rushes occur in the damp patches of a field) or because they represent the offspring of an original colonist and have not moved far from the parent. Another pattern is known as random distribution because each organism is independent of the other (as when thistledown settles on a plowed field). A regular pattern describes the situation when organisms are virtually equidistant from each other. It is often a reflection of competition.

Such patterns are normally determined by sampling the organisms from unit areas—generally squares known as quadrats. For vegetation sampling, long pins resembling knitting needles are dropped into the ground and plants that touch them are recorded. These pins are regarded as infinitely small quadrats and are termed point quadrats. Various statistics, such as the mean ($\overline{x}$) and the variance (s^2), may be calculated from the results of such sampling, and their magnitudes and relationships to each other will show whether the pattern is clumped, random, or regular: if the ratio $s^2/\overline{x} \sim 1$ the individuals are randomly distributed; if the ratio is greater than 1 they are clumped; but if less than 1 they are in a regular pattern. *See* STATISTICS.

Estimation of populations. For plants, individuals (such as trees) may be counted directly, and the main problem is often (as with grasses) deciding what is a plant. However, the populations of only a few animals can be estimated by direct counting; in these cases, the organism is either sessile (like barnacles) or the individuals are very large, and all the individuals in an area can be seen (or at least recorded on a photograph) at the same time. Aerial photography and remote sensing techniques have provided powerful approaches for the estimation of the numbers of large animals (such as cattle or elephants) over wide areas or the extent and type of vegetation cover. However, most animals are active and small, and only a part (sample) of the population can be seen at any one time. Alternatively, it may be tedious and time-consuming to count the animals, so only those in small parts (samples) can be counted (for example, mites per leaf rather than per tree). *See* AERIAL PHOTOGRAPHY; REMOTE SENSING.

One approach to estimating population size is to mark a number of the animals (a), perhaps with a colored dye or paint, and release these into the population. If it can be assumed that they have mixed completely (and the marking has not affected their behavior), the proportion that they constitute of the next sample will be the same as the proportion that the original number marked constitutes of the

total population (N). This is shown in the equation below, where r is the number of marked individuals recaptured, n is the total number in the sample after marking, a is the number originally marked, and N is the population. This calculation also assumes that there are no births or deaths, or any migration in the period between the samples. Often these assumptions are not justified, and many more complex systems, of both sampling and marking (on several occasions) and of calculation, have been devised.

$$\frac{r}{n} = \frac{a}{N} \quad \text{or} \quad N = \frac{an}{r}$$

An alternative approach is to take a number of samples and determine the total number of animals in them. The average number per sample is then multiplied by the total number of sample units in the habitat (for example, leaves on a tree) to obtain the population for that habitat. Special techniques are often needed to determine the total animals in a sample. For example, mites and insects in samples of soil (or litter) are usually extracted by placing the soil in a special apparatus, the Berlese-Tüllgren funnel, where a light bulb or other heat source slowly dries out the soil so that the animals move down and eventually drop through the funnel to a collecting tube. Other soil animals (for example, many types of worms) require moist conditions, and they can be driven out of wet samples by lowering the oxygen level (again by heat) or by chemicals. There is a wide variety of methods for different organisms and for various habitats. Special methods may also be necessary to take the samples: a common method of sampling insects from grassland is the D-vac sampler, basically a large, specially designed vacuum cleaner. *See* MATHEMATICAL ECOLOGY.

Population dynamics. The aim of population estimation is often to contribute toward an understanding of how the population is changing over time, such as to determine trends or the major mortalities. Methods are devised to measure the amount of mortality (for example, the remains of dead animals or plants can be found, or the level of parasitism assessed), the numbers born (natality), and the organisms moving in (immigration) and out (emigration) of the habitat. Special computational techniques are used to estimate the numbers of individuals in a population and to construct life tables. All these methods make certain assumptions, but since these are not always justified, there is great value in making population estimates by a number of different methods. *See* POPULATION ECOLOGY.

Energy flow. The contribution of organisms of different sizes to the functioning of an ecosystem may be best compared in terms of their total weight—the biomass. The biomass in a habitat is termed the standing crop. It is measured in wet weight or dry weight, the latter obtained by freeze drying. The dynamics of an ecosystem, that is, the exchanges between the different components, may be expressed in terms of energy transfers to form a picture of the energy flow. The energy content of a given biomass is often determined by burning a sample in a bomb calorimeter; the heat rise is a measure of the calorific value of the sample. Then it is possible to determine the amount of animal or plant material produced (the production) and, for animals, the amount of food consumed. *See* BIOLOGICAL PRODUCTIVITY; BIOMASS.

Two equations are particularly useful for determining the energy relationships of organisms:

Consumption − (fecal and urinary waste) = assimilation

Assimilation = respiration and production

Consumption and production can usually be measured in terms of biomass and calorific value. Respiration is often measured with special respirometers, and assimilation can sometimes be measured directly by using an inert marker in the food.

Other methods are used to measure the flow of specific nutrients in ecosystems; special computational methods are used to calculate indices to describe the diversity of animal and plant communities; many types of traps or similar devices are used to capture animals; and radioactive isotopes can be used as markers for organisms or materials. Ecologists are always devising new methods for the many and varied problems they encounter. *See* ECOLOGICAL ENERGETICS; ECOLOGY; ECOSYSTEM. T. R. E. Southwood

Bibliography. R. Clarke (ed.), *The Handbook of Ecological Monitoring*, 1986; P. A. Erickson, *Environmental Impact Assessment: Principles and Applications*, 1987; J. A. Ludwig and J. F. Reynolds, *Statistical Ecology: Primer on Methods and Computing*, 1988; R. Southwood and P. A. Henderson, *Ecological Methods*, 3d ed., 2000; H. Walter and S. W. Breckle, *Ecological Systems of the Geobiosphere*, 1985.

Ecological modeling

The use of computer simulations or mathematical equations to address questions that cannot be answered solely by experiments or observations. Ecological models have two major aims: to provide general insight into how ecological systems or ecological interactions work; and to provide specific predictions about the likely futures of particular populations, communities, or ecosystems.

Models can be used to indicate general possibilities or to forecast the most likely outcomes of particular populations or ecosystems. Models differ in whether they are "basic" or are intended to address management decisions. As ecology has grown in its sophistication, models are increasingly used as decision support tools for policy-makers. *See* POPULATION ECOLOGY.

Areas of modeling. The earliest ecological models were produced by mathematicians attempting to explain cycles of predators and prey, disease epidemics, pest outbreaks, and human population

growth. These early models are the foundations of ecology and exemplify the value of mathematics as a tool for formulating biological questions. For example, the Italian mathematician Vito Volterra was asked to explain why the fish that are eaten by sharks decreased in relative abundance when fishing halted in the Mediterranean Sea during World War I. The models developed by Volterra not only accounted for this decline in shark prey with the war-related curtailment of harvest but also indicated that insect pests might actually increase in numbers if chemical insecticides killed both their predators and the insects themselves [a phenomenon observed 40 years later after the invention of dichlorodiphenyltrichloroethane (DDT) and other chemical insecticides]. More generally, Volterra's models predicted a tendency for predators and their prey to fluctuate even if the environment stayed the same from year to year—another prediction that is generally upheld by data.

Models of human population growth stimulated the development of mathematical tools for converting schedules of birth and death into predictions about annual rates of population change. Although these demographic models originated over 50 years ago, modifications of them provide the foundation for much modern conservation biology. For instance, by inputting rates of mortality and reproduction for loggerhead sea turtles into a simple model, it was found that extinction of the turtles could best be prevented if large young adults were protected from shrimp trawlers; legislation for this was passed by the U.S. Congress.

Models of virtually every possible type of ecological interaction have been developed (competition, parasitism, disease, mutualism, plant-herbivore interactions, and so forth). The models vary in their level of detail. Some models simply keep track of the density of organisms, treating all organisms of any species as identical (mass action models). At the other extreme, the movement and fate of each individual organism may be tracked in an elaborate computer simulation (individual behavior models).

Building a model. The best way to understand how ecological modeling works is to actually build a model. One of the most important general models in ecology describes the consequences of habitat loss for species that require specific types of habitat. The type of model described below is called a metapopulation model because it keeps track of many populations in an archipelago of patches. In order to model how such species are likely to be influenced by losses of habitat, ecologists have examined models that range from simple ("pencil-and-paper") exercises that can be expressed with a few equations to intricate computer models that involve hundreds of complex mathematical relationships and lengthy computer programs.

A useful pencil-and-paper model about the consequences of habitat destruction may be devised by carefully expressing ideas in mathematical language. For example, begin by simplifying the world to patches of habitat that are either occupied by the species of interest (a fraction x) or not occupied by the species ($1 - x$). No species will occupy habitat forever, since there will inevitably be local death or even disappearance of an entire population within some patch of forest or alpine meadow. Conversely, empty or unoccupied habitat will not remain vacant forever, as long as there is some source of colonists that can find their way to the open habitat. A simple equation (1) summarizes the change

$$\textbf{change} = \begin{pmatrix}\text{gains in } x\\ \text{due to colonization}\\ \text{of empty patches}\end{pmatrix} - \begin{pmatrix}\text{losses in } x\\ \text{due to disappearance}\\ \text{from occupied patches}\end{pmatrix} \quad (1)$$

in species abundance expected in such a world of habitat patches (occupied or unoccupied) with perpetual turnover; here **change** represents the change in the fraction of occupied patches per unit time.

This expression may be made more precise by hypothesizing that there is a constant loss or disappearance of the species from occupied patches at a rate e; this means the "minus term" of Eq. (1) becomes $-ex$. It is reasonable to assume that vacant patches of habitat ($1 - x$) get colonized in proportion to the fraction of occupied habitat from which colonists might arise (c times x, where c indicates the dispersal or colonization ability of the species, and x is the proportion of habitat occupied as mentioned above). This is a specific model that predicts the fraction of habitat patches that should be occupied by an organism, and how that fraction changes through time [Eqs. (2)–(4)]. The species will increase as long

$$\textbf{change} = cx(1 - x) - ex \quad (2)$$

$$\textbf{change} = cx(1 - x - D) - ex \quad (3)$$

$$(1 - D - e/c) \quad (4)$$

as **change** is greater than zero. **Change** is greater than zero if x is less than ($1 - e/c$); and **change** is less than zero if x is greater than ($1 - e/c$). Hence x will tend to move toward the fraction given by ($1 - e/c$), and it should be clear why ($1 - e/c$) can be called the equilibrium fraction of habitat occupied by the species of interest. As long as colonization rates are sufficiently high (that is, c is greater than e), the species will persist forever. Habitat destruction permanently removes a fraction D of the vacant habitat; thus, instead of ($1 - x$) fraction of habitat being available to be colonized, only ($1 - x - D$) habitat is available to be colonized. With habitat destruction, Eq. (2) is modified to Eq. (3). With habitat loss incorporated through the D term, the equilibrium fraction of habitat occupied by a species is now (4). Thus, if habitat loss D is too large, the species will become extinct because the equilibrium will be less than zero, which means the species will decline until it has totally disappeared. This model makes several important points. For example, even though there is empty habitat (which means $1 - x$ is greater than 0), the species may still relentlessly

decline simply because so much habitat has been destroyed that the colonization of empty habitat occurs too infrequently to sustain the species. This exposes the fallacy of arguments that have been made regarding old growth forests and the endangered spotted owl species in the Pacific Northwest: the notion that there is ample old growth forest for this endangered species because some of the forest is devoid of owls is clearly wrong. The model also shows that there is a threshold for habitat destruction (for example, when habitat loss D becomes greater than $1 - e/c$), which when crossed implies inevitable gradual decline of a species to extinction.

A key feature of ecological modeling is a parameter estimation. The parameters in the model related to Eq. (4) are e, c, and D. If the value of these parameters was known, it would be possible to predict whether or not the species of interest could persist, and if it could persist, what fraction of habitat would be expected to be occupied by that species. Thus, in order to apply the model even crudely to any situation, it is necessary to estimate the extinction rate or disappearance rate (fraction lost per year or week), the colonization rate, and the fraction of habitat that has been lost altogether. This might seem like an easy task, but in most circumstances the factor that most thwarts ecological modelers is estimating the parameters. Because parameters may be difficult to estimate, there is typically a great deal of uncertainty surrounding the application of any model. For example, suppose it is known that the disappearance rate for spotted owls in old growth forest patches is 3% loss per year and the colonization rate is 5%, but it is uncertain whether or not 35% or 45% of the habitat has been destroyed. If $D = .35$ the species could persist (because $1 - .35 - .03/.05$ is greater than 0), but if $D = .45$ the species would go extinct (because $1 - .45 - .03/.05$ is less than 0). Any uncertainty about habitat loss D would yield uncertainty about the fate of the species.

Simple algebraic models such as Eqs. (1)–(4) are very useful for indicating general principles and possibilities. They were part of the testimony considered in court cases regarding the halting of logging in Washington and Oregon so that spotted owls would not cross the threshold for habitat loss, at which point the species would be doomed. But something as simple as Eq. (4) could never be practically applied to the management of any specific forest and spotted owl population. In order to be a management tool, the model must be more complicated and detailed to reflect the specific situation under examination. For example, instead of a few equations, ecologists have modeled spotted owl populations and old growth forests in Washington using a detailed computer simulation that keeps track of habitat in real maps at the scale of hectares. In these simulation models, owls are moved as individuals from one hectare to another, and their fate (survival, death, or reproduction) is recorded in the computer's memory. By tracking hundreds or even thousands of owls moving around in this computer world, different forestry practices corresponding to different logging scenarios can be examined. *See* ECOLOGY, APPLIED; MATHEMATICAL ECOLOGY; SYSTEMS ECOLOGY.

Advantages and limitations. A model is a formal way of examining the consequences of a series of assumptions about how nature works. Such models refine thinking and clarify what results are implied by any set of assumptions. As models become more complicated and specific, they can also be used as surrogate experimental systems; in other words, ecologists may use models to conduct experiments that are too expensive or impractical in the field. For example, to protect against extinction of northern spotted owls in the northwestern United States, investigators do not have the time or the resources to engage in thousand-acre experiments that entail different forestry practices (for example, logging in a way that leaves behind clusters of old growth forest versus logging randomly across the landscape). However, if enough is known about how owls move through forests, a simulation model can be devised that "does the experiment," which may suggest the best logging practice.

One danger of ecological modeling is the uncertainty of the models and the shortage of supporting data. Properly used, models allow exploration of a wide range of uncertainty, pointing out the limits of current knowledge and identifying critical information required prior to management decision making. Increasingly, the public will be asked to contribute to decisions about logging, sprawling urban development, removing dams from rivers, and so forth; critical information regarding these decisions will be provided by ecological models. It is therefore essential that citizens understand how models are formulated and applied. Conclusions based on models should not be quickly dismissed because the model may be the best line of reasoning available; however, it would not be prudent to rely solely on the output of any model. Ecological modeling is a branch of science that requires expertise in ecology, mathematics, statistics, and usually computer programming. Done well, ecological modeling can be tremendously valuable as a source of general scientific principles and as a decision-support tool in arenas of resource management and conservation. *See* ECOLOGY. Peter Kareiva

Bibliography. N. Gotelli, *A Primer of Ecology*, 2d ed., Sinauer Press, Sunderland, MA, 1998; J. Maynard Smith, *Models in Ecology*, Cambridge University Press, Cambridge, 1974; J. Roughgarden, *A Primer of Ecological Theory*, Prentice Hall, Upper Saddle River, NJ, 1998.

Ecological succession

A directional change in an ecological community. Populations of animals and plants are in a dynamic state. Through the continual turnover of individuals, a population may expand or decline depending on the success of its members in survival and reproduction. As a consequence, the species composition of communities typically does not remain static with

time. Apart from the regular fluctuations in species abundance related to seasonal changes, a community may develop progressively with time through a recognizable sequence known as the sere. Pioneer populations are replaced by successive colonists along a more or less predictable path toward a relatively stable community. This process of succession results from interactions between different species, and between species and their environment, which govern the sequence and the rate with which species replace each other. The rate at which succession proceeds depends on the time scale of species' life histories as well as on the effects species may have on each other and on the environment which supports them. *See* ECOLOGICAL COMMUNITIES; POPULATION ECOLOGY.

The course of ecological succession depends on initial environmental conditions. Primary succession occurs on novel areas such as volcanic ash, glacial deposits, or bare rock, areas which have not previously supported a community. In such harsh, unstable environments, pioneer colonizing organisms must have wide ranges of ecological tolerance to survive. In contrast, secondary succession is initiated by disturbance such as fire, which removes a previous community from an area. Pioneer species are here constrained not by the physical environment but by their ability to enter and exploit the vacant area rapidly.

As succession proceeds, many environmental factors may change through the influence of the community. Especially in primary succession, this leads to more stable, less severe environments. At the same time interactions between species of plant tend to intensify competition for basic resources such as water, light, space, and nutrients. Successional change results from the normal complex interactions between organism and environment which lead to changes in overall species composition. Whether succession is promoted by changing environmental factors or competitive interactions, species composition alters in response to availability of niches. Populations occurring in the community at a point in succession are those able to provide propagules (such as seeds) to invade the area, being sufficiently tolerant of current environmental conditions, and able to withstand competition from members of other populations present at the same stage. Species lacking these qualities either become locally extinct or are unable to enter and survive in the community.

Primary succession. In some cases, seres may take hundreds of years to complete, and direct observation at a given site is not possible. Adjacent sites may be identified as successively older stages of the same sere, if it is assumed that conditions were similar when each seral stage was initiated.

Glacier Bay. In the Glacier Bay region of Alaska glaciers have retreated, in phases, some 61 mi (98 km) since 1750, leaving a series of moraines of known ages supporting a range of seral vegetational types. Soil factors become modified by vegetation, enabling new species to become established (**Table 1**). Acidic decomposition products of alder leaves sharply reduce soil pH, whereas virtually no change occurs on bare glacial till or under other vegetation. Pioneer species must tolerate low nitrogen levels, but alder is able to fix atmospheric nitrogen (N) by the presence of microbial symbionts in root nodules and is correlated with a rapid increase in soil nitrogen by way of leaf fall. After invasion by spruce, levels of accumulated nitrogen fall as nitrogen becomes incorporated in forest biomass, and the annual additions from alder are reduced. Soil organic matter increases progressively, and influences the structural development of the soil. The mature forest remains only on well-drained slopes. In areas of poorer drainage, invasion by *Sphagnum* moss leads to replacement of trees by wet acidic bog or muskeg.

Lake Michigan. Another example of primary succession has been recorded on a sequence of dune ridges bordering Lake Michigan, ranging in age to 12,000 years. Succession is initiated on a bare sand surface either following a fall in lake level, as occurred in phases since the last glaciation, or due to wind erosion of an existing dune redepositing sand (**Table 2**). Marram grass impedes the transport of sand across bare dune surfaces and so promotes accretion of sand. At the same time, it grows and branches vigorously, thus maintaining cover on the expanding dune (**Fig. 1**).

TABLE 1. Successional changes in vegetation and soils observed on an aged series of moraines at Glacier Bay, Alaska

Years	Vegetation	Soil environment
0	Initial colonizers: mosses, fireweed, horsetail, *Dryas*, dwarf willows; later, willows form dense scrub	Initially pH 8.0–8.4 due to $CaCO_3$: soil N and organic matter lacking
50	Alder invades, forming dense thickets less than 33 ft (10 m) tall	pH falls to 5.0 in 30–50 years due to acidic products of alder leaf decomposition; marked increase in soil N due to fixation in alder root nodules; soil organic matter accumulates
170	Sitka spruce invades, forming dense forest	Reduction in soil N by incorporation in forest biomass; progressive increase in soil organic matter
250+	Western and mountain hemlock enter (climax forest on well-drained slopes)	
	Sphagnum bog with occasional pines replaces forest in poorly drained areas	Soil becomes waterlogged, deoxygenated, acidified

TABLE 2. Succession on sand dunes of Lake Michigan

Years	Vegetation	Soil environment	Ground invertebrates
0	Marram grass (*Ammophila*) invades by rhizome migration	Initial sand pH 7.6	
6		Sand accumulates, dune builds up	White tiger beetle (*Cicindela lepida*) Sand spider (*Trochosa cinerea*)
20	Marram declines in vigor	Sand surface stabilized	White grasshopper (*Trimerotropis maritima*) Longhorn grasshopper (*Psinidia fenestralis*)
50	Jack pine and white pine become established, with light-demanding understory		Burrowing spider (*Geolycosa pikei*) Digger wasps (*Bembex, Microbembex*)
100	Black oak becomes established, with shade-tolerant understory		Bronze tiger beetle (*Cicindela scutellaris*) Ant (*Lasius niger*) Migratory locust (*Melanoplus*) Sand locusts (*Agoeneotettix, Spharagemon*) Digger wasp (*Sphex*)
150 • •			Ant lion (*Cryptoleon*) Flatbug (*Neuroctenus*) Grasshoppers (six species) Wireworms (*Elateridae*) Snail (*Mesodon thyroides*)
1000 •		Soil N builds up to 0.1%	
10,000		Sand pH drops to 4.0 due to $CaCO_3$; soil matures to a nutrient-poor brown humic sand	

Succession on Lake Michigan dunes demonstrates that local factors may modify the typical pattern (**Fig. 2**). In damp depressions with impeded drainage, a grassland community develops. Sheltered pockets on lee slopes have a moister microclimate and tend to accumulate leaf litter from more exposed parts of the dunes, as well as receiving protection from frequent fires. As a result, a more nutrient-rich soil can develop, and succession proceeds via basswood to more nutrient-demanding oak-hickory and finally maple-beech woodland, typical of normal soils of the region. The black oak association appears to be stable on dune soils, since it is tolerant of low nutrients and water limitation, and tends to maintain these conditions by returning few nutrients in the leaf litter.

Climax community. Early stages of succession tend to be relatively rapid, whereas the rates of species turnover and soil changes become slower as the community matures. Eventually an approximation to the steady state is established with a relatively stable community, the nature of which has aroused considerable debate. Earlier, the so-called climax vegetation was believed to be determined ultimately by regional climate and, given sufficient time, any community in a region would attain this universal condition. This unified concept of succession, the monoclimax hypothesis, implies the ability of organisms progressively to modify their environment until it can support the climatic climax community. Although plants and animals do sometimes ameliorate environmental conditions, evidence suggests overwhelmingly that succession has a variety of stable end points. This hypothesis, known as the polyclimax hypothesis, suggests that the end point of a succession depends on a complex of environmental factors that characterize the site, such as parent material, topography, local climate, and human influences.

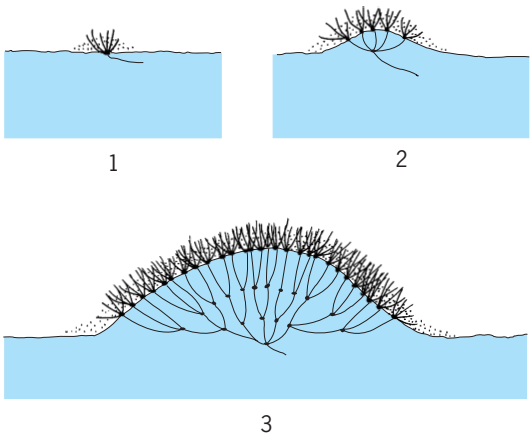


Fig. 1. Dune formation by the gradual deposition of wind-carried sand particles around aerial shoots of *Ammophila arenaria*. 1, 2, and 3 indicate successive periods of time. (*After K. A. Kershaw, Quantitative and Dynamic Ecology, 2d ed., Arnold, 1973*)

In the Lake Michigan sand dunes, the course of succession and its climax appear to be determined by physiographic conditions at the start (Fig. 2). Similarly, the transformation of glacial moraine forest to muskeg depends on local drainage. In the tropical rainforest of Moraballi Creek, Guyana, five apparently stable vegetation types have been distinguished on different soil types under the same climate. A mixed forest is present on red loam, whereas the Wallaba forest occurs on bleached sand, with the Mora forest type in areas liable to flooding.

Autogenic vs. allogenic factors. In the examples of succession discussed above, the chief agent in modifying the environment is the community itself: thus marram stabilizes the sand dune surface, and alder increases the soil nutrient status of moraine soil. These actions of the community on the environment, termed autogenic, provide an important driving force promoting successional change, and

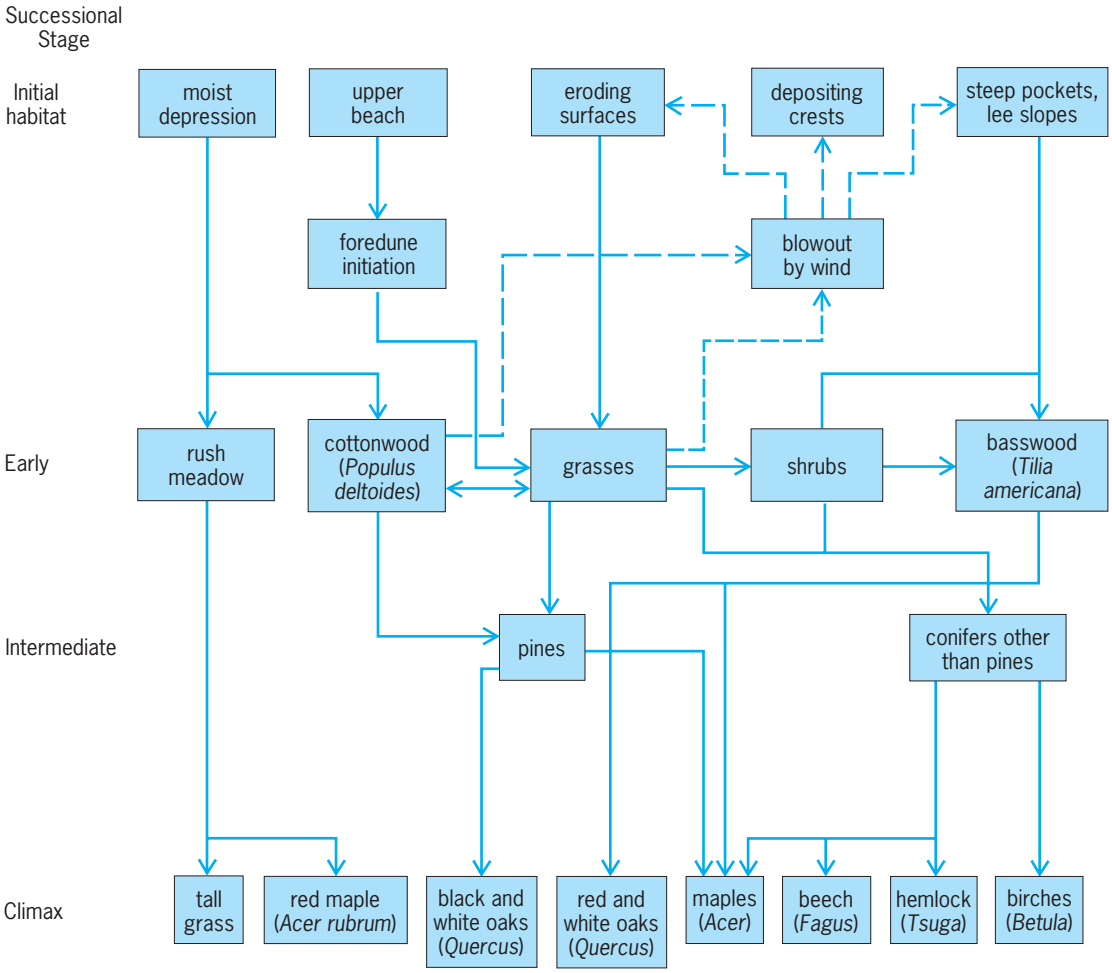


Fig. 2. Alternative patterns of primary plant succession on Lake Michigan dunes, depending on initial conditions, extrinsic variables, and colonization-invasion patterns. (*After S. J. McNaughton and L. L. Wolf, General Ecology, 2d ed., Holt, Reinhart, and Winston, 1979*)

are typical of primary succession where initial environments are inhospitable. Alternatively, changes in species composition of a community may result from influences external to the community called allogenic. For example, in aquatic ecosystems (the hydrosere) the community commonly develops from open water with submerged and floating aquatic plant species toward a swamp community in which rooted emergent plants dominate in shallower water, until finally the marsh is colonized by land plants from the surrounding area as sediment deposition continues and the soil dries out. Reduction in water depth, enabling colonization by marsh species and finally terrestrial species, occurs with input of waterborne and airborne sediments—thus the aquatic phase of the hydrosere is controlled by input of materials from outside the system. Similarly, lakes are typically subject to enrichment of nutrients from surrounding areas, resulting in increased productivity. Extremely high production occurs in culturally eutrophic lakes, which receive nutrient inputs from human activities. In aquatic systems where the influence of allogenic factors such as siltation are apparently minimal, vegetation tends to develop via a series of productive intermediate steps toward an oligotrophic community, that is, one with low productivity, dominated by *Sphagnum* moss (**Fig. 3**). *See* EUTROPHICATION.

Whereas intrinsic factors often result in progressive successional changes, that is, changes leading from simple to more complex communities, external (allogenic) forces may induce retrogressive succession, that is, toward a less mature community. For example, if a grassland is severely overgrazed by cattle, the most palatable species will disappear. As grazing continues, the grass cover is reduced, and in the open areas weeds characteristic of initial stages of succession may become established.

Heterotrophic succession. In the preceding examples of succession, the food web was based on photosynthetic organisms and there was a slow accumulation of organic matter, both living and dead. This is termed autotrophic succession. In some instances, however, addition of organic matter to an ecosystem initiates a succession of decomposer organisms which invade and degrade it. Such a succession is called heterotrophic. In an Illinois pasture, cow dung is degraded by a seral community of some 40 to 60 invertebrate species over a period of 30 days. The newly deposited cow pat is immediately visited by

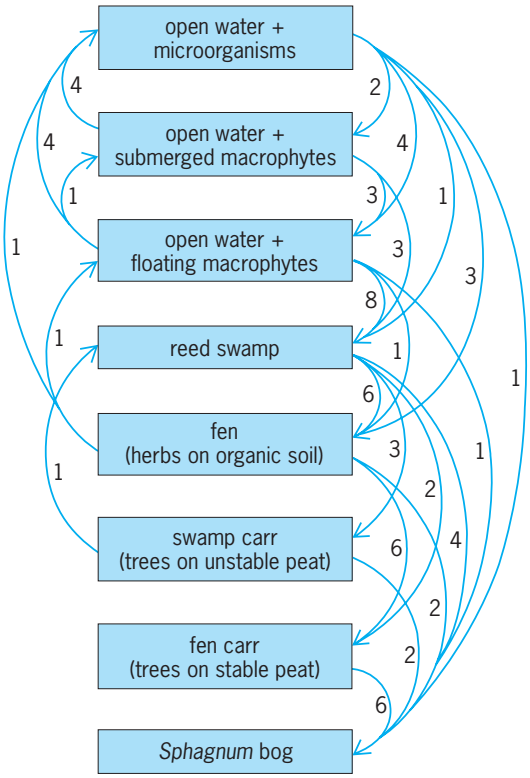


Fig. 3. Transitions between vegetation "stages" free from obvious allogenic influences derived from 20 pollen diagrams for lake sediments and peat. Figures show numbers of observed transitions.

the horn fly (*Haematobia irritans*), which quickly lays eggs and returns to the cow. It is followed by several other dung flies whose larvae are eaten by beetles such as *Sphaeridium scaraboides* which burrow through the dung and lay their eggs. A parasitic wasp (*Xyalophora quinquelinata*) deposits eggs inside maggots of *Sarcophaga* flies. As the dung ages and dries out, it is inhabited by a wider variety of species. Little is known in detail of the importance of the various saprovores in degrading the dung and in modifying the microhabitat, or of their dependence on the activities of saprophytic fungi.

Discharge of organic effluent into a river is detectable downstream by a progression in chemical factors and in the biota. Succession in time is here equivalent to the change in species composition resulting from the decline in organic effluent that corresponds to the distance from the discharge. Marked reduction in dissolved oxygen directly below an outfall results from respiration of microorganisms as they degrade organic matter. Detritivores tolerant of low oxygen concentrations, such as *Tubificidae* and *Chironomidae*, attain high population densities in the bottom sediments. Subsequently, a bloom of algae is typical, utilizing released nitrate and phosphate. As the river flows downstream, the aquatic food web progressively changes from a heterotrophic to an autotrophic basis and productivity declines to its normal level as the "clean water" community returns. *See* BIOLOGICALS; FOOD WEB; PRODUCTIVITY.

Secondary succession. Following the partial or complete destruction of an established community by disturbing events such as fire or clearfelling, and similarly on the cessation of grazing or tillage, a sequence of species invasion and replacement ensues. Compared to the slow initial progress of primary succession in which amelioration of the environment plays an important part, secondary succession is characterized initially by rapid turnover of typically opportunist species which invade relatively congenial habitats.

Piedmont. Abandoned fields in Piedmont area of North Carolina show a rapid sequence of replacement of herbaceous species, apparently related to the life histories of the plants (**Table 3**). Horseweed produces seeds in late summer which germinate immediately so that the plant overwinters as a juvenile, grows rapidly the following year, and dies after seeding in the summer. Aster seeds do not germinate until the following spring, and seedlings grow slowly due to shading by established horseweed plants. In addition, decaying horseweed roots inhibit its own growth and, to a greater extent, that of aster. Horseweed attains dominance in the first year by efficient seed dispersal and rapid establishment. Being a perennial, aster is able to outcompete horseweed in the second year despite the inhibitory effect of the latter. Seedlings of aster are present in abundance in third-year fields but are less drought-resistant than those of broomsedge, which outcompetes aster except in fields with more available water, where aster survives the competition longer. *See* ALLELOPATHY.

Broomsedge seeds are not available to colonize initially because seeds are not produced until the end of the plant's second year and require a period of cold dormancy before germination. Seedling growth is apparently enhanced by decomposition products of the previous colonists. The late establishment of broomsedge in the succession is dependent not on changes brought about by earlier colonists but on the life history of the plant.

TABLE 3. Secondary succession on abandoned fields in the Piedmont area of North Carolina*

Years after last cultivation	Dominant plant	Other common species
0 (autumn)	Crabgrass (*Digitaria sanguinalis*)	
	↓	
1	Horseweed (*Erigeron canadensis*)	Ragweed (*Ambrosia elatior*)
	↓	
2	Aster (*Aster pilosus*)	Ragweed
	↓	
3	Broomsedge (*Andropogon virginicus*)	
	↓	
5–15	Shortleaf pine (*Pinus echinata*)	Loblolly pine
	↓	
50–150	Hardwoods (oaks)	Hickory

*From C. Krebs, *Ecology*, 2d ed., Harper and Row, 1978.

After broomsedge, shortleaf pine invades the herb community. Pine seeds require mineral soil and minimal root competition to become established, and the seedlings are not shade-tolerant. Hence after about 20 years, under a dense pine canopy, reproduction of pines is almost lacking. Accumulation of litter and shade under pines causes the old-field herbs to die out.

Oak seedlings become established after about 20 years, when the depth of litter is adequate to prevent desiccation of acorns. Organic matter in the soil surface layer also increases, improving its water-holding capacity. After about 50 years, several oak species become established and gradually assume dominance as the pines fail to reproduce. Unlike pine, which is capable of germinating on bare soil, oaks and other hardwoods require changes in the soil resulting from pine litter before their seedlings can establish successfully.

While plant species are the main indicators of succession, it is important to note that animal species are also changing over time. In the Piedmont, the changes in bird species as succession proceeds have been well documented. Just a few species, such as meadowlarks and grasshopper sparrows, are found in the initial stages, while more complicated assemblages of species are common in the latter forest stages.

Nova Scotia forest. Following clearfelling in a Nova Scotia forest, the course of secondary succession involves invasion by shrubs (raspberry) followed by understory trees (pincherry, aspen), followed by shade-intolerant species (red maple, paper birch), and finally a shade-tolerant community (hard maple, yellow birch, white ash). Perhaps shade-tolerant species such as red maple, which are of low commercial value, could be inhibited if strip felling were practiced by the forestry industry as an alternative to clearfelling in large blocks. Commercially desirable shade-tolerant species such as white ash would be favored where there was greater local shading of the regenerating community.

Mechanisms of species replacement. Observed changes in the structure and function of seral communities result from natural selection of individuals within their current environment. Three mechanisms by which species may replace each other have been proposed; the relative importance of each apparently depends on the nature of the sere and stage of development.

1. The facilitation hypothesis states that invasion of later species depends on conditions created by earlier colonists. Earlier species modify the environment so as to increase the competitive ability of species which are then able to displace them. Succession thus proceeds because of the effects of species on their environment.

2. The tolerance hypothesis suggests that later successional species tolerate lower levels of resources than earlier occupants and can invade and replace them by reducing resource levels below those tolerated by earlier occupants. Succession proceeds despite the resistance of earlier colonists.

3. The inhibition hypothesis is that all species resist invasion of competitors and are displaced only by death or by damage from factors other than competition. Succession proceeds toward dominance by longer-lived species.

None of these models of succession is solely applicable in all instances; indeed most examples of succession appear to show elements of all three replacement mechanisms. In secondary succession on North Carolina croplands, stimulation of broomsedge growth by decomposition products of previous colonists, and the requirement of oak seedlings for a deep litter layer in which to germinate, exemplify facilitation. The ability of broomsedge to displace aster in competition for water suggest the tolerance mechanism, whereas the inhibition hypothesis is supported by the greater tolerance of horseweed seedlings than aster seedlings to horseweed decomposition products.

Deterministic vs. stochastic succession. Succession has traditionally been regarded as following an orderly progression of changes toward a predictable end point, the climax community, in equilibrium with the prevailing environment. This essentially deterministic view implies that succession will always follow the same course from a given starting point and will pass through a recognizable series of intermediate states (such as in Fig. 2). In contrast, a more recent view of succession is based on adaptations of independent species. It is argued that succession is disorderly and unpredictable, resulting from probabilistic processes such as invasion of propagules and survival of individuals which make up the community. Such a stochastic view reflects the inherent variability observed in nature and the uncertainty of environmental conditions. In particular, it allows for succession to take alternative pathways and end points dependent on the chance outcome of interactions among species and between species and their environment.

Consideration of community properties such as energy flow supports the view of succession as an orderly process. Early in autotrophic succession gross primary productivity (P_g) increases rapidly with community biomass (B), whereas community respiration (R) increases more slowly (**Fig. 4**). As a result, net primary productivity (P_n, where $P_n = P_g - R$) builds up early in succession, and the ratio P_g/B is at its

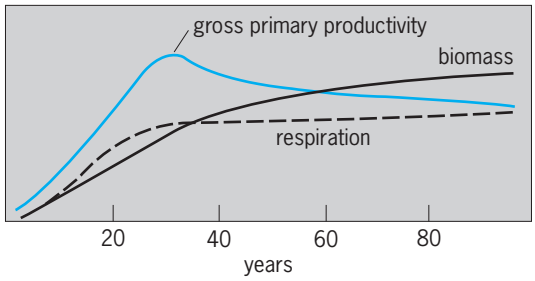


Fig. 4. The energetics of ecosystem development in a forest. The difference between gross primary productivity and respiration is the net primary productivity. (*After E. P. Odum, Ecology, 2d ed., Holt, Reinhart, and Winston, 1975*)

highest in the initial stages. As the community increases in biomass and complexity over time, more complete overall utilization of basic resources such as light limits further increase in primary productivity, whereas R continues to increase because of the increase in tissue to support. Hence zero and the biomass of a mature forest community no longer accumulates. The rate of gross primary productivity typically becomes limited also by the availability of nutrients, now incorporated within the community biomass, and declines to a level sustainable by release from decomposer organisms. Species diversity tends to rise rapidly at first as successive invasions occur, but declines again with the elimination of the pioneer species by the climax community.

Trends in community function, niche specialization, and life history strategy are summarized in **Table 4**. As the community acquires increasing maturity, P_n declines to zero, nutrients become incorporated in biotic pools, broad-niched species are replaced by those with more specific requirements, and the structural organization of the community increases.

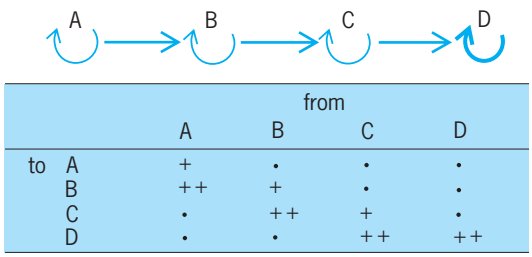
Regeneration of an area of subtropical rainforest in Queensland was observed after the vegetation and surface litter were removed with a bulldozer. Because of small environmental differences in the 10 quadrats observed, succession took four directions after the demise of the first ubiquitous colonizing species, resulting in four apparently stable plant associations. This divergence may result merely from small-scale variation in topography within the 65×130 ft (20×40 m) experimental site and from differing efficiencies of removal of surface litter between the quadrats. Hence the different plant associations detected could be interpreted as divergent products of succession or as phases within a larger-scale vegetation unit.

Stochastic aspects of succession can be represented in the form of models which allow for transitions between a series of different "states." Such models, termed Markovian models, can apply

TABLE 4. Proposed successional trends in ecosystem structural and functional organization, species characteristics, evolutionary factors, and homeostasis*

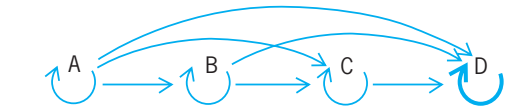
Ecosystem property	Ecosystem stage: Successional (immature)	Ecosystem stage: Climax (mature)
	Energy Flow	
Gross productivity/respiration (P_g/R)	Autotrophic >1 Heterotrophic <1	
Biomass supported/unit energy flow (B/P_g)	Low	High
Net productivity (P_n)	High	Low
Type of food chains	Linear, grazing	Webs, detritus
	Nutrient Flow	
Mineral cycles	Open	Closed
Flow rate: organism-environment	Rapid	Slow
Role of detritus	Unimportant	Important
	Community Structure	
Total organic matter	Little	Much
Location of chemicals	Habitat pools	Biotic pools
Species richness	Low	High
Species evenness	Low	High
Biochemical diversity	Low	High
Spatial heterogeneity	Low	High
	Species Life History Characteristics	
Niche breadth	Broad	Narrow
Organism size	Small	Large
Life cycles	Short, simple	Long, complex
	Selection Pressure	
Growth form	Rapid growth (r-selection)	Feedback control (K selection)
Production	Quantity	Quality
	Overall Homestasis	
Internal symbiosis	Undeveloped	Developed
Nutrient conservation	Poor	Good
Resistance to perturbations	Poor	Good
Entropy	High	Low
Information	Low	High

*After Odum, 1976.



		from A	B	C	D
to	A	+	•	•	•
	B	++	+	•	•
	C	•	++	+	•
	D	•	•	++	++

(a)



		from A	B	C	D
to	A	+	•	•	•
	B	+	+	•	•
	C	+	+	+	•
	D	+	+	+	++

(b)

		from A	B	C	D
to	A	+	•	•	•
	B	+	++	•	•
	C	+	•	+	•
	D	+	•	+	++

(c)

Fig. 5. Schematic representation of three postulated mechanisms for species replacement in succession: (*a*) facilitation, (*b*) tolerance, and (*c*) inhibition. Arrows indicate the direction in which the systems tend to move between different states, A–D. Relative probabilities of movement are indicated by thickness of arrows and by symbols in the accompanying transition matrix, where • = close to zero, + = moderate, and ++ = high.

at various levels: plant-by-plant replacement, changes in tree size categories, or transitions between whole communities. A matrix of replacement probabilities defines the direction, pathway, and likelihood of change, and the model can be used to predict the future composition of the community from its initial state. With alternative transition matrices, this simple model could represent a linear progression toward a stable end state, or a cyclical, recursive sequence of communities. The three postulated mechanisms for species replacement discussed above can be illustrated by the topology of alternative Markovian models (**Fig. 5**). The facilitation model of succession is represented as a linear sequence with greater probabilities of progression toward the final state (D) than maintenance of intermediate states. In the tolerance model, later stages may develop from earlier stages depending on the availability of propagules of subsequent stages and their competitive ability. Again, only state D has a high probability of self-replacement. In the inhibition model, there is a high probability that an intermediate state (in this case, B) will persist by strong self-replacement, thereby truncating the normal succession toward state D. A very low probability that B will change to C or D is assumed. Hence, a high degree of realism can be achieved with a simple model system, and alternative predictions of successional changes can be compared with observed data. *See* ECOLOGY; STOCHASTIC PROCESS.

Peter Randerson

Bibliography. D. H. Boucher (ed.), *The Biology of Mutualism: Ecology and Evolution,* 1985, reprint 1989; R. E. Ricklefs and G. L. Miller, *Ecology*, 4th ed., 1999; P. Yodis, *Introduction to Theoretical Ecology*, 1988.

Ecology

The subdiscipline of biology that concentrates on the relationships between organisms and their environments; it is also called environmental biology. Ecology is concerned with patterns of distribution (where organisms occur) and with patterns of abundance (how many organisms occur) in space and time. It seeks to explain the factors that determine the range of environments that organisms occupy and that determine how abundant organisms are within those ranges. It also emphasizes functional interactions between co-occurring organisms. In addition to its character as a unique component of the biological sciences, ecology is both a synthetic and an integrative science since it often draws upon information and concepts in other sciences, ranging from physiology to meteorology, to explain the complex organization of nature.

Environment is all of those factors external to an organism that affect its survival, growth, development, and reproduction. It can be subdivided into physical, or abiotic, factors, and biological, or biotic, factors. The physical components of the environment include all nonbiological constituents, such as temperature, wind, inorganic chemicals, and radiation. The biological components of the environment include the organisms. A somewhat more general term is habitat, which refers in a general way to where an organism occurs and the environmental factors present there. *See* ENVIRONMENT.

A recognition of the unitary coupling of an organism and its environment is fundamental to ecology; in fact, the definitions of organism and environment are not separate. Environment is organism-centered since the environmental properties of a habitat are determined by the requirements of the organisms that occupy that habitat. For example, the amount of inorganic nitrogen dissolved in lake water is of little immediate significance to zooplankton in the lake because they are incapable of utilizing inorganic nitrogen directly. However, because phytoplankton are capable of utilizing inorganic nitrogen directly, it is a component of their environment. Any effect of inorganic nitrogen upon the zooplankton,

then, will occur indirectly through its effect on the abundance of the phytoplankton that the zooplankton feed upon. *See* PHYTOPLANKTON; ZOOPLANKTON.

Just as the environment affects the organism, so the organism affects its environment. Growth of phytoplankton may be nitrogen-limited if the number of individuals has become so great that there is no more nitrogen available in the environment. Zooplankton, not limited by inorganic nitrogen themselves, can promote the growth of additional phytoplankton by consuming some individuals, digesting them, and returning part of the nitrogen to the environment.

Ecology is concerned with the processes involved in the interactions between organisms and their environments, with the mechanisms responsible for those processes, and with the origin, through evolution, of those mechanisms. It is distinguished from such closely related biological subdisciplines as physiology and morphology because it is not intrinsically concerned with the operation of a physiological process or the function of a structure, but with how a process or structure interacts with the environment to influence survival, growth, development, and reproduction.

Scope. There are a wide variety of approaches to ecology because of its broad, comprehensive character. Ecological studies can be characterized by the type of organisms studied, the habitat where studies take place, the level of organization that is of interest, and the methodology used. These are nonexclusive categories, and ecologists combine them in various ways while doing research. Major subdivisions by organism include plant ecology, animal ecology, and microbial ecology. Subdivisions by habitat include terrestrial ecology, the study of organisms on land; limnology, the study of freshwater organisms and habitats; and oceanography, the study of marine organisms and habitats.

The levels of organization studied range from the individual organism to the whole complex of organisms in a large area. Autecology is the study of individuals, population ecology is the study of groups of individuals of a single species or a limited number of species, synecology is the study of communities of several populations, and ecosystem, or simply systems, ecology is the study of communities of organisms and their environments in a specific time and place.

Higher levels of organization include biomes and the biosphere. Biomes are collections of ecosystems with similar organisms and environments and, therefore, similar ecological properties. All of Earth's coniferous forests are elements in the coniferous forest biome. Although united by similar dynamic relationships and structural properties, the biome itself is more abstract than a specific ecosystem. The biosphere is the most inclusive category possible, including all regions of Earth inhabited by living things. It extends from the lower reaches of the atmosphere to the depths of the oceans. *See* BIOME; BIOSPHERE.

The principal methodological approaches to ecology are descriptive, experimental, and theoretical. Much of ecology through the first half of the twentieth century was descriptive, concentrating on describing the variety of populations, communities, and habitats throughout Earth. Experimental ecology, which involves manipulating organisms or their environments to discover the underlying mechanisms governing distribution and abundance, has become of increasing importance since the mid-1960s. Theoretical ecology uses mathematical equations based on assumptions about the properties of organisms and environments to make predictions about patterns of distribution and abundance. It also has become of increasing importance. All of these approaches, however, are evident in the origins of ecology. The science was more descriptive in its early phases and has become more experimental and theoretical in recent decades.

Ecosystem. An ecosystem is the organisms and physical factors in a specific location that are interrelated through the flow of energy and chemicals to form a characteristic trophic structure (**Fig. 1**). The concept of the ecosystem is fundamental to ecology and may be applied at various levels of organization, although it commonly encompasses several different species. The trophic structure of an ecosystem characterizes organisms according to their feeding level and how those feeding relationships of species result in specific patterns of energy flow and chemical cycling. The living mass of a given population or trophic level at any given time is called biomass. A change in mass with time is referred to as net productivity. *See* BIOLOGICAL PRODUCTIVITY; BIOMASS.

Primary producers, largely green photosynthetic plants, utilize the energy of the Sun and inorganic molecules from the environment to synthesize organic molecules. Those organic molecules serve as food for higher trophic levels. Primary consumers, or herbivores, feed on the producers. Secondary consumers, or carnivores, feed on other consumers, while omnivores feed at several trophic levels. Decomposers feed on the dead tissues of other organisms.

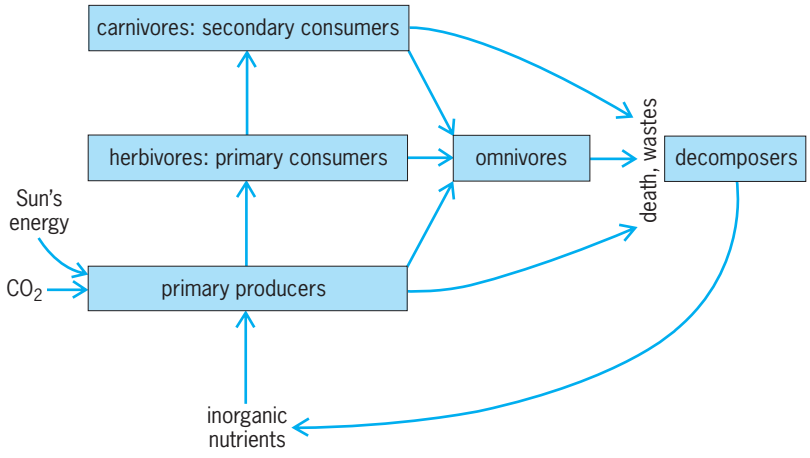


Fig. 1. Generalized trophic of ecosystems. Arrows indicate flow of energy and chemicals.

Each single pathway of energy and chemical flow is referred to as a food chain; the entire collection of food chains in an ecosystem is a food web, or trophic web. Since energy is expended to do work and chemicals are released back into the environment at each step of the trophic web, energy flow in ecosystems is unidirectional, constantly diminishing up the trophic web, while chemicals are recycled to the environment where they can be reutilized in organic syntheses. *See* FOOD WEB.

The ultimate energy source of all ecosystems is the radiant energy of photons from the Sun. Two major types of ecosystems can be distinguished, however, by their proximate energy sources. Autotrophic ecosystems have primary producers as a principal component and sunlight as the major initial energy source. Heterotrophic ecosystems depend upon preformed organic matter that is imported from autotrophic ecosystems elsewhere. The organisms in a stream, for example, may be highly dependent upon organic matter which has eroded from surrounding terrestrial ecosystems.

The ecosystem concept is flexible in application, but the way that it is applied must be clearly specified. An ecosystem study might concentrate on all the organisms and their environments in a specified forest. Alternatively, it might confine itself to the forest floor and soil, concentrating on trophic relationships within that more restricted habitat. *See* ECOSYSTEM.

Autecology

Autecology is the study of particular organisms, typically directed toward determining the traits they possess, and the traits they lack, that allow them to occupy, or be excluded from, certain environments. Early in the twentieth century, autecology placed considerable emphasis upon the morphology and anatomy of organisms, but the field has become diversified in the last few decades into a variety of different approaches. Autecology now also includes physiological ecology, which emphasizes the role of physiological processes, and chemical ecology, which emphasizes the role of biochemical traits. *See* PHYSIOLOGICAL ECOLOGY (PLANT).

One of the earliest generalizations of ecology was the principle of limiting factors, which states that organisms are limited by the factor or combination of factors that are farthest from the requirements of the organism. It was first stated by Liebig based on his studies of soil fertility. He recognized that there usually was a single chemical that would promote yield when it was added to the soil. When that initial limiting factor had been enriched, he found that addition of another nutrient might promote yield even more. The British physiologist Blackman restated the principle in its modern form based on his studies of photosynthesis revealing that the process sometimes could be limited simultaneously by more than one factor.

Though autecology is concerned with individual organisms, that information is often used to contrast the traits of organisms that occupy different environments. Thus there is a strong comparative character to autecological studies; two major habitat comparisons, discussed below, are marine vs. freshwater, and terrestrial vs. aquatic.

Marine vs. freshwater habitats. Life first evolved in the sea; the colonization of freshwater habitats that followed was accompanied by the evolution of traits associated with exploiting the more dilute solutions of freshwater habitats. Seawater is over 3.6% dissolved solids, largely sodium chloride, while freshwater contains 0.01–0.05% dissolved solids. Colonization of freshwater was coupled with the evolution of relatively impervious body surfaces, mechanisms allowing organisms to accumulate ions against a concentration gradient, and the production of large quantities of very dilute urine or other excretory waste. For example, many marine invertebrates are isotonic (equal in osmotic pressure) to seawater so they have no tendency to either lose or gain water. Elasmobranchs have high concentrations of urea in the blood, and although the concentrations of ions is lower than is seawater, the concentrations of urea plus ions is isotonic. Marine teleosts have blood osmotic concentrations well below that of seawater, and they therefore tend to lose water to the sea. They counteract this loss by drinking large quantities of seawater, absorbing it through the gut, and secreting the salt from special cells in the gills. Most freshwater animals are hypotonic (lower osmotic concentration) to the solution that bathes them so they tend to gain water. *See* OSMOREGULATORY MECHANISMS.

Terrestrial vs. aquatic habitats. The colonization of terrestrial habitats was accompanied by the evolution of a large number of traits that allowed organisms to occupy habitats where water was a comparatively rare component of the environment and the atmosphere and soils became principal media surrounding organisms.

Since the body was no longer bathed in fluid, plants and animals both evolved traits allowing them to conserve water and obtain it when it was scarce. Plants developed extensive absorptive surfaces which allowed them to grow into moist regions of the soil and obtain both water and nutrients from the soil solution. Animals often drank water directly and developed outer tissue layers that were extremely impervious to water. The evolution of vascular systems in both animals and plants allowed them to transport water from limited sites of acquisition to the rest of the body. Since the atmosphere is a much less dense and buoyant medium than water, terrestrial plants and animals also evolved traits providing rigid support, such as the skeletons of vertebrates and the strengthened woody stems of vascular plants.

Terrestrial habitats are also characterized by much greater temperature fluctuations than occur in aquatic habitats. Animals therefore evolved physiological traits that allowed them to maintain their body temperatures constant, as well as behavioral traits that involved the ability to move into habitats where favorable temperatures occurred. Physiologically,

evolution culminated in homeothermy in birds and mammals; these organisms expend energy to maintain body temperature within narrow limits. Other animals, called heterotherms, which are incapable of maintaining their body temperature over a wide range of environmental temperatures, often move into localized environmental areas providing appropriate temperatures, and spend unfavorable seasons in quiescent metabolic states. *See* THERMOREGULATION.

Finally, the colonization of terrestrial habitats was associated with many changes in reproductive methods due to the inability of organisms to effectively disperse gametes through the surrounding medium. Such adaptations include the behavioral mating systems of birds and mammals that allow mates to find one another, and the elaborate pollination mechanisms of flowering plants that may involve specific transfer of pollen between individuals of the same species by animal vectors. *See* POLLINATION; REPRODUCTIVE BEHAVIOR.

Physiological and chemical ecology. The postwar development of analytical instrumentation that was much more sensitive to many chemicals than previous analytical techniques led directly to the development of physiological and chemical ecology, which applies that instrumentation to studying organisms in their natural environments. The development of infrared gas analyzers that detect carbon dioxide and water vapor at concentrations that occur in nature led to many studies of photosynthesis, respiration, and transpiration by plants. Those studies have revealed two different types of photosynthesis that occur in plants occupying different types of environments. C_3 plants, so named because the initial stable product of carbon fixation is a three-carbon molecule, are characteristic of almost all aquatic habitats and of terrestrial environments with lower temperatures, less arid climates, and lower solar radiation intensities. C_4 plants, so named because the initial stable product of photosynthesis is a four-carbon molecule, are more common in arid, hotter environments with more intense solar radiation. Those differences are due to the fact that C_3 photosynthesis is intrinsically more efficient but is also more temperature-sensitive and generally more likely to be light-saturated well below the intensity of full sunlight. Thus, C_4 plants are common in tropical and subtropical, arid terrestrial ecosystems, while C_3 plants are more common in most other ecosystems. *See* PHOTOSYNTHESIS.

Chemical ecology studies have been largely devoted to studying rare molecules that have major biological effects. Many animals, for example, produce pheromones that are extremely dilute in the environment but can attract mates, repel predators, or kill pathogens. Many plants also produce organic molecules that are present in their tissues in very small concentrations but act to deter herbivores and pathogens. The evolution of highly toxic molecules by plants sometimes has led to the evolution of specific mechanisms of avoidance or detoxification by a limited number of animals. That type of coupled evolution in which populations change genetically in response to the properties of each other is called coevolution. For example, some animals, such as the monarch butterfly whose larvae feed on milkweeds, have even appropriated aspects of their food plant's chemistry to protect them from predators. Milkweeds produce cardiac glycosides, chemicals that poison the heart muscles of vertebrates. The larvae of the monarch sequester those chemicals in their body, which protects them from such potential predators as birds and small mammals. *See* CHEMICAL ECOLOGY.

Population Ecology

Population ecology is the study of the vital statistics of populations, and the interactions within and between populations that influence survival and reproduction. Population ecologists are concerned with the balances between births and deaths that determine the rate of change of population size through time. When births exceed deaths, of course, population size increases, and when deaths exceed births, population size declines. Vital statistics are often expressed on a per capita basis or, in human populations, on a per-10,000-persons basis. Density, the number of organisms per unit area, is a fundamental property of populations.

One of the principal concerns of population ecologists is identifying and determining the importance of factors that control population densities. Since A. van Leeuwenhoek first calculated the massive potential of flies to reproduce—that is, one pair of flies could produce 746,496 flies after only 3 months of breeding—population ecology has been concerned with the factors that limit the complete realization of this reproductive potential. There are two fundamentally different types of factors limiting population density. Density-independent factors are environmental factors that reduce reproduction or increase death rate independently of the number of organisms in the population. Weather is believed to be a major density-independent factor that often affects populations in a catastrophic way to reduce density. Severely cold winters, deep snowfalls, exceptionally wet growing seasons, and extreme drought occur with a frequency and intensity that are independent of population size. Although these factors can determine population density, they cannot regulate that density since regulation involves maintaining density within certain limits. Density-dependent factors influence the survival or reproduction of individuals in a way that is proportional to density. They can, therefore, regulate population density. Food supply, predators, disease, and such behavioral interactions as territoriality can limit survival and reproduction with increasing intensity as population density increases.

Two vital statistics of populations of principal interest to population ecology are survivorship and reproductive schedules (**Fig. 2**): It is the balance between the probability that an individual will live to reach a certain age, and the average number of

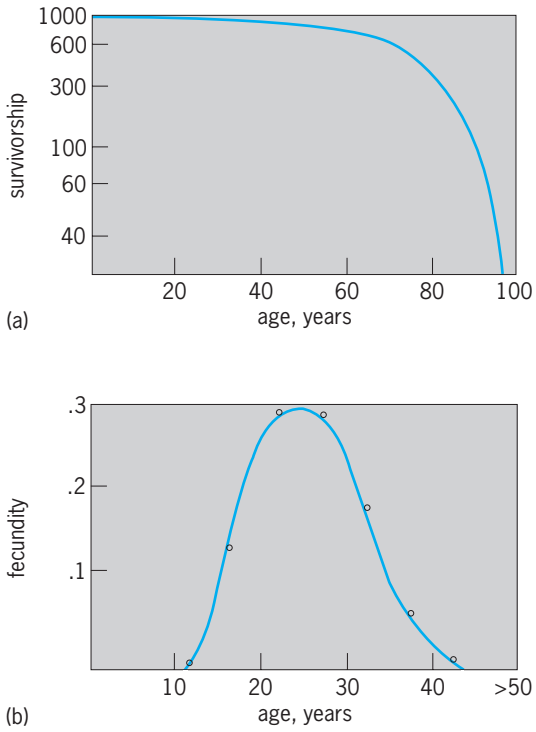


Fig. 2. Demographic statistics of females in the United States. (*a*) Survivorship, where the vertical axis is the number surviving, of 1000 born alive (on a logrithmic scale). (*b*) Reproductive schedule, where the vertical axis is the number of female offspring per female of given age, at 5-year intervals.

offspring produced by individuals of that age that determine population growth rate.

Population growth. One of the earliest quantitative expressions in ecology was a description of density-dependent population growth called the logistic equation. It was first described by P. F. Verhulst in 1838, but it was completely forgotten until its rediscovery by R. Pearl and L. Reed in 1920. Population growth rate is a consequence of population size (N) and per capita birth (b) and death (d) rates, as shown by Eq. (1), where dN/dt is rate of change

$$\frac{dN}{dt} = (b - d)N \qquad (1)$$

of population size. If the symbol r is used to represent the realized per capita reproductive rate, that is, $b - d$, then Eq. (1) can be written in the form of Eq. (2). If r is a constant, a population described

$$\frac{dN}{dt} = rN \qquad (2)$$

by Eq. (2) would grow exponentially with time. The logistic equation recognizes two factors lacking in Eq. (2) that can influence population growth: the maximum per capita reproductive potential of the population, r_m, and the carrying capacity of the environment, K. Maximum reproductive potential would only be realized in the most favorable environment when population density was low enough that individuals did not make simultaneous demands on the same resources. Carrying capacity is a measure of both the amount of resources in the environment and the efficiency with which organisms use those resources. A population is at its carrying capacity when no additional individuals can be supported by an environment. Incorporating those constants into the growth equation above gives Eq. (3), the logistic

$$\frac{dN}{dt} = r_m N\left(1 - \frac{N}{K}\right) \qquad (3)$$

equation. Maximum per capita reproductive potential is determined by the genetic ability to reproduce in an unlimited environment. A population growing according to this equation has a sigmoid, or S-shaped, pattern of population size through time (**Fig. 3**). The maximum population growth rate, which, again, is the product of population size and per capita reproductive rate, is reached when the population is halfway to K.

Dividing Eq. (3) through by N and rearranging gives Eq. (4). Plotting r against N gives a straight line

$$\frac{dN}{Ndt} = r_m - \left(\frac{r_m}{K}\right)N \qquad (4)$$

with an intercept on the y axis that is equal to r_m and an intercept on the x axis that is equal to K.

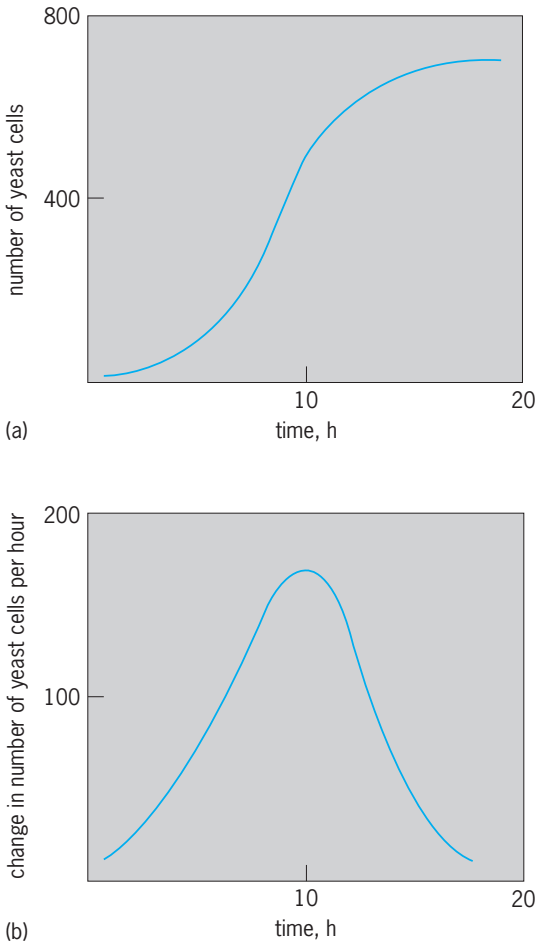


Fig. 3. A yeast population growing according to the logistic equation. (*a*) Population size. (*b*) Growth rate.

An equation as simple as the logistic equation provides merely a yardstick against which the growth of real populations can be compared since it contains a number of assumptions that will not be met in nature. Principal among those assumptions in the logistic equation are that carrying capacity is a constant; that each individual in the population is equivalent enough to all other individuals to be represented in the average population statistic, r_m; that all individuals have an equal probability of mating; and that there are no time lags so that birth and death rates are adjusted immediately for each increment of N as population size increases to K. Much of the study of the ecology of real populations involves understanding how those assumptions are violated.

Competition is inherent in the logistic equation, and in population growth in nature. Competition occurs when the resources in the environment are in short supply relative to the demands that individuals make on those resources. In the logistic equation, that competition is expressed in the decline of r as N increases. This type of competition between members of the same species population is called intraspecific competition.

Population interactions. Competition also is a principal type of interaction between the members of different species, when it is termed interspecific competition. Another important type of interaction between populations is predation, when the individuals of one population feed on the individuals of another population. Both of these interactions were described mathematically in the 1920s by two ecologists working independently, A. Lotka and V. Volterra. The Lotka-Volterra competition equations are Eq. (5) and the exact equivalent, Eq. (6), where 1 and 2 refer

$$\frac{dN_1}{dt} = r_{m1}N_1\frac{K_1 - N_1 - \alpha_{12}N_2}{K_1} \tag{5}$$

$$\frac{dN_2}{dt} = r_{m2}N_2\frac{K_2 - N_2 - \alpha_{21}N_1}{K_2} \tag{6}$$

to the different species, with α_{12} being the competitive effect of an individual of species 2 on species 1, and α_{21} being the competitive effect of an individual of species 1 on species 2. These Lotka-Volterra competition equations recognize that each species depresses the growth rate and carrying capacity of the other. Depending upon the respective competition coefficients, carrying capacities, and maximum per capita reproductive rates, there are four potential outcomes of competition: (1) species 1 may go extinct; (2) species 2 may go extinct; (3) the species may have an unstable equilibrium in which they coexist for a while until chance events allow one species to increase a bit above the equilibrium and the other species then goes extinct; or (4) there may be a stable equilibrium in which both species coexist at some population size below the carrying capacity for either species alone.

The Lotka-Volterra predation equations describe population ecology when members of a predator population, P, feed on members of a prey population, H. The prey equation is Eq. (7), where k_1 is

$$\frac{dH}{dt} = r_mH\left(1 - \frac{H}{K}\right) - k_1PH \tag{7}$$

a constant that is the predation rate per capita of predator. The equation for the predator is Eq. (8),

$$\frac{dP}{dt} = k_2PH - k_3P \tag{8}$$

where k_2 is the efficiency with which predators convert their food into more offspring and k_3 is the per capita predator death rate. Unlike the logistic and Lotka-Volterra competition equations, which predict that populations will reach some final constant size, the Lotka-Volterra predator-prey equations predict that predator and prey populations will tend to oscillate through time. At low densities of both predator and prey, the values of H/K and k_1PH in the prey population equation will both approach zero, and the prey will grow according to the unmodified logistic, approaching the maximum per capita reproductive rate. As the prey population increases, the predator population also can increase, and there is a tendency for predator density to overshoot prey density, leading to oscillations through time. Such oscillations are, in fact, often observed in nature, with a classic example being Canadian populations of snowshoe hare, the prey, and predators, such as lynx (**Fig. 4**). Whether such oscillations in real populations are due to the instability of predator-prey relations inherent in the Lotka-Volterra predation equations, or to the fluctuations of other factors such as weather or food supply available to the hares, has not yet been proven. *See* MATHEMATICAL ECOLOGY; POPULATION ECOLOGY.

Parasitism and disease are additional types of population interactions involving detrimental effects on one participant and beneficial effects on the other participant. A parasite is an organism that grows, feeds on, and is sheltered by another organism, the host. Disease is an abnormal condition of an organism, often due to infection, that impairs normal physiological activity. Parasitism and disease often are described mathematically by equations similar to the Lotka-Volterra predation equations, although they

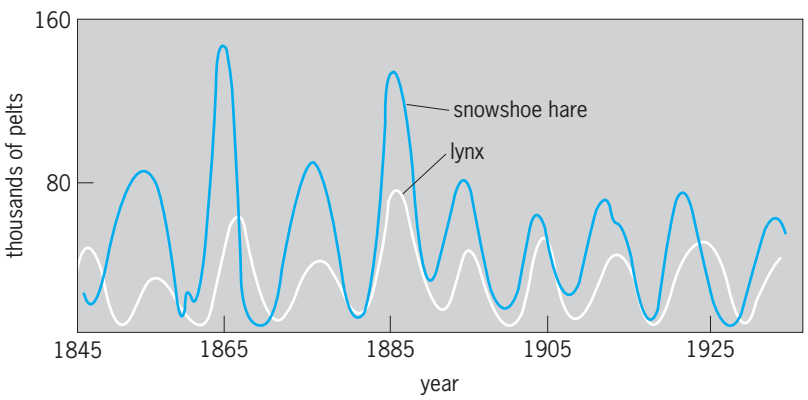


Fig. 4. Oscillations of predator and prey populations. Data are from pelt records of lynx and snowshoe hare of the Hudson's Bay Company, in Canada.

are commonly modified to account for the dispersal of parasites or pathogens between different hosts. These population interactions have been of growing interest in ecology, although they have been the subject of the entire field of epidemiology, an important aspect of medicine. The importance of parasitism and disease in natural ecosystems is evident from the extinctions of two important tree species, the American chestnut (*Castanea dentata*) and American elm (*Ulmus americana*), from the forests of eastern North America during this century. Those extinctions were caused, respectively, by chestnut blight and Dutch elm disease, both due to pathogens introduced into North America from Europe. *See* EPIDEMIOLOGY.

Symbiosis is a type of interaction between populations in which a participant is promoted by the interaction. There are two types of symbiosis: commensalism, which occurs when one participant is promoted by the interaction and the other is unaffected, and mutualism, which occurs when both participants are promoted by the interaction. Mutualism often involves organisms that are very different and have coevolved such a strong mutual interdependence that neither can exist in natural environments in the absence of the other. Symbiosis commonly is described mathematically by variations of the Lotka-Volterra competition equations in which the sign preceding the α becomes positive for one (commensalism) or both (symbiosis) populations. This transforms the competition coefficients into mutualism coefficients.

Mutualism is an extremely important type of ecological interaction since it often allows the symbionts to perform functions that have proven impossible for any single organism to accomplish with the same efficiency. Lichens are symbiotic associations between fungi and algae. They are capable of colonizing some of the most inhospitable environments on Earth, such as the bare faces of boulders and extremes of drought, cold, and heat. The fungal and algal species are so closely integrated that lichens have been classified by the binomial systems that are applied to other true species.

An example of the functional importance of symbiosis to ecology is the digestion of cellulose, which is accomplished largely by symbiotic associations. Cellulose is the most abundant organic molecule on Earth, but this vast energy source is largely inaccessible to living organisms except to symbionts. Among the most important symbiotic associations capable of digesting cellulose are ruminant animals and their gut flora, and termites and their gut protozoa. In the absence of these symbiotic associations, Earth would be covered with vast residues of undecomposed cellulose.

Synecology

Synecology is the study of groups of populations cooccurring in time and space; it is also called community ecology. Synecology is concerned with factors controlling the species composition of communities and why species composition changes in different environments and with time in the same habitat. Species composition of a community may include density, biomass, or productivity of individuals of different species.

The study of communities is one of the oldest branches of ecology, arising strongly in the studies of naturalists in the sixteenth to nineteenth century, and one of the most descriptive. Only since the mid-1960s have ecologists begun to turn significantly to experimental studies that manipulate either the environment or the species composition of communities. Those experiments are designed to understand mechanisms responsible for the species compositions of communities and the interactions that occur between different species in the community. *See* ECOLOGICAL COMMUNITIES.

Community organization. A principal concern of community ecology is the life forms of organisms, the taxonomic composition of co-occurring species groups, and the spatial arrangement of species. Plant life forms are relatively easy to characterize, and communities in widely separated geographic areas but similar environments commonly have a similar life-form composition. Most of the species in tropical rainforests, for example, are trees and shrubs, while most of the species in grasslands are small shrubs, and perennials with buds belowground. Although the species compositions of communities in different geographic areas may be quite different, the life forms are often quite similar in similar environments.

Another aspect of community organization that is a concern of synecology is spatial patterns, or dispersion. Ecologists typically distinguish between dispersal, the process of movement, and dispersion, the arrangement of individuals in space. The three types of dispersion patterns are random, regular, and clumped (aggregated). In a random distribution, the probability of an individual occupying a given location in a community is not influenced by the presence of another individual at a nearby location, which implies that no important factor, such as environment or parental proximity, has influenced the dispersion of individuals. In a regular distribution, the probability of an individual occupying a location is reduced if another individual occupies a nearby location; this distribution may be due to competition for an evenly spread environmental resource. In a clumped distribution, which may be due to proximity to seed source, or environmental heterogeneity, the probability of an individual occupying a location increases if another individual occupies a nearby location. *See* POPULATION DISPERSAL; POPULATION DISPERSION.

Temporal distributions are also a concern of synecology. By substituting the phrase "at a given time" for the phrase "at a given location" in the above definitions, temporal distributions are judged by the same criteria. Spatial and temporal patterns of individuals are important aspects of community ecology because they can provide insight into the underlying distributions of limiting factors in the environments and how species are related to those factors. Species

with distributions in which individuals are clumped in space or time often utilize resources that also are clumped in space or time. Limited dispersal also can lead to a clumped distribution due to the clustering of offspring around their parents. Regular distributions can often result from competition between individuals for a resource that is evenly distributed in space or time. Desert plants, for example, frequently have a regular distribution pattern because they compete for soil moisture, and each individual must utilize the moisture available in a certain minimum soil volume if it is to survive.

Niche. The niche concept refers to the environmental factors that control a species, and its distribution in relation to those factors. A species' niche is a consequence of the genetic properties of the individuals of that species, the environmental factors that affect it, and interactions with other species. G. E. Hutchinson described the niche as an n-dimensional hypervolume occupied by a population, where n is the number of environmental factors affecting it. He also distinguished between a fundamental niche, which is the genetic potential of the species, and a realized niche, which is the range of environmental factors occupied in the presence of other organisms.

Species are often distributed along environmental gradients as a series of bell-shaped curves (**Fig. 5**). Those curves represent the realized niches of species. Removal of other species, a type of experiment that is becoming increasingly common in synecology, would be required to reveal a species' fundamental niche. A species that is narrowly distributed has a narrow niche and is often referred to as an ecological specialist, while a widely distributed species has a broad niche and is often referred to as an ecological generalist. It is often inferred that intense interspecific competition will lead to specialization, restricting a species' realized niche to only that portion of the fundamental niche where it is a strong competitor. Conversely, intense intraspecific competition will tend to lead to a broader niche that reduces the effects of different members of the same species upon each other. However, there are as yet insufficient experimental studies to strongly support either of these inferences.

Species diversity. Species diversity is a property of a community that encompasses both the number of species present (species richness) and the equality of relative abundances of the species present (equitability of evenness). One of the principal concerns of synecology since its origin has been why communities differ so dramatically in the number of species present and the similarities of their relative abundances. Naturalists attached to exploratory expeditions during the historical colonial period were particularly overwhelmed by the variety of plant and animal life in the wet tropics. Tropical forests seemed to consist of an almost countless number of plant and animal species compared with communities in the temperate climates with which the naturalists were familiar. Indeed, there is a general gradient of increasing diversity from the poles to the Equator.

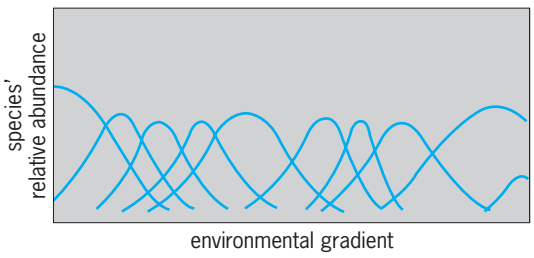


Fig. 5. A continuum of species distributions along an environmental gradient, where each line represents a different species. Note that peaks of species abundance do not overlap, and that breadth of curves differs between species.

Various explanations for the latitudinal differences in biotic diversity have been proposed, but none is wholly satisfactory and none has gained prominence. Chief among those explanations are: (1) Nontropical habitats have been subject to severe climatic fluctuations, including glaciation, that lead to repeated species extinctions. (2) The moderate climates of tropical locations allow species to become more specialized for narrow ranges of climate and other environmental factors. (3) Biological interactions, such as competition and predation, are stronger in tropical locations, leading to a proliferation of different adaptive types. (4) Productivity is higher in tropical locations, allowing more species to coexist due to greater available food. (5) Spatial variation of environments is greater in the tropics, leading to a proliferation of specialists in different areas. (6) There is greater environmental complexity in the tropics, partly because of more species, leading to greater specialization among species in a sort of positive feedback. *See* BIOGEOGRAPHY.

Community: reality vs. abstraction. One of the most contentious issues throughout the history of ecology has been the degree to which the species that co-occur in space and time represent an organized, meaningful biological unit and to what extent they are merely a fortuitous collection of individuals. Two of the principal ecologists in American ecology in the early decades of the twentieth century argued these points. F. Clements referred to communities as being like an organism, while H. A. Gleason said that communites were merely a happenstance collection of species with similar ecological tolerances. Clements's view prevailed until the middle of the twentieth century, and then Gleason's view gained ascendency, largely because of quantitative studies of community composition revealing that species distributions along environmental gradients were commonly the type of bell-shaped curves in Fig. 5. Those bell-shaped curves suggested that each species population had its own, individualistic response to the environment, and that the species that occurred at a given point on the gradient did so fortuitously.

Still, there is no gainsaying that similar communities occur in similar habitats. Although the species compositions of communities may vary considerably

in different geographic areas, the occurrence of communities with very similar life forms in geographically separated locations with similar environments suggests that a community is not merely a random collection of species. Many experiments in marine intertidal communities indicate that the removal of one species has a cascading effect that alters the abundances of many other species, even leading to the extinction of some species and invasion of the community by other species that were previously absent. These results indicate that the species in a community are interconnected through the functional processes of competition, predation, and symbiosis in ways producing a community that is not just a random collection of species.

Succession. Succession is the process of change in community composition and environmental properties over time on a site. Primary succession occurs in newly formed habitats, such as glacial moraines, river levees, or volcanic ash. Secondary succession occurs in habitats where the previous community has been destroyed or severely disturbed, such as following forest fire, abandonment of agricultural fields, or epidemic disease or pest attack.

A general consequence of succession is an amelioration of physical factors and a reduction of their importance as controlling factors, and an increase in the complexity and importance of biological factors. In newly formed ash fields or newly abandoned fields, sunlight is intense, nutrients are often in poor supply, winds are strong, and evaporative stress is severe. As such sites are colonized by plants, all of these physical factors tend to be modified and trophic webs develop that depend upon the plants. Species diversity increases as more species colonize, and increasingly complex trophic patterns develop. There often is an increase in the stature of the vegetation, which reduces sunlight, wind, and evaporative stress below the canopy. That creates new conditions that allow species to invade that were incapable of withstanding the physically harsh conditions early in succession.

Early in succession, there is directional change in both environmental factors and species composition of the community. Ultimately a relatively constant environment is reached and species composition no longer changes in a directional fashion but, instead, fluctuates about some mean, or average, community composition. That stage is called the climax.

As in the nature of the community itself, there is disagreement among ecologists about the exact nature of succession, and of the climax. Nevertheless, there is a general tendency for communities to become more complex and for biological factors to increase in importance and the effects of physical factors to be reduced as time passes. *See* ECOLOGICAL SUCCESSION.

Systems Ecology

Systems ecology is the study of dynamic relationships between the units of the ecosystem, particularly those relationships that influence rates of energy flow and chemical cycling. The ecosystem concept, as pointed out above, can be applied to many different levels of organization and varying degrees of complexity. Ecosystem science, or systems ecology, can be broadly divided into approaches concerned with relationships between trophic levels and those that employ with specific details of population dynamics of community organization within trophic levels. In both approaches the goal is to develop mechanistic explanations for the observed behavior of the system as a whole. Often this goal is accomplished through simulation modeling.

Simulation modeling involves the use of mathematical equations to characterize the state and dynamics of a system. Mathematical approaches to population dynamics, as seen earlier in this article, commonly use simplified equations, such as the logistic, to model the dynamics of one or a few populations. Simulation modeling differs principally in using a set of coupled equations to characterize the processes occurring in ecosystems. Those models often have much more detail than the simplified models of population dynamics, and require numerical solution using digital computers. A simulation model typically begins with a conceptual model or diagram consisting of boxes that represent components of system state and connecting arrows representing processes that interconnect the components. Simulation modeling uses rate equations to describe the processes and evaluate the mechanisms that are important to understanding both the dynamics and functional properties of ecosystems. Simulation models have as goals both isolating the general features of a broad range of similar ecosystems and providing sufficient detail to accurately predict how a specified system will respond to perturbations.

A system can be defined as a group of interacting and interdependent elements that form a collective entity. An ecosystem, therefore, is a collection of environmental factors and organisms that interact through specific, dynamic relationships defined by the food web of which they are members. Systems may be open or closed. A closed system is completely self-contained and does not interact with any other system; an example would be molecules of a gas enclosed in a container that was a perfect insulator. Those molecules would function independently of all other things. An open system interacts with other systems. All ecosystems are open systems. In particular, they are open to energy, whose ultimate source is the Sun. Proximally, however, energy can be imported into an ecosystem from some other system where the energy of photons from the Sun is converted into organic energy by photosynthesis.

Energy flow. Energy flow in ecosystems is approached in a way closely related to the laws of thermodynamics, that branch of physics that deals with the relationships among different forms of energy. The first law of thermodynamics states that the energy input to a system either is stored or is used to do work. It is generally given as Eq. (9), where Q

$$Q = \Delta E + W \tag{9}$$

is energy input, ΔE is a change in the energy content of the recipient system, and W is work done by that system. Ecologists restate the law as Eq. (10),

$$P_g = P_n + R \tag{10}$$

where P_g is gross productivity, P_n is net productivity, and R is respiration. The change in energy content, P_n, is therefore equal to the energy input minus the work done. That work has two components: one is maintenance respiration, the cost of maintaining the structure of the system at the time of energy input, and the other is growth respiration, the cost of synthesizing new biomass.

The second law of thermodynamics also is important to energy balance in ecosystems. That law states that the entropy of isolated systems always tends to increase. Entropy (S) is a measure of the randomness, disorder, or lack of organization of a system. In energy terms it is measured as the heat capacity of the system (joules $\cdot$ degree^{-1} $\cdot$ mole^{-1}) at a particular temperature (T). In isolated systems, entropy always increases. Because ecosystems are open systems, they are able to maintain their organization and avoid decaying to a state of maximum entropy and zero free energy. The maintenance component of respiration is the cost of maintaining present structure, that is, replacing the free energy degraded to heat. The growth component of respiration is the cost, that is, the free energy intake necessary to build structure and organization. Those costs are expressed both as an increase in the entropy (thus an increase in heat content) of the environment to which organisms in an ecosystem are coupled as well as the loss of heat produced by "friction" representing the degree of thermodynamic inefficiency of the metabolic energy transformations. *See* OPEN-SYSTEMS THERMODYNAMICS (BIOLOGY); THERMODYNAMIC PRINCIPLES.

A fundamental generalization of systems ecology is that net productivity diminishes about one order of magnitude at each level in the trophic web. If primary P_n averaged 100 g/m^2 $\cdot$ yr, secondary P_n at the herbivore level would be about 10 g/m^2 $\cdot$ yr, and at the carnivore level would be about 1 g/m^2 $\cdot$ yr. This so-called energy pyramid has two causes. First, the above-mentioned maintenance costs of respiration at each trophic level diminish the energy available to higher levels. Second, the ecological harvest of energy by any trophic level is not fully efficient. Therefore, the energy flow must diminish at each successive trophic level in an ecosystem.

In most ecosystems, certainly, the biomass of primary producers is greater than that of the herbivores which, in turn, is greater than carnivore biomass. Sometimes, however, particularly in aquatic habitats, biomass increases up the trophic web. That is possible only when lower trophic levels consist of organisms with very high rates of net productivity per unit of biomass. The ratio of energy flow to biomass is referred to as turnover time. If turnover time is short, as it is for small organisms with rapid life cycles, it is possible for the biomass pyramid to be inverted; that is probably most common in the open ocean where large, long-lived organisms are at upper trophic levels and small, short-lived organisms are at lower trophic levels. *See* ECOLOGICAL ENERGETICS.

Chemical cycling. Another fundamental concern of systems ecology is the pattern of chemical flow that is coupled to the process of energy flow. Each time that energy is used to do work, some chemicals are released back into the environment. Death also releases chemicals back into the environment.

There are two different types of trophic webs in ecosystems. The grazing food web is based on the consumption of the tissues of living organisms. The detritus food web is based on the consumption of dead organic material, called detritus in aquatic systems and sometimes referred to as litter in terrestrial ecosystems. Chemicals are recycled to the environment from the grazing food web each time that work is done, and as the excretory wastes of living organisms. Chemicals are recycled to the environment from the detritus food web as organisms utilize excretory wastes and the dead tissues of organisms.

Decomposition is the process of degrading the energy content of dead tissues and simultaneously releasing chemicals back into the environment. When those chemicals are released in inorganic forms, the process is called mineralization. Decomposition and mineralization typically involve many steps and take place in a trophic web fully as complex as those of the grazing food web. There is a size-dependent hierarchy as particles are broken into progressively smaller particles; that subdivision increases surface-volume ratios, exposing progressively more of the substance to decay. There also typically is a chemical hierarchy. Fats, proteins, and simple carbohydrates are the most easily utilized organic chemicals, and they generally are rapidly attacked. Wood, chitin, and bones are only slowly decomposed. More resistant substances accumulate in terrestrial ecosystems as soil humus and in aquatic ecosystems as sediments, or ooze, beneath the water. In geological time, those sediments were transformed into the hydrocarbon deposits that serve as sources of coal, oil, and gas for modern industrial society.

Microorganisms are exceptionally important in the processes of decomposition and mineralization. The final steps of mineralization are almost invariably accomplished by microorganisms. For example, the cycling of nitrogen in ecosystems involves microorganisms at several critical steps (**Fig. 6**). The only significant pathway introducing nitrogen into trophic webs is the process of nitrogen fixation in which gaseous, molecular nitrogen is converted into organic nitrogenous compounds by nitrogen-fixing bacteria and blue-green algae. Some of these bacteria and algae are free-living, and some participate in symbiotic associations with other organisms. The organic molecules are degraded in the detritus food web by ammonifying bacteria that release ammonium ion or ammonia. Nitrite bacteria convert the ammonia into nitrite, and nitrate bacteria oxidize the nitrogen further into nitrate; those conversions are referred to collectively as nitrification. Both

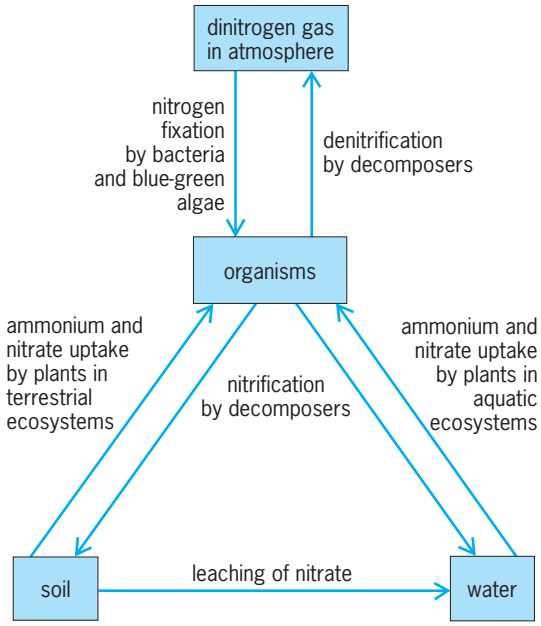


Fig. 6. Generalized nitrogen cycle. Arrows represent flow of nitrogen.

nitrate and ammonium ion can reenter the food web via uptake by plants. Denitrifying bacteria convert the nitrate into nitrogen gas, a process called denitrification that releases gaseous nitrogen back into the atmosphere. *See* BIOGEOCHEMISTRY; MATHEMATICAL ECOLOGY.

Applied Ecology

Applied ecology is that branch of ecology dealing with practical problems of immediate social importance. Ecology is often confused in the minds of laypersons with the environmental movement, a social interest group concerned with environmental degradation and problems of resource supply to human societies. Applied ecology makes a fundamental contribution to those social and political concerns by identifying environmental problems, gaging their significance, and suggesting potential solutions. Among the environmental issues to which ecology has made an important contribution have been problems of population growth and resource supply, acid rain, eutrophication, consequences of pollution, biological control of crop pests, range management, forestry, and the ecological consequences of nuclear war. From T. R. Malthus's essay on human population growth to recent scientific evaluations of nuclear war, ecology has always been concerned with problems that are important to the affairs of humans. *See* ECOLOGY, APPLIED.

Samuel J. McNaughton

Bibliography. B. Glaeser (ed.), *The Green Revolution Revisited*, 1987; H. Howe and L. Westley, *Ecological Relationships of Plants and Animals*, 1988, reprint 1990; J. Keating, *Interdependence in the Natural World*, 1987; R. P. McIntosh, *The Background of Ecology: Concept and Theory*, 1985, reprint 1986; M. B. Rambler et al. (eds.), *Global Ecology: Towards a Science of the Biosphere*, 1988.

Ecology, applied

The application of ecological principles to the solution of human problems and the maintenance of a quality life. It is assumed that humans are an integral part of ecological systems and that they depend upon healthy, well-operating, and productive systems for their continued well-being. For these reasons, applied ecology is based on a knowledge of ecosystems and populations, and the principles and techniques of ecology are used to interpret and solve specific environmental problems and to plan new management systems in the biosphere. Although a variety of management fields, such as forestry, agriculture, wildlife management, environmental engineering, and environmental design, are concerned with specific parts of the environment, applied ecology is unique in taking a view of whole systems, and attempting to account for all inputs to and outputs from the systems—and all impacts. In the past, applied ecology has been considered as being synonymous with the above applied sciences. *See* SYSTEMS ECOLOGY.

Ecosystem ecology. Ecological systems, or ecosystems, are complexes of plants, animals, microorganisms, and humans, together with their environment. Environment includes all those factors, physical, chemical, biological, sociocultural, which affect the ecosystem. The complex of life and environment exists as an interacting system and is unique for each part of the Earth. The unique geological features, soils, climate, and availability of plants, animals, and microorganisms create a variety of different types of ecosystems, such as forests, fields, lakes, rivers, and oceans. Each ecological system may be composed of hundreds to thousands of biological species which interact with each other through the transfer of energy, chemical materials, and information. The interconnecting networks which characterize ecosystems are often called food webs (**Fig. 1**). It is obvious from this structural feature of interaction that a disturbance to one population within an ecosystem could potentially affect many other populations. From another point of view, ecosystems are composed of chemical elements, arranged in a variety of organic complexes. There is a continual process of loss and uptake of chemicals to and from the environment as populations are born, grow and die, and are decomposed. Ecosystems operate on energy derived from photosynthesis (called primary production) and from other energy exchanges. The functional attributes of ecosystems, such as productivity, energy flow, and cycling of chemical elements, depend upon the biological species in the ecosystem and the limiting conditions of the environment. *See* BIOLOGICAL PRODUCTIVITY; ECOLOGICAL ENERGETICS; FOOD WEB.

Ecological systems develop in accord with the regional environment. Although these systems have evolved to resist the normal expected perturbations encountered in the environment, unusual disturbances and catastrophic events can upset and even destroy the system. In this case, recovery can occur

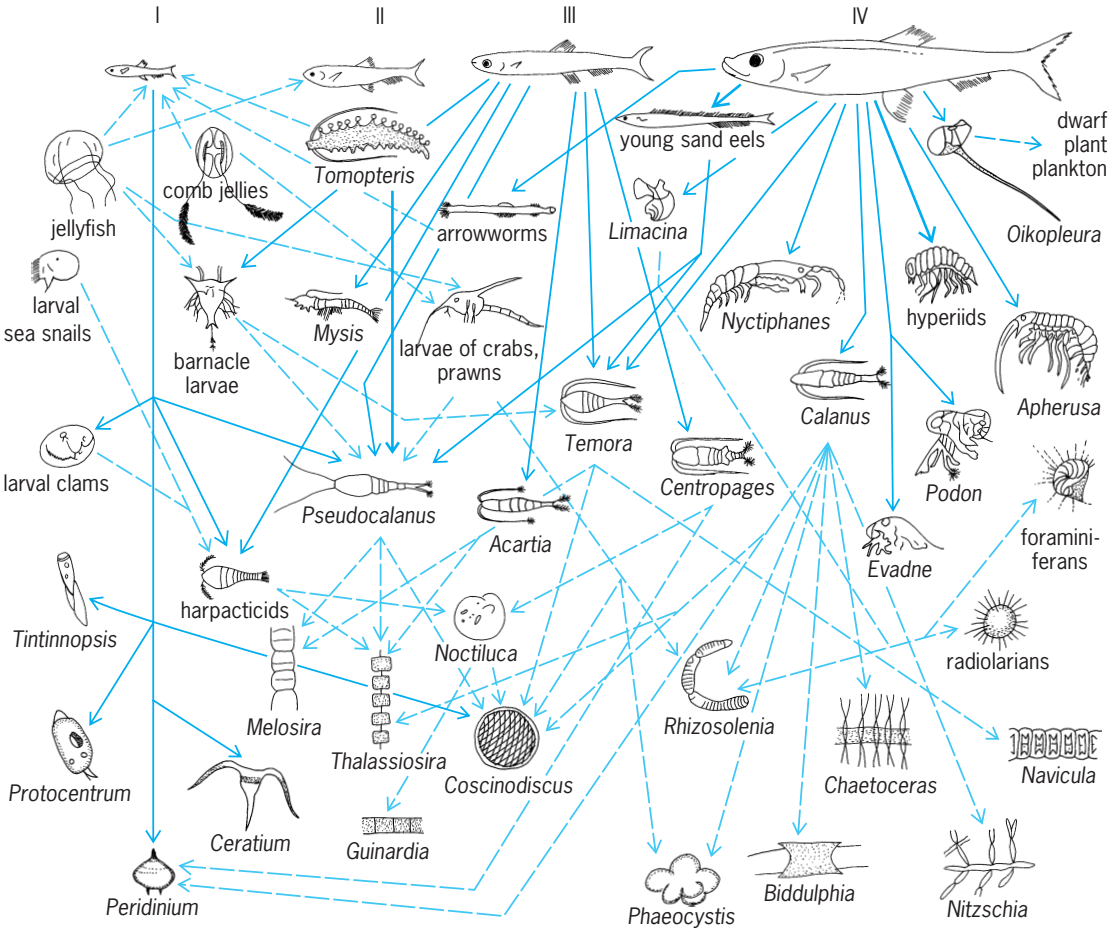


Fig. 1. Food relations of the herring at different stages in its life. Sizes of herring are (I) 0.6 to 1.3 cm, (II) 1.3 to 4.5 cm, (III) 4.5 to 12.5 cm, and (IV) over 12.5 cm. Solid lines indicate food eaten directly by herring. 1 cm = 0.4 in.

after the disturbance stops. Recovery is termed ecological succession since it comprises a sequence of communities which succeed each other until a dynamic steady state is reestablished. *See* ECOLOGICAL SUCCESSION.

Populations in these ecosystems fill a variety of structural and functional roles within the system. Often, groups of populations coevolve, so that they form a more or less isolated subunit. For example, the pollinators of a plant species, and their predators and parasites, form such a guild. Populations continually adapt and develop through natural selection, expanding to the limit of their resources. Population growth is, therefore, due to an increase in resources or a relaxation of limiting factors. *See* ECOSYSTEM; POPULATION ECOLOGY.

Ecosystem management theory. The objective of applied ecology management is to maintain the system while altering its inputs or outputs. Often, ecology management is designed to maximize a particular output or the quantity of a specific component. Since outputs and inputs are related (**Fig. 2**), maximization of an output may not be desirable; rather, the management objective may be the optimum level. Optimization of systems can be accomplished through the use of systems ecology methods which consider all parts of the system rather than a specific set of components. In this way, a series of strategies or scenarios can be evaluated, and the strategy producing the largest gain for the least cost can be chosen for implementation.

The applied ecology management approach has been partially implemented through the U.S. National Environmental Policy Act. This act requires that an environmental-impact analysis be carried out by an interdisciplinary team of specialists

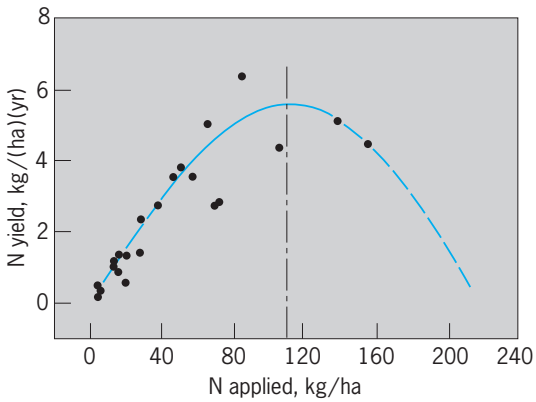


Fig. 2. Relationship between N yield and N consumption (1964–1969).

representing the subjects necessary for an ecosystem analysis. This team is required to evaluate a project on the basis of its environmental, social, and cultural features. One alternative that must be considered is that of no alteration of the system.

A variety of general environmental problems within the scope of applied ecology relate to the major components of the Earth: the atmosphere, water, land, and the biota. The ecological principles used in applied ecology are discussed elsewhere; a sequence of environmental problems of special importance to applied ecology is discussed below.

Atmospheric problems. The atmosphere is one of the most important components of the environment to consider from the viewpoint of applied ecology since it connects all portions of the Earth into one ecosystem. The atmosphere is composed of a variety of gases, of which oxygen and nitrogen make up the largest percentage. It is not uniform in its depth or its composition, but is divided into several layers or zones which differ in density and composition. Although most interaction with humans occurs in the zone nearest Earth, the most distant parts of the atmosphere are also important since they affect the heat balance of the Earth and the quality of radiant energy striking the surface. Disturbances to these portions of the atmosphere could affect the entire biosphere. *See* ATMOSPHERE.

The composition of the atmosphere varies according to location. The qualities of minor gases such as carbon dioxide, the amounts of various metallic elements, and the quantity of water vapor and dust all may differ, depending on the relative distance from land or sea. But, in addition, the atmospheric composition may change in time. For example, over the history of the planet, the percentage composition of oxygen has changed from a very oxgen-poor environment to the present atmosphere, with 20.95% oxygen by volume. *See* ATMOSPHERE, EVOLUTION OF; ATMOSPHERIC CHEMISTRY.

Human activities may introduce a variety of pollutants into the atmosphere. The principal pollutants are carbon dioxide, sulfur compounds, hydrocarbons, nitrogen oxides, solid particles (particulates), and heat. The amounts of pollutants that are produced may be quite large, especially in local areas, and have increased in amount as industrialization has become more widespread. Industrial and domestic activities have also been estimated to put 2.18×10^8 tons (1.96×10^8 metric tons) of sulfur into the atmosphere per year. In most cases, these pollutants have increased during the recent past, and in many areas have become a serious problem. *See* AIR POLLUTION.

Atmospheric problems can have a variety of impacts on humans. Numerous observers have attributed climatic change to atmospheric alteration, since any change in the gaseous envelope of the Earth could alter the heat balance and the climate. The Earth's climate is not constant, and it is difficult to establish an exact correlation between pollution and variation in temperature or solar radiation at the Earth's surface. The pollutants most likely to have an effect on climate are carbon dioxide and solid particles. Pollution may also affect the chemical balance of regions of the Earth. These effects may be extremely complicated. For example, it has been reported that sulfur oxides produced in the industrial districts of northern Europe have moved north in the atmosphere over Scandinavia. The sulfur oxides react with water to form sulfuric acid, which rains out in precipitation. The acid rain changes the acidity of the soil in this region and may depress the activity of blue-green algae, which fix nitrogen from the atmosphere. Acid rain appears to cause reduction in the growth of trees. Other pollutants may act in a similar complicated fashion through the network of interactions in the Earth ecosystem. *See* ACID RAIN.

Finally, atmospheric pollution has direct effects on plants and animals and human activities. Pollutants, like other materials, can act as limiting influences on the growth, reproduction, and survival of plants and animals. A variety of plants, such as lichens and mosses, that are extremely sensitive to pollution can be used to indicate the degree of atmospheric deterioration. In some severe cases, all vegetation and animal life may be destroyed in the vicinity of the polluting industry. Gases and solids are taken into the lungs of humans, and cause disease or discomfort. In cities, such pollutants as asbestos and lead are exceedingly dangerous to the population. And finally, the impact of pollutants on buildings, clothes, artwork, and machines is costly.

Control of atmospheric pollution requires interception of pollutants at the point of discharge. Industrial control can be achieved by the use of special filters, precipitators, and other devices. Control of pollution for automobiles also may involve special equipment, as well as redesign of engines and fuel. Reduction in dust and similar general sources of air pollution may demand a change in the operation producing the problem.

Water problems. The aquatic environment is of equal importance to applied ecology since most of the Earth's surface is covered by the oceans, and the land is connected to the oceans by streams and rivers. Thus, like the atmosphere, the waters are a connection between distant parts of the biosphere and can carry a disturbance from its origin to another region.

The composition of waters varies widely, and it is essential in evaluating aquatic health to establish the base-line conditions which are stable under the normal or undisturbed conditions. Water pollution arises from a variety of sources—industry, domestic sewage, agricultural fertilizers and feedlots, construction activities, and forest practices. Principal pollutants are sediments, organic pollutants containing nitrites, and phosphates, lead, mercury, hydrocarbons, pesticides, and various synthetic chemicals. The impact of the pollutant depends both on its chemical nature and on the quantity released. All water bodies receive quantities of chemicals and solid materials; a variety of organisms which break down and utilize these inputs have evolved. Serious problems arise when the inputs to the water body become unusually large or contain synthetic materials

which cannot be decomposed by the extant organisms. Further problems may develop through concentrations of pollutants in the food web of the aquatic system. For example, if a chemical is not metabolized by organisms, but is concentrated in their tissues, as are some pesticides, then as each organism is eaten by another, the body burden of chemicals can increase. In this way, predators may obtain very large and dangerous amounts of pollutants. The decline in population of certain fish-eating birds has been attributed to this process of transfer and concentration.

Aquatic pollution has many consequences of significance to humans. An excess of chemical materials which enrich plant and animal growth can cause rapid increase in life. This process, termed eutrophication, may entail dramatic increases in the algal, planktonic, or rooted aquatic plant populations, with the result that the water body appears green in color or becomes clogged with vegetation. Toxic chemicals released to water bodies may directly kill aquatic life or, if present in sublethal amounts, may change the species of plants and animals present. Often aquatic pollution is not a dramatic either-or proposition, with all fish or other aquatic life killed; rather, more commonly a trend toward an increase in the more resistant species is seen, with the elimination of those forms which are especially susceptible to the pollutant. Aquatic pollution also involves heat, especially that derived from industrial activities, including nuclear power and fossil fuel plants. In these instances, water is used to cool the machines or reactors and is exhausted to the environment at elevated temperatures. Since all metabolic and chemical processes are influenced by heat, thermal pollution should have a significant effect on aquatic systems, but thus far it has been difficult to prove that such an impact occurs.

Other aquatic problems of interest to applied ecology concern alteration of water channels by impoundments or channelization and irrigation. In each instance the natural pattern of water movement is altered, and deterioration of the environment may result. Impoundments limit the natural movement of sediment and chemical elements; production patterns in the water and lands below the impoundment may be altered, and other changes may occur. But, on the plus side, impoundments often provide fisheries, electrical energy, recreation, and other advantages. Irrigation problems may involve the movement of salts from depths in the soil, with deposition near the surface. Disturbance of the chemical equilibrium of the soil, in turn, interferes with plant growth. *See* WATER POLLUTION.

Terrestrial and soil problems. Terrestrial environments constantly undergo a degrading and decomposing process owing to the action of water, frost, wind, and other environmental processes on the surface which involve the linkages joining land, water, and atmosphere. Human activities may accelerate these natural processes. In addition, the use of chemical materials on the land may have effects similar to those resulting from their addition to water. Most terrestrial environmental problems are caused by agricultural, grazing, or forest practices. Probably the most serious effect concerns practices which increase the rate of surface erosion. Only a small percentage of the Earth's surface is suitable for agriculture, and the loss of soil from these areas is extremely serious. As a consequence, certain regions have been denuded and are no longer productive. Overgrazing may also remove the cover of vegetation and allow water and wind to erode the soil. Deserts have increased in extent almost everywhere because of overgrazing, and in India the increase in the Rajasthan Desert can be measured in feet per year. Dust from this desert blows as far east as Thailand. Overcutting trees and lack of reforestation programs also may increase soil erosion and nutrient losses in forest regions. These impacts are not solely the mark of modern civilization. Misuse of the land has been noted in many past civilizations and can even be a problem for present-day primitive societies. *See* DESERTIFICATION; EROSION; FOREST AND FORESTRY.

However, modern agriculture has added new problems to those of the primitive farmer. Various chemicals such as fertilizers, pesticides, and herbicides are used to increase agricultural production. These chemicals may be needed because of past misuse of the land or because of economic demands in a society that does not recognize the need to maintain and protect terrestrial resources. Organic and ecological agricultural practices seek to reestablish a pattern of land use without causing deterioration of the soil and biotic resources. Although reestablishment of the pattern may result in somewhat lower productivity, proponents of "ecoagriculture" argue that high productivity can be maintained without loss of soil through erosion and without a reduction in fertility. *See* AGRICULTURE.

Probably the most serious short-term impact arises from the use of chemicals on the land. In the most extreme cases, the health of the agriculturalist may be affected by the materials. But more commonly, the pesticide or chemical in the soil is taken up by the crop and then enters the human food chain. Many modern governments maintain agencies to advise farmers on the proper amounts of chemicals to apply so that buildup does not occur; other agencies periodically sample and analyze foodstuffs for residues. In this way, the consumer can be protected from misuse of chemicals or from excessive concentrations. Unfortunately, these agencies seldom consider the impacts of agricultural chemicals on other animal food chains. For example, the soil fauna and the natural nitrogen-fixing organisms present in the soil, as well as the terrestrial faunas living near the agricultural or forest plantations, can be significantly affected; and populations of animals, even beneficial species, may be reduced through misuse of chemical materials. However, extinction of plants and animals, which is also a serious applied ecology problem, usually is due to the destruction of their habitat. Pollution, disturbance of the land, and overhunting may provide the final cause of the destruction of a particular living species.

Nuclear energy. Industrialized societies require large quantities of energy. Energy production from nuclear reactors has been enthusiastically developed in many regions of the biosphere. However, nuclear energy also has environmental consequences that are of concern to applied ecology and that must be considered when these facilities are designed and operated, so that negative impacts on the environment do not occur.

Nuclear energy has three primary environmental consequences: the storage of radioactive products, release of radioactive material to the environment, and, as mentioned above, release of heat.

There are various kinds of radioactivity associated with the particular elements used in the reactor. In the process of the generation of energy, these fuel elements are changed into a suite of radioactive materials. Although these materials, in turn, form a new source of chemicals, the process of separation and concentraton is very costly and dangerous. In either case, however, the processes result in radioactive waste that must be stored for periods of hundreds, even thousands of years. The fact that the potential danger of these wastes will require technical attention for periods of time longer than the histories of many modern societies is a prime argument against the widespread use of nuclear energy. However, proponents of the use of nuclear energy state that certain geological structures such as salt mines can be safely used for storage indefinitely.

A second environmental problem concerns the loss of relatively small quantities of radioactive materials to the environment during chemical processing in reactors or chemical plants. If these materials enter the body, they can cause disease and death. Like pesticides, radioactive chemical materials may be concentrated in food chains and can appear in relatively large concentrations. In certain fragile environments such as the arctic tundra, the food chains are very short. Thus, radioactive chemicals derived from testing atomic weapons pass through lichens or reindeer or caribou to humans. Concentrations in certain localities may be high enough to cause concern to public health authorities. *See* NUCLEAR REACTOR; RADIOACTIVITY.

Population problems. Applied ecology also is concerned with the size of the human population, since many of the impacts of human activities on the environment are a function of the number and concentration of people. The human population on Earth has increased exponentially, and in many countries this increase poses almost insurmountable problems. Control of environmental degradation, even a concern for environment, is nonexistent when the population is undernourished, starving, ill-housed, and underemployed. Social disorder, alienation, psychological disturbances, physical illness, and other problems have been correlated with overpopulation. Population also places demands on the resources of the Earth; as the standard of living rises, these demands increase.

The human population problem is exceedingly complicated because the growth of population is controlled largely by the decisions of individual families. The family may visualize several different strategies—the number of children that is best for the family, best for their social group or tribe, or best for the human race—or have no strategy at all. Families may decide that a large family is best even under serious conditions of overpopulation. Considerable evidence indicates that the size of the family declines as the population becomes less rural and more industrial. Thus, some specialists urge economic development, regardless of environmental impact, as a means of solving the population problem. Others urge that the family be more directly influenced to reduce the number of children. Direct action might entail birth control advice, medical abortion and sterilization, and taxation. Yet others argue that no measures such as these can be significant and that the human population will be controlled by famine, war, or disease. Each of these positions leads toward a set of social policies, all of which have an environmental impact which, in turn, affects the human society. *See* HUMAN ECOLOGY.

Environmental planning and design. The foregoing discussion suggests that there is an optimum environment for the human race which is influenced by a variety of population densities and activities. Thus, although the population of the United States is relatively sparse, it has a large environmental impact. This, in turn, suggests that the human environment and society could be designed in such a way to minimize the negative impacts and provide a satisfactory productive life for the population. Environmental design considers economic and social policy, as well as the impact of designed rural, urban, transport, industrial, and other systems. It also considers the design of the individual environment of house, furniture, clothes, and so on. Considering the often violent impact modern society has had on the environment, a design revolution is required to reorganize the environment created by society so that these impacts can be reduced.

Environmental planning and design obviously have a deep political component, since the methods used to redesign society depend upon the control of individual demands. At one extreme, individual demand is allowed to express itself without limit, and education is used to create in the individual a realization that environmental constraints must be recognized. At the other extreme, the government or party controls demand through regulation. Most societies operate somewhere between these extremes.

Throughout human history, utopian designs have been developed for human societies and environments. Today these designs pay more attention to environmental features of society and are often labeled as ecological. Applied ecology, thus, considers not only the alteration of specific features of the modern industrial society to correct some environmental defect but also the fundamental reorientation of society to achieve a balance between humans and the natural world on which they depend. *See* ECOLOGY.

Frank B. Golley

Ecosystem

A functional system that includes an ecological community of organisms together with the physical environment, interacting as a unit. Ecosystems are characterized by flow of energy through food webs, production and degradation of organic matter, and transformation and cycling of nutrient elements. This production of organic molecules serves as the energy base for all biological activity within ecosystems. The consumption of plants by herbivores (organisms that consume living plants or algae) and detritivores (organisms that consume dead organic matter) serves to transfer energy stored in photosynthetically produced organic molecules to other organisms. Coupled to the production of organic matter and flow of energy is the cycling of elements. *See* ECOLOGICAL COMMUNITIES; ENVIRONMENT.

Autotrophic production. All biological activity within ecosystems is supported by the production of organic matter by autotrophs (organisms that can produce organic molecules such as glucose from inorganic carbon dioxide; **Fig. 1**). More than 99% of autotrophic production on Earth is through photosynthesis by plants, algae, and certain types of bacteria. Collectively these organisms are termed photoautotrophs (autotrophs that use energy from light to produce organic molecules). In addition to photosynthesis, some production is conducted by chemoautotrophic bacteria (autotrophs that use energy stored in the chemical bonds of inorganic molecules such as hydrogen sulfide to produce organic molecules). The organic molecules produced by autotrophs are used to support the organism's metabolism and reproduction, and to build new tissue. This new tissue is consumed by herbivores or detritivores, which in turn are ultimately consumed by predators or other detritivores.

Terrestrial ecosystems, which cover 30% of the Earth's surface, contribute a little over one-half of the total global photosynthetic production of organic matter (approximately 60×10^{15} grams of carbon per year; **Fig. 2**). The global rate of photosynthetic production of organic matter in terrestrial ecosystems is regulated by temperature, water availability, and nutrient concentrations. Correspondingly, the rate of photosynthetic production in the tropics (average rate = 1800 g C m^{-2} y^{-1}) is 25-fold greater than in deserts (average rate = 70 g C m^{-2} y^{-1}) and 13-fold greater than in tundra and alpine ecosystems (average rate = 140 g C m^{-2} y^{-1}).

Oceans, which cover 70% of the Earth's surface, produce approximately 51×10^{15} g C y^{-1} of organic matter (Fig. 2). Photosynthesis in oceans varies greatly across the Earth in response to nutrient availability (primarily nitrogen and iron). Photosynthesis is greatest along continental margins and in regions where deep, nutrient-rich water upwells into surface waters receiving light. Photosynthetic production is greatest in regions of deep upwelling (averaging 420 g C m^{-2} y^{-1}), intermediate in coastal margins (averaging 250 g C m^{-2} y^{-1}), and lowest in the open ocean (averaging 130 g C m^{-2} y^{-1}).

Whereas most ecosystems receive sunlight to support photosynthesis, photosynthetic production does occur in the deep ocean and in caves. In a few unique ecosystems, such as deep-sea thermal vents and caves, the basis of organic matter production is by chemoautotrophic bacteria. These ecosystems harbor surprisingly diverse communities that function in the complete absence of sunlight.

Heterotrophic production. The organic matter produced by autotrophic organisms is ultimately consumed by heterotrophs (organisms that obtain energy and nutrients by consuming other organisms; Fig. 1). Globally, most organic matter produced by plants and algae is consumed by detritivores. Herbivores consume only about 5% of photosynthetic production, fire destroys another 5%, and a small fraction of autotrophic production is lost through burial within the Earth's crust (to potentially become fossil fuels). The proportion of photosynthetic production consumed by herbivores can vary considerably among ecosystems. In terrestrial ecosystems, much plant tissue is structural molecules such as cellulose and lignin that are difficult or impossible for most organisms to digest. In aquatic ecosystems, in which algae dominate, a much greater proportion of photosynthetic production is consumed by herbivores. Algae have little in the way of structural molecules and contain more molecules that can be consumed by grazing organisms. *See* BIOMASS; ECOLOGICAL SUCCESSION.

Food webs. Organisms are classified based upon the number of energy transfers through a food web (Fig. 1). Photoautotrophic production of organic matter represents the first energy transfer in ecosystems and is classified as primary production. Consumption of a plant by a herbivore is the second energy transfer, and thus herbivores occupy the second trophic level, also known as secondary production. Consumer organisms that are one, two, or three transfers from photoautotrophs are classified as primary, secondary, and tertiary consumers. Moving through a food web, energy is lost during each transfer as heat, as described by the second law of thermodynamics. Consequently, the total number of

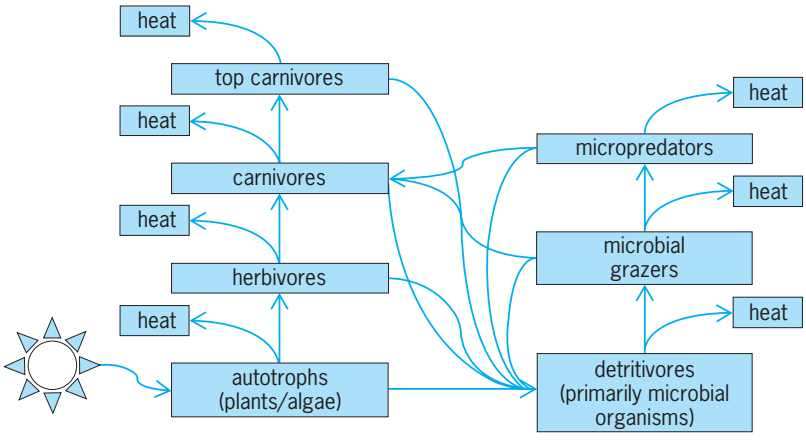


Fig. 1. General model of energy flow through ecosystems.

energy transfers rarely exceeds four or five; with energy loss during each transfer, little energy is available to support organisms at the highest levels of a food web. *See* ECOLOGICAL ENERGETICS; FOOD WEB.

Biogeochemical cycles. In contrast to energy, which is lost from ecosystems as heat, chemical elements (or nutrients) that compose molecules within organisms are not altered and may repeatedly cycle between organisms and their environment. Approximately 40 elements compose the bodies of organisms, with carbon, oxygen, hydrogen, nitrogen, and phosphorus being the most abundant. If one of these elements is in short supply in the environment, the growth of organisms can be limited, even if sufficient energy is available. In particular, nitrogen and phosphorus are the elements most commonly limiting organism growth. This limitation is illustrated by the widespread use of fertilizers, which are applied to agricultural fields to alleviate nutrient limitation. *See* BIOGEOCHEMISTRY.

Nitrogen cycle. Nitrogen commonly limits the rate of primary production in terrestrial, fresh-water, estuarine, and oceanic ecosystems (**Fig. 3**). In one turn of the biogeochemical cycle of nitrogen, (1) cyanobacteria and certain types of eubacteria transform atmospheric dinitrogen into organic molecules such as protein (nitrogen fixation); (2) organisms are consumed by other organisms and organic nitrogen is converted into ammonium as a waste product (decomposition); (3) ammonium either is assimilated by plants, algae, and bacteria into organic forms or is converted to nitrate by chemoautotrophic bacteria (nitrification); and (4) nitrate is converted back to dinitrogen by anaerobic bacteria (denitrification). Whereas nitrogen is very abundant in the atmosphere as dinitrogen (N_2), only a select group of organisms can transform dinitrogen into organic forms (amino acids), and the transformation requires a lot of metabolic energy. Humans have greatly impacted the global nitrogen cycle by doubling the amount of ammonium, nitrate, and organic nitrogen in the environment through the widespread use of fertilizers and combustion of fossil fuels. This massive change in nitrogen forms has led to problems with acid precipitation, alterations of global patterns of primary production rates, increased production of nitrous oxide (a greenhouse gas), and pollution of aquatic ecosystems.

Phosphorus cycle. Phosphorus also commonly limits primary production. In contrast to nitrogen, phosphorus lacks an atmospheric phase. The primary source of phosphorus to ecosystems is the weathering of the Earth's crust. This inorganic phosphorus is then assimilated by plants, algae, and bacteria and is incorporated into organic forms such as phospholipids and adenosine triphosphate (ATP). As organisms are consumed by other organisms, organic phosphorus is converted back to inorganic forms. Both organic and inorganic forms are slowly carried by water from plants and soils to streams and rivers, and ultimately to the world's oceans where phosphorus is buried in sediments. The global phosphorus cycle is completed over long, geologic time scales

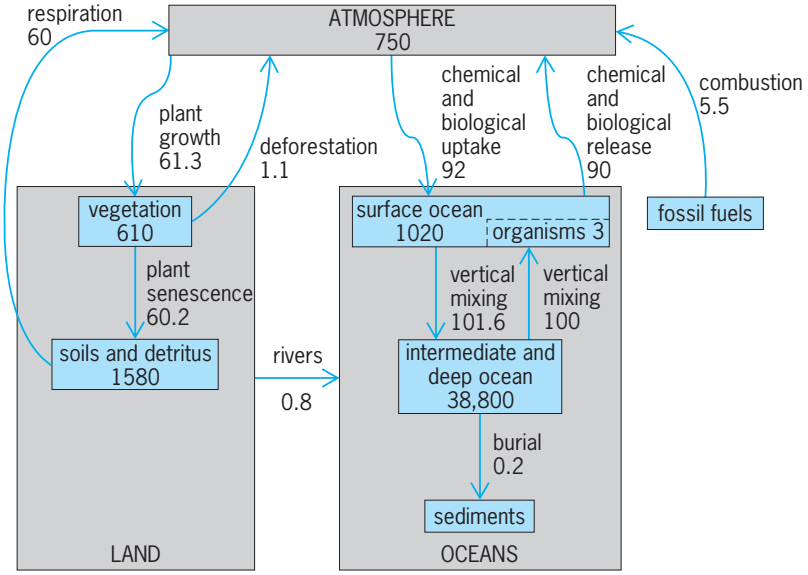


Fig. 2. Model of the global carbon cycle illustrating movement of carbon between the atmosphere and terrestrial and oceanic ecosystems. The storage of carbon in these ecosystems is in picograms (1 pg = 10^{15} g), and fluxes of carbon between boxes are in pg y^{-1}.

when ocean sediments are compacted into rock and this new rock is uplifted onto continents. As with the nitrogen cycle, humans have dramatically impacted the phosphorus cycle by application of fertilizers. The major source of phosphorus fertilizers is from the mining of guano. Current estimates project readily accessible sources of guano being depleted by the end of the twenty-first century.

Carbon cycle. The carbon cycle has also received considerable research focus, especially with the increase of carbon dioxide in the atmosphere due to the combustion of fossil fuels, and the impact on global climate. Carbon cycles between the atmosphere and terrestrial and oceanic ecosystems. This cycling results, in part, from primary production and decomposition of organic matter. Rates of primary production and decomposition, in turn, are regulated

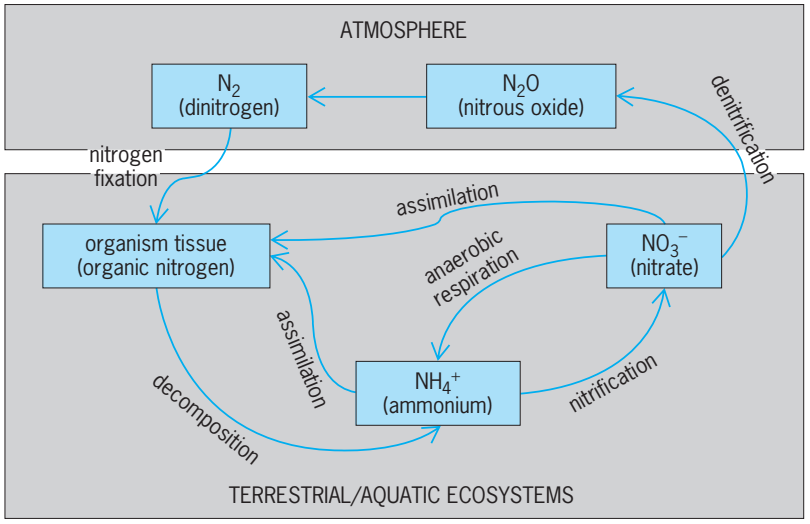


Fig. 3. General model of the nitrogen cycle in ecosystems.

by the supply of nitrogen, phosphorus, and iron. The combustion of fossil fuels is a recent change in the global cycle that releases carbon that has long been buried within the Earth's crust to the atmosphere. Carbon dioxide in the atmosphere traps heat on the Earth's surface and is a major factor regulating the climate. This alteration of the global carbon cycle along with the resulting impact on the climate is a major issue under investigation by ecosystem ecologists. *See* AIR POLLUTION; CONSERVATION OF RESOURCES; ECOLOGY, APPLIED; HUMAN ECOLOGY; WATER POLLUTION.

Jeremy B. Jones

Bibliography. F. B. Golley, *A History of the Ecosystem Concept in Ecology*, Yale University Press, 1993; J. B. Hagen, *An Entangled Bank: The Origins of Ecosystem Ecology*, Rutgers University Press, 1992; W. H. Schlesinger, *Biogeochemistry: An Analysis of Global Change*, 2d ed., Academic Press, 1991; R. H. Waring and S. W. Running, *Forest Ecosystems: An Analysis at Multiple Scales*, Academic Press, 1998.

Ecotone

A geographic boundary or transition zone between two different groups of plant or animal distributions. The term has been used to denote transitions at different spatial scales or levels of analysis, and may refer to any one of several attributes of the organisms involved. For example, an ecotone could refer to physiognomy (roughly, the morphology or appearance of the relevant organisms), such as between the boreal forest and grassland biomes; or it could refer to composition, such as between oak-hickory and maple-basswood forest associations; or it could refer to both. Ecotones are generally distinguished from other geographic transitions of biota by their relative sharpness. The ecotone between boreal forest and prairie in central Saskatchewan occurs over a hundred kilometers or so, in contrast to the transition from tropical forest to savanna in South America or Africa that is associated with increasing aridity and is dispersed over hundreds of kilometers. The "tension zone" between broadleaf deciduous forests in south-cental Michigan and mixed forests to the north is similarly sharp. Ecotones are thought to reflect concentrated long-term gradients of one or more current environmental (rather than historical or human) factors. Though often climatic, these factors can also be due to substrate materials, such as glacial sediments or soils. Regardless of their specific environmental basis, most ecotones are thought to be relatively stable.

A good example of an ecotone is the abrupt prairie-forest transition in North America that extends southeastward from south-central Alberta, Canada, through central Minnesota, and ultimately (in more diffuse form) to approximately southern Lake Michigan. Separating the grassland biome in the western Midwest and Great Plains from the boreal and deciduous forest biomes to the northeast, this ecotone is thought to reflect long-term patterns of precipitation and evaporation (warmer, drier to the southwest). These, in turn, are produced by modal (most frequent) atmospheric circulation patterns over North America as they repeatedly steer daily weather features across the region and ultimately generate environments, either directly or indirectly through patterns of disturbance, that are conducive to prairie or forest persistence in different parts of the region. Typically, specific sites within an ecotone will support communities that are more like one or the other broad groupings of biota the ecotone separates, depending on local factors including environmentally related patterns of disturbances such as fire.

Ecotones are often reflected in the distributions of many biota besides the biota used to define them. The prairie-forest ecotone, for example, is defined not only by the dominant vegetation components but also by many faunal members of the associated ecosystems, such as insects, reptiles and amphibians, mammals, and birds, that reach their geographic limits here. Across the glaciated midwestern United States, where landscapes of different geologic ages co-occur, the associated soil patterns are sometimes reflected in similarly bold ecological patterns. In central Illinois, for example, the transition from a Wisconsinan-aged glacial landscape (deposited approximately 20,000 years ago) in the north to an Illinoisan-aged landscape (several hundred thousand years older) to the south corresponds to a transition to older (better-developed) soils, different dominant tree species in the upland wooded tracts, and even contrasts in the herpetofauna, with many species in one area historically absent from the other. *See* ALPINE VEGETATION; BIOME; ECOLOGICAL COMMUNITIES; ECOSYSTEM; FOREST ECOSYSTEM; GRASSLAND ECOSYSTEM; LIFE ZONES; SAVANNA; ZOOGEOGRAPHY.

Jay R. Harman

Bibliography. J. A. Harrington and J. R. Harman, Climate and vegetation in central North America: Natural patterns and human alterations, *Great Plains Quart.*, 11:103–112, 1991; J. B. Lachavanne and R. Juge (eds.), *Biodiversity in Land/Inland Water Ecotones*, CRC Press–Parthenon, 1997; K. E. Medley and J. R. Harman, Relationships between the vegetation tension zone and soils distribution across central Lower Michigan, *Mich. Bot.*, 26:78–87, 1987.

Eddy current

An electric current induced within the body of a conductor when that conductor either moves through a nonuniform magnetic field or is in a region where there is a change in magnetic flux. It is sometimes called Foucault current. Although eddy currents can be induced in any electrical conductor, the effect is most pronounced in solid metallic conductors. Eddy currents are utilized in induction heating and to damp out oscillations in various devices.

Causes. If a solid conductor is moving through a nonuniform magnetic field, electromotive forces (emfs) are set up that are greater in that part of the conductor that is moving through the strong part of

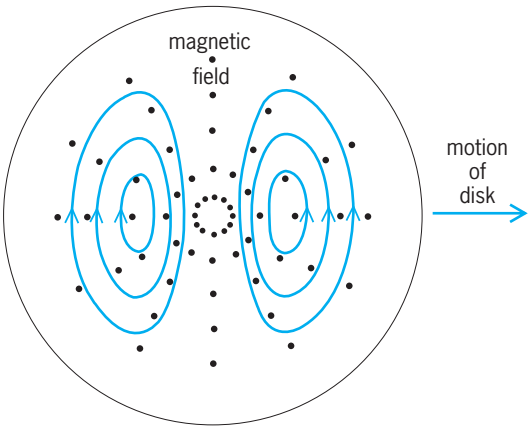


Fig. 1. Eddy currents which are induced in a disk moving through a nonuniform magnetic field.

the field than in the part moving through the weaker part of the field. Therefore, at any one time in the motion, there are many closed paths within the body of the conductor in which the net emf is not zero. There are thus induced circulatory currents that are called eddy currents (**Fig. 1**). In accordance with Lenz's law, these eddy currents circulate in such a manner as to oppose the motion of the conductor through the magnetic field. The motion is damped by the opposing force. For example, if a sheet of aluminum is dropped between the poles of an electromagnet, it does not fall freely, but is retarded by the force due to the eddy currents set up in the sheet. If an aluminum plate oscillates between the poles, it will be stopped quickly when the switch is closed and the field set up. The energy of motion of the aluminum plate is converted into heat energy in the plate. *See* ELECTROMAGNETIC INDUCTION; LENZ'S LAW.

Eddy currents are also set up within the body of material when it is in a region in which the magnetic flux is changing rapidly, as in the core of a transformer. As the alternating current changes rapidly, there is also an alternating flux that induces an emf in the secondary coil and at the same time induces emfs in the iron core. The emfs in the core cause eddy currents that are undesirable because of the heat developed in the core (which results in high energy losses) and because of an undesirable rise in temperature of the core. Another undesirable effect is the magnetic flux set up by the eddy currents. This flux is always in such a direction as to oppose the change that caused it, and thus it produces a demagnetizing effect in the core. The flux never reaches as high a value in the core as it would if there were no eddy currents present.

Laminations. Induced emfs are always present in conductors that move in magnetic fields or are present in fields that are changing. However, it is possible to reduce the eddy currents caused by these emfs by laminating the conductor, that is, by building the conductor of many thin sheets that are insulated from each other rather than making it of a single solid piece. In an iron core the thin iron sheets are insulated by oxides on the surface or by thin coats of varnish. The laminations do not reduce the induced emfs, but if they are properly oriented to cut across the paths of the eddy currents, they confine the currents largely to single laminae, where the paths are long, making higher resistance; the resulting net emf in the possible closed path is small. Bundles of iron wires or powdered iron formed into a core by high pressure are also used to break up the current paths and reduce the eddy currents. *See* CORE LOSS.

Kenneth V. Manning

Testing. The testing of metallic materials through the detecting of eddy currents induced within them is widely used. A high-frequency alternating current in a coil placed close to a metallic material induces eddy currents within the material through electromagnetic induction. The flux caused by the eddy currents threads the coil producing the original flux and alters its apparent impedance. This alteration is a measure of the strength of the eddy currents. The flow of the eddy current is impeded by the material's electric resistance, which reflects characteristics such as hardness and chemical composition. Therefore, eddy current measurements reveal various properties of materials.

The hardness of aluminum alloys used in the aerospace industry can be related to their electrical conductivity. Eddy-current conductivity meters are widely used to nondestructively monitor these alloys at various stages of aircraft production, in particular after critical heat treatment processes. The rapid measurement of electrical conductivity is one of the many checks made by coin vending machines, so the conductivity of alloys used in the manufacture of coins is often determined using an eddy-current conductivity meter to assure close quality control.

Flaws such as open cracks within materials cause eddy currents to detour around them and effectively enhance electric resistance; accordingly, eddy current testing can be used to detect flaws. Differences in thickness of foils and plating cause differences in the electric resistance; hence, eddy current testing is also convenient for estimating thickness. Stronger

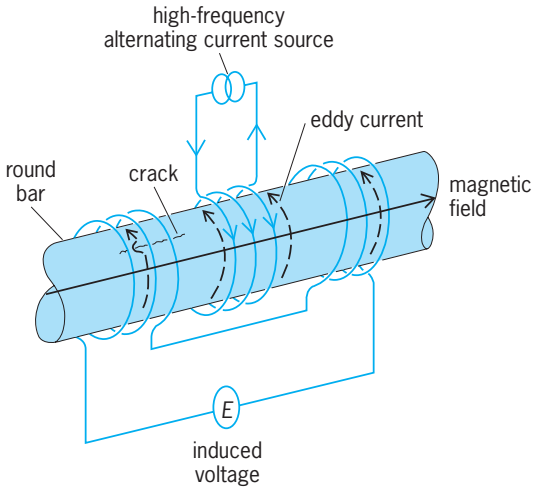


Fig. 2. The alignment of coils used to inspect a long bar.

magnetic fields are induced within ferromagnetic materials undergoing testing, so that the magnetic permeability also affects the eddy current. The evaluation of the hardness of ferritic steel through eddy current testing is based upon this effect of permeability, which is very susceptible to hardness. *See* FERROMAGNETISM; MAGNETIC MATERIALS; MAGNETISM.

Figure 2 shows the general alignment of coils to detect flaws within long bars. Annular coils encircle the round bar. The alternating current applied in the central coil generates a magnetic field inside the coil and within the bar in the direction of their axis, and thereby induces eddy currents along concentric circular paths within the bar. Simultaneously this eddy current produces a counter magnetic field which is superimposed on the primary one. The combined magnetic field gives rise to electromotive forces, that is, voltages between terminals of each coil. Two coils on either side of the central coil are wound in opposite directions, connected in series. The voltage between terminals is monitored continuously. When a flaw within a moving bar enters one of the two coils, the voltage reading corresponds to the difference between the voltages of the coils. This alignment permits the detection of flaws with high sensitivity, because it eliminates noises that originate from irregularities in diameter or physical properties of materials, whose variation is negligibly small between the closely spaced coils.

Since eddy current testing is of the noncontacting type, it can detect flaws within materials moving at high speed and even those raised to red heat in hot-rolling mills. Inspection of the latter is advantageous to exclude the influence of variation in permeability, because ferromagnetism vanishes at temperatures above the Curie point (1416°F or 769°C for iron). *See* CURIE TEMPERATURE; NONDESTRUCTIVE EVALUATION.

Kazuo Watanabe; A. E. Drake

Braking and damping. The principle outlined above, whereby a conductor which is moving through a nonuniform magnetic field experiences a retarding force, is used to brake or damp the oscillations of moving objects. In these applications, the field is that produced by a permanent magnet or an electromagnet supplied with direct current. The retarding force is proportional to the product of the flux produced by the magnet ϕ and the effective eddy current I produced by it. The effective eddy current, however, results from the rate of flux cutting, that is, the flux ϕ multiplied by the velocity of movement υ. The retarding force is thus proportional to $\phi^2\upsilon$. For a constant flux, which applies with a permanent magnet or electromagnet with constant current, the retarding force is proportional to the velocity of the moving object. With an electromagnet, the current can be varied, to adjust the ratio of retarding force to velocity. This ratio is also affected by the physical dimensions and materials used, for example, the resistivity of the moving object. Brakes are often fitted to rotating objects, and in these applications the retarding torque is proportional to the angular velocity ω.

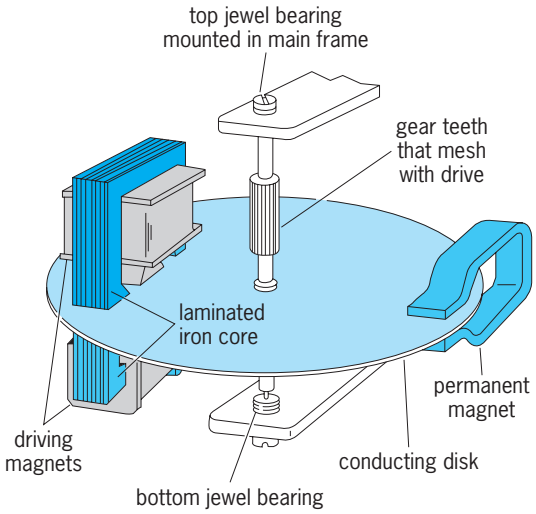


Fig. 3. Basic construction of electromechanical energy meter.

Eddy-current brakes are probably most widely used in the electromechanical energy meters which are installed in consumers' premises by electricity supply companies. In this application, the meters (**Fig. 3**) contain electromagnets which produce fluxes proprotional to the system voltage and current. These fluxes induce emf's and eddy currents in a conducting disk, which is mounted in jewel bearings so that it may rotate freely. The eddy currents react with the fluxes to produce a torque proportional to the power being supplied. Eddy-current braking, produced by a flux from a permanent magnet cutting the disk, provides a retarding torque proportional to the disk speed. Because the frictional torques are negligible, the disk runs at a speed at which the driving and retarding torques are almost equal, and as a result the speed is proportional to the power being supplied to the consumer. A geared drive to indicating dials meshes with gear teeth on the disk's spindle. *See* WATT-HOUR METER.

The same form of braking is also employed widely in the induction-type relays to provide time-graded protection. Eddy-current brakes incorporating electromagnets are used to provide controllable load torques on test equipment for small machines. *See* RELAY.

Eddy-current damping is used to suppress undesirable mechanical oscillations and is used widely in deflection-type indicating instruments. *See* DAMPING.

Arthur Wright

Bibliography. Atomic Energy of Canada, Ltd, *Eddy Current Testing*, 2 vols., 1987; G. Birnbaum and G. Free (eds.), *Eddy-current Characterisation of Materials and Structures*, ASTM Spec. Publ., no. 722, American Society for Testing and Materials, 1981; B. I. Bleaney and B. Bleaney, *Electricity and Magnetism*, 2 vols., 3d ed., 1976, paper, 1989; D. Hagemaier, *Fundamentals of Eddy Current Testing*, 1990; A. E. Drake and A. C. Lynch, AC conductivity standards for the calibration of eddy-current conductivity meters, *J. Phys. E*, 120:137–139, 1987;

H. L. Libby, *Introduction to Electromagnetic Nondestructive Test Methods*, 1971, reprint, 1979; E. M. Purcell, *Electricity and Magnetism*, 2d ed., 1985; S. S. Udpa and P. O. Moore (eds.), *Eddy Current Testing*, vol. 5 of *Nondestructive Testing Handbook*, 3d ed., American Society for Nondestructive Testing, 2004.

Edema

An abnormal accumulation of fluid in the cells, tissue spaces, or cavities of the body, also known as dropsy. An excess of fluid in the pleural spaces is referred to as hydrothorax, in the pericardial sac as hydropericardium, and in the peritoneal cavity as ascites. Anasarca is a generalized subcutaneous edema.

There are three main factors in the formation of generalized edema and a fourth which plays an important role in the formation of local edema. They are (1) permeability of the capillary wall, (2) colloid osmotic pressure of the plasma proteins, (3) hydrostatic pressure in the capillaries, and (4) lymphatic obstruction.

Permeability of capillary wall. Normally the capillary walls are freely permeable to water, salts, and dissolved gases but are almost impermeable to proteins. When the vessel wall is injured by toxins, anoxia, or paralytic dilatation, the capillary endothelium becomes permeable to proteins. With a diffusion of protein into the tissues, the plasma osmotic pressure is lowered and the osmotic pressure of the tissue is increased. Under these circumstances fluid collects in the tissue spaces.

Such a condition plays an important role in inflammatory edema. It is a factor in the edema of severe infections, metabolic intoxications, asphyxia, anaphylactic reactions, secondary shock, and acute nephritis. It also contributes to the edematous conditions when there is a fall in the level of plasma proteins.

Osmotic pressure of plasma proteins. A fall in plasma proteins tends to decrease the forces tending to reabsorb and hold fluid in the vascular compartment. Albumin is the protein of greatest importance in this regard. When the plasma protein level drops below 3 g/100 ml, the colloid osmotic pressure is no longer sufficient to maintain a balance with the hydrostatic pressure of the blood, which tends to drive fluid out into the tissue spaces. Hence more fluid goes out into the tissue spaces and remains there until a new equilibrium is reached.

This form of edema is seen in association with prolonged malnutrition (nutritional edema) as during a famine or with chronic nutritional or metabolic defects. With a marked loss of albumin in the urine as in nephrotic syndrome, there follows a lowering of the plasma albumin fraction and the development of edema.

In edema of kidney disease the protein content and specific gravity of the edema fluid are low. The blood reveals a markedly elevated cholesterol and a drop in total protein level with a relatively greater drop in the albumin fraction. As the plasma osmotic pressure drops, water passes into the tissues and with it crystalloids. These substances, especially sodium chloride, are retained in the tissues. Thus as water is taken in, it passes rapidly into the tissues and is not eliminated into the urine. *See* OSMOREGULATORY MECHANISMS.

Capillary hydrostatic pressure. Under normal conditions the hydrostatic pressure in the arterial end of the capillary is sufficient to overcome the plasma osmotic pressure and drives fluid out of the capillary into the tissue spaces. During passage through the capillary the pressure drops to a level low enough to allow the osmotic pressure of the proteins to draw fluid back into the vascular compartment. An increase in hydrostatic pressure at the venous end of the capillary will upset the balance, resulting in a decreased absorption of tissue fluid by the osmotic pressure of the plasma proteins. Under these circumstances an increased amount of fluid will be returned via the lymphatics, but as the condition progresses edema will develop. Such a situation can follow venous congestion of long duration. *See* CIRCULATION.

Cardiac edema following the generalized venous congestion of cardiac failure is the commonest form of this type of edema. The fluid which collects in the tissue spaces is affected by changes in position, being more marked in the dependent portions of the body. Fluid also collects in the serous cavities. The lymphatics empty into the venous system, and therefore a rise in pressure in the venous system results in an increased pressure within the lymphatics, which also contributes to the edema formation. As the capillary walls are distended, they become more permeable. In addition, a state of chronic hypoxia may exist, causing a further insult to the capillary endothelium with a loss of protein into the tissue spaces.

Other factors seem to play a role in cardiac edema. With decreased cardiac output there is reduced renal blood flow and glomerular filtration rate, with a consequent reduced excretion of salt and water. This may be responsible for an increased volume of extracellular fluid and plasma which, in turn, is followed by a rise in venous pressure.

Pulmonary edema (edema within the lung) is usually a form of cardiac edema but may be secondary to other factors such as inflammation. *See* RESPIRATORY SYSTEM DISORDERS.

Another condition resulting from an increased hydrostatic pressure in the capillaries is postural edema. This occurs when an individual has been standing motionless for a long period of time; the fluid collects in the subcutaneous tissues of the feet and ankles.

Cirrhosis of the liver causes an impediment to the flow of blood through the portal circulation. There is a consequent rise in venous pressure and ascites forms. *See* CIRRHOSIS.

Lymphatic obstruction. A portion of the intercellular tissue fluids returns to the circulation via the lymphatics. Obstruction to this channel will contribute to local edema. Infestation by filaria is one cause of lymphatic obstruction, particularly in the

tropics. Lymphatic channels may be destroyed or obstructed by surgical procedures, resulting in localized edema. Milroy's disease is a chronic hereditary edema thought to be due to lymphatic obstruction. *See* LYMPHATIC SYSTEM.

Management of edema. The management of patients with edema is directed toward the treatment of the underlying medical condition. Diuretics are often used, most successfully in cardiac edema. Romeo A. Vidone; Irwin Nash

Bibliography. A. C. Guyton and J. E. Hall, *Textbook of Medical Physiology*, 9th ed., 1996; J. M. Kissane (ed.), *Anderson's Pathology*, 9th ed., 1989; S. L. Robbins et al., *Pathologic Basis of Disease*, 6th ed., 1999; J. B. Wyngaarden and L. H. Smith (eds.), *Cecil Textbook of Medicine*, 17th ed., 1985.

Edentata

The former name of an order that included the Xenarthra (sloths, anteaters, armadillos), Pholidota (pangolins), and Tubulidentata (aardvarks). The order was so named because of the apparent lack of some or all teeth, although most of these species are not toothless but have at least vestigial teeth. It was subsequently realized that Edentata was polyphyletic—that it contained unrelated families and was thus invalid. It is now thought that any similarities between these groups is the result of similar adaptations to a common way of life and are not indicative of actual relationships. Thus, pangolins and aardvarks are now placed in separate orders, and the order Xenarthra was established to group the remaining families (which are all related). The term edentate is now used only as a term of convenience when referring to these unusual mammals; it has no taxonomic validity. *See* AARDVARK; ANTEATER; ARMADILLO; DENTITION; MAMMALIA; PHOLIDOTA; SLOTH; TOOTH; TUBULIDENTATA. Donald W. Linzey

Bibliography. R. M. Nowak, *Walker's Mammals of the World*, 6th ed., Johns Hopkins University Press, 1999.

Ediacaran biota

A widely distributed group of soft-bodied marine organisms that are preserved as fossils in rocks of latest Proterozoic age (600–543 million years ago, or Ma). The biota characterizes a geological period, known as the Ediacarian or the Vendian, which precedes the widespread appearance of animals with mineralized skeletons. The name Ediacara refers to an abandoned mining area about 380 mi (600 km) north of Adelaide, South Australia.

Discovery and distribution. Although Ediacaran fossils were described in the 1930s from southwest Africa (Namibia), it was the discovery in 1946 of abundant fossil "jellyfish" at Ediacara that sparked international interest in this biota. Subsequently, similar or identical fossils were found in central England (1957); southeastern Newfoundland (1967); northern Russia, the Ukraine, northern Siberia, and the Ural Mountains (1960s–1990s); North Carolina (1966); northwestern Canada (1979); southern Nevada (1997); and elsewhere. About 30 localities have been described, with the most diverse biotas found at Ediacara, Namibia, and on the coast of the White Sea in northern Russia. As some of these sites could not have been less than about 6000 mi (10,000 km) apart, no matter how the continents were arranged at the time, there is no doubt that the Ediacaran biota was a globally distributed marine biota. Ediacaran fossils have also been found in both shallow-water and deep-water facies. The biota appears to show a degree of biogeographic provinciality, but it has proved difficult to untangle the effects of geographic separation, paleoenvironmental differences, age differences, and sampling and preservational biases on the composition and preservation of the biota at each major site. A number of taxa are found in rocks formed under widely different conditions, and these, at least, may have been ecological generalists.

Age of assemblage. Recent radiometric dating has bracketed the Ediacaran biota between 600 and 543 Ma, with most sites between 565 and 543 Ma. Most known occurrences of Ediacaran organisms precede the earliest great radiation of skeletal fossils (archaeocyaths, trilobites, mollusks, brachiopods), or else they can be placed as latest Proterozoic on other evidence. Although unfossiliferous strata separate the Ediacaran biota from the earliest Cambrian fossils at many localities, Ediacaran biotas in Namibia and Nevada are found in rock sequences extending right up to the base of the Cambrian. Radiometric dates have confirmed that there was no significant time gap between the disappearance of the Ediacaran organisms and the Cambrian radiations. In fact, a few Ediacaran fossils have been found in Cambrian strata; the biota did not entirely die out before the Cambrian. *See* CAMBRIAN.

Fossil preservation. At almost all sites, the fossils are preserved as impressions in some kind of detrital sedimentary rock. Commonly, they are found on the bases of sandstone beds (South Australia), within sandstone beds (Namibia), or below volcanic ashes swept into deep water by wind and turbidity currents (Newfoundland). The organisms appear to have lived in continental shelf to slope environments and are normally preserved in sediments that were deposited under fairly quiet conditions below normal wave base. Organisms with resistant bodies are usually preserved as concave impressions in the casting medium, but other fossils appear to be the sediment fillings of either soft bodies or the cavities left by their decay. Most strata where Ediacaran fossils have been found also show evidence of extensive microbial growth that formed firm mats on the sediment surface. By stabilizing and firming the sediment, and possibly by speeding up its lithification, these mats are now thought to have been crucial to the fossilization of these soft-bodied organisms.

Taxonomic diversity. More than a hundred scientific names have been given to various kinds of Ediacaran fossils, but many of these are based on questionable material, or they are unnecessary synonyms for the same genus or species preserved in different ways. Nevertheless, at a conservative estimate there are probably 40–50 distinct genera or probable genera of Ediacaran organisms worldwide. Assigning these genera to higher taxa, however, has been controversial. Prior to 1985, many scientists who worked on the Ediacaran fossils attempted to relate them to the phyla of living invertebrates. A real problem is that few of these fossils can unequivocally be referred to living or extinct animal taxa. Because many of the fossils are simply circular structures with or without radial or concentric markings, they impart little information and are difficult to interpret. Some of these circular "medusoids" could be the remains of cnidarian jellyfish (Hydrozoa or Scyphozoa), but very few show diagnostic cnidarian characters such as fourfold symmetry, or even diagnostic metazoan characters such as definite guts. The recognition of many critical features is hampered by the nature of the preservation. In addition, many Ediacaran fossils, notably most forms from Namibia, are so unusual in shape that they cannot be placed firmly in any modern group. *See* CNIDARIA.

Many workers have placed the more unusual organisms in an extinct higher taxon of phylum grade, commonly named the Petalonamae. Others have used differences in symmetry and body organization to identify and characterize several major taxonomic groups regarded, in general, as extinct higher taxa of phylum or class grade. Beginning in the mid-1980s, much controversy and interest were sparked by proposals that all of the Ediacaran organisms were constructed on a single basic plan that was radically different from any other animal. According to this hypothesis, the Ediacarans were "quilted" organisms lacking heads, muscles, or digestive systems, and consisted of parallel sheetlike walls held together by regularly spaced internal partitions, and the whole organism was inflated by body fluids. Without digestive systems, these organisms would have lived by taking up dissolved nutrients or perhaps by photosynthesis. Proponents of this hypothesis classify the Ediacaran organisms in an extinct kingdom of multicellular life, the Vendobionta. Still others have proposed that the Ediacaran organisms belonged to extant kingdoms other than the animals; they were lichens, algae, or single-celled or colonial protists.

Controversy still reigns, but in all likelihood no hypothesis is entirely incorrect. Ediacaran fossils have been described which do show clear features diagnostic of living animal phyla. However, many of the Ediacaran organisms remain impossible to definitively link with living animals, and they may well represent extinct phyla or even kingdoms.

Distinctive genera such as the discoidal "worm" *Dickinsonia* (**Fig. 1**), the sac-shaped *Ernietta*, the paddle-shaped "sea pen" *Charnia*, and the trifoliate *Pteridinium* seem to form a coherent group of organisms with a Vendobionta-like "quilted" or "air-mattress" type of construction. Whether this common structural pattern is due to convergent evolution in response to particular environmental conditions remains uncertain, and so the Petalonamae or even the Vendobionta may be an unnatural group of unrelated taxa. There is, for example, evidence that "sea pens" such as *Charniodiscus* and *Rangea* (**Fig. 2**) lived attached to the ocean floor by

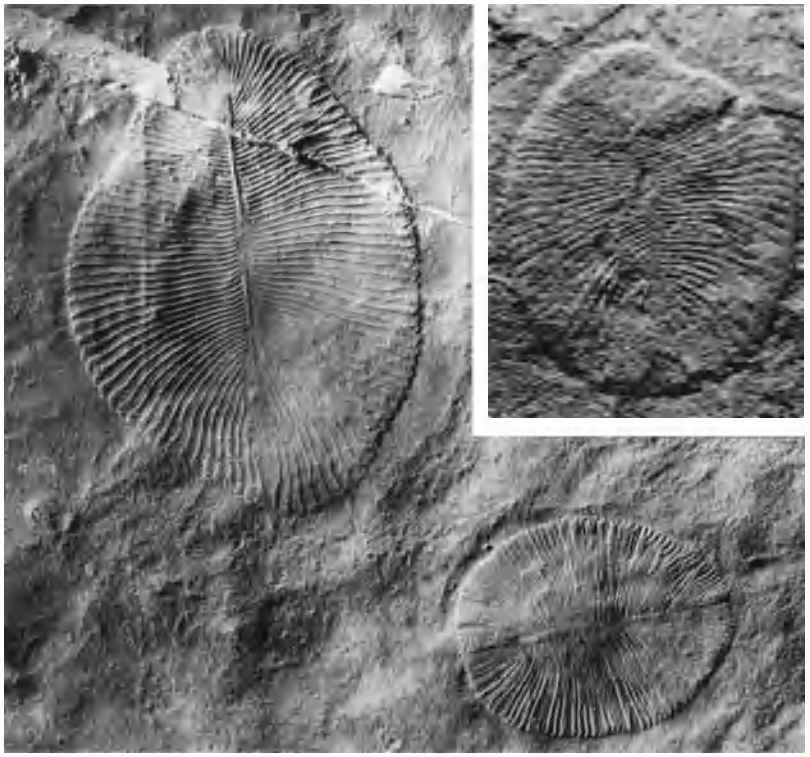


Fig. 1. Natural impressions of *Dickinsonia* from Australia and from Russia (inset). The Australian specimens, 5 and 3 in. (13 and 8 cm) in length, have about the same number of segments and may be regarded as the contracted and expanded states of a single individual.

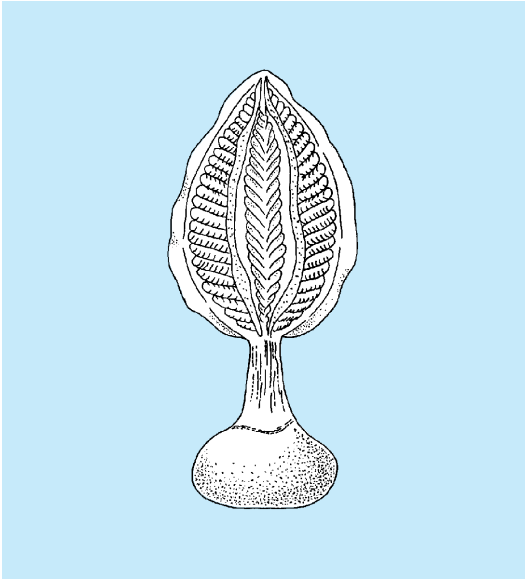


Fig. 2. Reconstruction of *Rangea*, a complex, sessile genus of "sea pens" from Namibia.

globular holdfasts, whereas *Ernietta* lived partially buried in sand with no holdfast. Unnamed spindle-shaped "sea pens" from Newfoundland also show no trace of an attachment structure. *Dickinsonia* seems to have been free-living and mobile and may have had a gut.

The members of a second group of Ediacaran animals are unified by their threefold symmetry. All have a circular outline and low conical shape reminiscent of modern jellyfish, and they have been regarded as an extinct class of the Cnidaria by some researchers. However, the morphological complexity of genera such as *Tribrachidium* and *Albumares* argues against such a view. It is plausible that the triradiate forms are members of a second extinct phylum, the Trilobozoa. True free-living cnidarians may be represented by "medusoids" such as *Ediacaria* and *Eoporpita*, which are now thought to be benthic polyplike organisms. A few forms, such as the tetraradial *Conomedusites* and the chondrophorine-like *Ovatoscutum* (**Fig. 3**), may have been planktonic cnidarians. The Australian *Palaeophragmodictyon* shows impressions of a spicule network and is almost certainly a sponge.

Although several higher invertebrate phyla such as the Annelida (segmented worms), Arthropoda, and Echinodermata have been identified from the Ediacaran biota, only a few of the known genera could be used to make a plausible case for the presence of these phyla. *Spriggina floundersi*, known from tens of specimens from Ediacara and nearby sites, had a long, bilaterally symmetrical, segmented body with a boomerang-shaped feature at one end. *Spriggina* has been placed with the annelids, but it is equally likely that it was a primitive soft-shelled arthropod. In this interpretation, the boomerang-shaped structure is considered to be homologous with the arthropod head. A few *Spriggina*-like fossils show traces of branching caecae within this head, closely resembling the anatomy of trilobites and other arthropods from Cambrian sites such as the Burgess Shale. Yet another fossil, *Kimberella quadrata*, appears to have borne a stiff shell on its dorsal side, which covered a soft muscular "foot" surrounded by numerous gill-like folds. *Kimberella* is likely to be closely related to the Mollusca, although it may not be closely related to any extant molluscan class.

Fig. 3. Discoidal fossil *Ovatoscutum* from South Australia may be a link between the Ediacaran fauna and younger biotas. It is probably an impression of the float of a chondrophorine cnidarian.

Fig. 4. Mold of *Cloudina* or a related genus, an annulated tubular fossil that was mineralized in life. *Cloudina* has been found together with soft-bodied Ediacaran fossils at several localities; this one is from Nevada.

Finally, a few organisms associated with the Ediacaran biota formed hard or even biomineralized parts or skeletons. Biomineralization did not absolutely originate in the Early Cambrian. The calcified tube *Cloudina* (**Fig. 4**) and the agglutinated tube *Archaeichnium* may have been formed by worms. Mineralized sponge spicules have been found in Ediacaran-age rocks in Mongolia. Other mineralized fossils are more difficult to decipher and include tubular, cup-shaped, conical, and hexagonal forms. Organic-walled tubes include the threadlike sabelliditids, possibly made by annelid or pogonophoran relatives; and the tetraradial *Corumbella*, which resembles the tubes made by certain scyphozoan polyps. Small, unmineralized toothlike fossils (*Redkinia*), which somewhat resemble annelid jaws, are known from Ediacaran assemblages in Russia. *See* ANNELIDA; ARTHROPODA; BURGESS SHALE; ECHINODERMATA.

Associated trace fossils. The marks left in soft sediments by otherwise unknown animals provide another source of knowledge of late Precambrian animal life. The figure-8-shaped trail in **Fig. 5** indicates that animals capable of directed, muscular, gliding motion (like that of a garden snail) coexisted with more typical members of the Ediacaran biota. In a similar fashion, strings of fecal pellets demonstrate the existence of animals with one-way guts, and closely meandering marks imply an ability for

Fig. 5. Ediacaran trace fossil *Gordia* found on the sole of a sandstone bed in South Australia gives clear evidence for the existence of mobile animals in the late Precambrian.

systematic grazing. There is little evidence for vertical burrowing in rocks of this age.

Significance. Although many Ediacaran fossils are enigmatic, there is sound evidence that sponges, cnidarians, bilaterian worms, and possibly arthropods and other phyla were present. This implies that the Animalia originated even farther back in time. The largest Ediacaran fossils, reaching up to 3 ft (1 m) in length, are flattened "fronds" that had large surfaces compared with their volumes. Proponents of the Vendobionta hypothesis have claimed that such large organisms, lacking guts or muscles, must have been photosynthetic or chemosynthetic and probably contained symbiotic, photosynthetic microorganisms. However, such a lifestyle is also found in modern reef corals and various other marine animals. Alternatively, the high ratio of surface to volume may represent adaptations to an atmosphere and hydrosphere relatively low in oxygen. *See* PALEONTOLOGY.

Bruce Runnegar; Ben Waggoner

Bibliography. J. G. Gehling, Microbial mats in terminal Proterozoic siliciclastics: Ediacaran death masks, *Palaios*, 14:40–57, 1999; J. P. Grotzinger et al., Bio-stratigraphic and geochronologic constraints on early animal evolution, *Science*, 270:598–604, 1995; G. M. Narbonne, The Ediacara biota: A terminal Neoproterozoic experiment in the evolution of life, *GSA Today*, 8(2):1–6, 1998; B. Runnegar, Vendobionta or Metazoa? Developments in understanding the Ediacara "fauna," *N. Jb. Geol. Paläontol. Abh.*, 195:303–318, 1995; A. Seilacher, Vendobionta and Psammocorallia: Lost constructions of Precambrian evolution, *J. Geol. Soc. London*, 149:607–613, 1992; R. C. Sprigg, Early Cambrian (?) jellyfishes from the Flinders Ranges, South Australia, *Trans. Roy. Soc. S. Aust.*, 71:212–224, 1947.

Eel

A fish in the order Anguilliformes, one of several orders of the superorder Elopomorpha, which share a leptocephalus (a slender, transparent larva with a long narrow head) [**Fig. 1**] larval development. The name eel also applies to other fishes similar in appearance to true eels, that is, having a serpentine body, lacking pectoral fins, and usually lacking scales. *See* ANGUILLIFORMES.

True Eels

The Anguilliformes consist of suborders Anguilloidei, Muraenoidei, and Congroidei.

Anguilloidei. This suborder comprises three families.

Freshwater eels (family Anguillidae). These eels (**Fig. 2**) are found in tropical and temperate seas except for the eastern Pacific and southern Atlantic; they consist of one genus and 15 species. They are catadromous (inhabiting freshwater and migrating to spawn in salt water), living practically all of their adult life in freshwater.

Mud eels (family Heterenchelyidae). These eels inhabit the tropical Atlantic, Mediterranean, and eastern Pacific; they consist of two genera and eight species. They typically burrow head-first into the substrate.

Spaghetti eels (family Moringuidae). These eels inhabit the tropical Indo-Pacific and western Atlantic; they consist of two genera and six species. Their bodies are extremely elongate and adapted for burrowing head-first into the substrate.

Muraenoidei. This suborder comprises three families.

False morays [family Chlopsidae (Xenocongridae)]. These eels inhabit the tropical and subtropical Atlantic, Indian, and Pacific oceans; they consist of eight genera and 18 species. Their gill openings are restricted to small rounded, laterally placed apertures; pectoral fins are absent in two of the eight genera; and lateral-line pores are restricted to the head.

Myroconger eels (family Myrocongridae). These eels inhabit the eastern tropical Atlantic and Pacific oceans; they consist of one genus and four species. Their gill

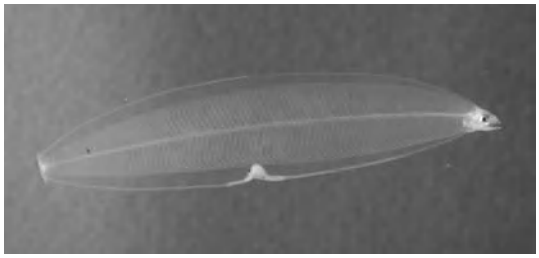


Fig. 1. Leptocephalus larva. (*Photo by Antonio J. Ferreira, © California Academy of Sciences*)

Fig. 2. *Anguilla anguilla*. (*Drawing by Robbie Cada, www.FishBase.org*)

openings are small but not greatly restricted; pectoral fins are present; their bodies are strongly compressed; and lateral-line pores are located in the branchial (pertaining to the gills) region above the pectoral fins.

Moray eels (family Muraenidae). These eels inhabit tropical and temperate seas of the world, with some species entering freshwater; they consist of 15 genera and about 185 species. Their gill openings are restricted to small rounded, laterally placed apertures; the pectoral fins are absent; and lateral-line pores are located on the head and branchial region. The giant moray (*Gymnothorax javanicus*) of the Indo-Pacific attains a length of 3 m (10 ft), making it one of the largest eels.

Congroidei. This suborder comprises nine families.

Cutthroat eels (family Synaphobranchidae). These eels inhabit the Atlantic, Indian, and Pacific oceans; they consist of 10 genera and 32 species. The gill openings are low, at or below the insertion of the pectoral fins; hence, the name cutthroat eels.

Snake eels and worm eels (family Ophichthidae). These eels inhabit tropical and warm temperate seas of the world, with some species in freshwater; they consist of 52 genera and about 290 species. The family differs from other eels in the unusual form of the branchiostegal (pertaining to the membrane covering the gills) rays, which broadly overlap those of the other side, leaving the branchiostegal region inflated or slightly bulbous. *Ophichthus rex*, known only in the Gulf of Mexico from Florida to Texas, and *O. ophis*, from the tropical and subtemperate western and eastern Atlantic, both with total lengths of about 210 cm (83 in.), are the largest ophichthids.

Shorttail eels (family Colocongridae). These eels inhabit the Atlantic, Indian, and western Pacific oceans; they consist of one genus and five species. Shorttail eels, with a blunt head and stubby body, are the least elongate of all the true eels.

Longneck eels (family Derichthyidae). These eels inhabit the Atlantic, Indian, and Pacific oceans; they consist of two genera and three species. These are mesopelagic [inhabiting the middle ocean region, from about 200 to 1000 m (650 to 3300 ft)] to bathypelagic [inhabiting the deep ocean region, from about 1000 to 4000 m (3300 to 13,000 ft)] eels, with a maximum length of about 60 cm (24 in.).

Pike congers (family Muraenesocidae). These eels inhabit the tropical Atlantic, Indian, and Pacific oceans; they consist of possibly six genera and 13 species. Pike congers have a large mouth equipped with large conspicuous teeth; their large eyes are covered with skin; and they have a conspicuous lateral line. Five species attain a total length of 200 cm (78 in.) or more, with the largest known being 250 cm (98 in.).

Snipe eels (family Nemichthyidae). These eels are mesopelagic in the tropical and temperate Atlantic, Indian, and Pacific oceans; they consist of three genera and nine species. Snipe eels are the most bizarre anguilliforms. Their bodies are extremely long and ribbonlike; one genus, *Nemichthys*, has over 750 vertebrae. The jaws are very long and narrow, and the teeth cannot be occluded (that is, the teeth do not come into contact with cusps of the opposing teeth fitting together), except in sexually mature males, which undergo an extreme shortening of the jaws and loss of teeth. The two sexes are so different that at one time some species of snipe eels were thought to be different taxa, even at the suborder level.

Conger eels (family Congridae). These eels inhabit the tropical to temperate Atlantic, Indian, and Pacific oceans; they consist of 32 genera and 160 species. Most congrid eels are dull colored, usually tan or brown on the back and silvery on the sides and belly, with only a few having spots or stripes; the lateral line is complete; the dorsal and anal fins are well developed and continuous with the caudal fin; and the pectoral fins are usually well developed, except in garden eels, in which the pectorals are reduced to a tiny fleshy flap. Garden eels live in colonies, where they partly submerge themselves in tubelike burrows. With their head bent facing the current, they sway back and forth picking plankton as it drifts by. Congrid eels vary greatly in size—the smallest, *Bathycongrus odontostomus*, of the Indo-West Pacific, is only 4.5 cm (1.8 in.) in total length, whereas *Conger conger* of the eastern Atlantic attains a total length of 300 cm (118 in.). The longest true eel in American waters is *C. oceanicus*, with a total length of 230 cm (90 in.).

Duckbill eels (family Nettastomatidae). These eels inhabit the Atlantic, Indian, and Pacific oceans; they consist of six genera and about 38 species. They have elongate heads, large mouths, and flattened snouts—hence the name duckbill; their tails are attenuated; and pectoral fins are usually absent in adults. Their maximum length is about 100 cm (39 in.).

Sawtooth eels (family Serrivomeridae). These eels are pelagic (living in the open ocean) in the tropical and temperate Atlantic, Indian, and Pacific oceans; they consist of two genera and about 10 species. Their jaws are extremely elongate and slender, but less so than in snipe eels; their gill openings are connected ventrally; the tails are attenuated; and the vomerine teeth (that is, teeth on the vomer, which is the median unpaired bone on the roof of the mouth) are lancet-like and arranged in a sawlike row. Their maximum total length is about 78 cm (31 in.).

Other "Eels"

Other "eels" are classified in orders Saccopharyngiformes, Gymnotiformes, and Siluriformes.

Order Saccopharyngiformes. This order comprises two families.

Bobtail snipe eel (family Cyematidae). These fish are bathypelagic in the Atlantic, Indian, and Pacific oceans; they consist of two monotypic genera. Their bodies are relatively short and compressed; they lack lateral-line pores; their eyes are small to vestigial; the tip of the tail is blunt; the caudal fin is present; and maxillae are present. The maximum length is about 15 cm (6 in.).

Pelican eels (family Eurypharyngidae). These eels inhabit the tropical and temperate Atlantic, Indian, and Pacific oceans; they consist of one species. They have a huge head, a pelicanlike mouth, jaws with numerous minute teeth, and minute pectoral fins. This fish

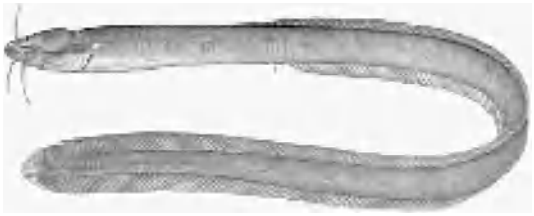


Fig. 3. Eel catfish (*Channallabes apus*). (*Drawing by P. Mertens, www.FishBase.org*)

is unique among teleosts in having five gill arches. The maximum length is about 74 cm (29 in.).

Order Gymnotiformes. This order comprises a single family.

Electric eel (family Gymnotidae). Another true-eel mimic is the electric eel (*Electrophorus electricus*), a member of the family Gymnotidae (nakedback knifefishes). It occurs in the Orinoco and Amazon rivers of northern South America and is unique among the gymnotiforms in having electric organs capable of delivering discharges up to 600 volts. The electric eel is further identified by the lack of a dorsal fin, while displaying a very long anal fin terminating at the tip of the tail. It has an oral respiratory organ for breathing air and a rounded body that lacks scales, and can continuously add vertebrae throughout life. It attains a total length of 220 cm (87 in.). *See* ELECTRIC ORGAN (BIOLOGY).

Order Siluriformes. A catfish that could easily be mistaken for an eel inhabits the Congo River Basin of Africa. It is called the eel catfish (*Channallabes apus*) [**Fig. 3**] and is the only one among about 2870 species of catfishes that bears the name eel. The eel catfish, belonging to the siluriform family Clariidae (air-breathing catfishes), has reduced eyes, lacks pectoral and pelvis fins, has long dorsal and anal fins, and has the habit of burrowing in the substrate, all characteristics of many species of true eels. *See* SILURIFORMES.

Herbert Boschung

Bibliography. 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.

Effective dose 50

This term is used chiefly to characterize the potency of a drug by the amount required to produce a response in 50% of the subjects to whom the drug is given. The term is also known as ED_{50} or median effective dose. At one time it was usual to try to measure the effect of a drug by noting the amount which was just sufficient to produce a particular response; but when it was realized that this amount varied greatly from subject to subject, attention was turned to measuring the effect on a group of subjects. Suppose, for example, that a drug is being used to relieve a certain type of pain; then the median effective dose is of such a size that it controls the pain in 50% of the sufferers and is insufficient to control it in the remaining 50%.

The term median effective dose is most commonly applied in connection with drugs, but it may be used when various other sources of stimuli, for example, x-rays, are under consideration. The response must be of the kind known as quantal, or all or nothing, where the investigator is simply able to report that the response either was or was not elicited; for example, convulsions did or did not occur, hemorrhage was or was not produced, pregnancy did or did not ensue, the animals did or did not survive.

The median effective dose is not a well-defined quantity until the test animal, the end response, and such factors as the route of injection of a drug and the state of nutrition of the animal are specified. In the work of an investigator who uses the appropriate controlled conditions, however, the ED_{50} is a reproducible measurement.

Determination of the ED_{50} of a drug requires the administration of at least two separate amounts of the drug, each one given to several subjects. Suppose that an animal physiologist wishes to measure the ED_{50} of a hormone, estrone, that causes estrus, or heat. He has specified that he will use spayed female rats as experimental animals, and has also laid down various conditions that he regards as important in controlling the results of his investigation: the age of the animals, perhaps, and the route by which the hormone will be administered. It is unlikely that a reliable measurement could be made with fewer than 30 animals. Suppose that a total of 60 were used, 20 at each of 3 doses, with the following results:

Dose, international units	*Animals showing estrus, %*
2	20
3	45
4	80

Inspection of the results suggests that a dose somewhat greater than 3 units will cause estrus in 50% of the animals. Methods based on the theory of probability can be used to give a more objective analysis of the data and to provide a measure of the error of the estimate, but it is sufficient to note for the present purpose that the ED_{50} is approximately 3. It is clear that in measuring the ED_{50} some choices of the dosage to be used will work out better than others. Doses giving a response in a very small or very large percentage of the subjects contribute little information to the measurement of ED_{50}. The ideal is to have doses on either side of the ED_{50}, and the closer they are to ED_{50}, the better. However, in practice a pharmacologist sometimes determines an ED_{50} as one part of a larger experiment to investigate the dose response curve of a drug, or to compare the potency of two preparations of a drug. In such a case, though he would still want to have doses on each side of the ED_{50}, he would not want them close to the ED_{50}. More information about the total curve would be given if the low dose had an effect in, say, 25% of the animals and the high dose in about 75%. *See* BIOASSAY.

Colin White

Bibliography. C. Janney and J. Timpke, *Calculation of Drug Dosages*, 4th ed., 1993; R. Wiederhold, *Dosages and Calculations*, 1991.

Effector systems

Those organ systems of the animal body which mediate overt behavior. Injury to an effector system leads to loss or to subnormal execution of behavior patterns mediated by the system, conditions termed paralysis and paresis, respectively.

Overt behavior consists of either movement or secretion. Movement results from contraction of muscle. Secretion is a function of glands. Neither muscular contraction nor glandular secretion is autonomous but is regulated by an activating mechanism which may be either neural or humoral. In neurally activated systems the effector organ, whether muscle or gland, is supplied by nerve fibers originating from cell bodies situated in the central nervous system or in peripherally located aggregates of nerve cell bodies known as ganglia. The nerve fibers make intimate contact with, but are not protoplasmically continuous with, the cells of the effector organ. Activation of the effector organ occurs when the nerve cell body is excited and generates a nerve impulse, an electrochemical alteration which is conducted along the nerve fiber to its terminations on the effector organ cells. Here the nerve impulse releases from the nerve ending a chemical transmitter that generates a similar electrochemical alteration in the effector cells, and this alteration, in turn, leads to either contraction or secretion. Such systems composed of a muscle or a gland along with their regulating nerves are termed neuromuscular or neuroglandular effector systems, respectively. Examples are the skeletal muscles together with their motor nerve supplies, and the medullae of the adrenal glands along with their innervation (splanchnic nerves). In many neurally regulated effector systems (for example, skeletal muscle and adrenal medulla), function is totally dependent on intact innervation, and denervation leads to functional paralysis. In other organs (for example, the salivary glands), denervation causes only temporary paralysis. When recovery occurs, the gland may oversecrete continuously (paralytic secretion), apparently because the denervated gland cells become unusually sensitive to certain blood-borne chemical agents (denervation supersensitivity). *See* BIOPOTENTIALS AND IONIC CURRENTS; ENDOCRINE MECHANISMS; MOTOR SYSTEMS.

In other effector systems (humeromuscular and humeroglandular), the activating agent is normally a blood-borne chemical substance produced in an organ distant from the effector organ. Uterine smooth muscle is uninfluenced by the uterine nerve activity but contracts vigorously when the blood contains oxytocin, a chemical substance elaborated by the posterior lobe of the hypophysis; sensitivity to oxytocin increases progressively during pregnancy. Similarly, secretion of pancreatic juice is independent of pancreatic innervation; the regulating agents are blood-borne substances (cholecystekinin and secretin) produced by cells in the wall of the small intestine. Generally, in such humorally regulated effector systems, activation is more delayed and more prolonged than in neurally regulated systems. *See* PANCREAS; UTERUS.

Finally, some effector systems are hybrid in the sense that both nerves and humors regulate their functions. The smooth muscles of arterioles contracts in response to either nerve stimulation or epinephrine, a substance secreted into the bloodstream by the adrenal medulla. Secretion of hydrochloric acid by the gastric mucosa is increased by activation of the vagus nerve or by the presence in the blood of histamine, a substance found in many tissues of the body. Effector systems with both neural and humoral regulation are never completely paralyzed by denervation but may be deficient in reaction patterns when the quick integrated activation provided by neural regulation is essential. For example, following extensive vascular denervation, humoral agents may maintain sufficient arteriolar constricton to sustain the blood pressure in static situations. However, the normal capacity to increase arteriolar constriction to offset the gravitational effects of rising from the prone to standing positions is permanently lost, with the result that such postural changes may lead to inadequate cerebral blood flow and consequent fainting. *See* HISTAMINE; NERVOUS SYSTEM (VERTEBRATE); PSYCHOLOGY.

Theodore C. Ruch; Harry D. Patton

Efficiency

The ratio, expressed as a percentage, of the output to the input of power (energy or work per unit time). As is common in engineering, this concept is defined precisely and made measurable. Thus, a gear transmission is 97% efficient when the useful energy output is 97% of the input, the other 3% being lost as heat due to friction. A boiler is 75% efficient when its product (steam) contains 75% of the heat theoretically contained in the fuel consumed. All automobile engines have low efficiency (below 30%) because of the total energy content of fuel converted to heat; only a portion provides motive power, while a substantial amount is lost in radiator and car exhaust. *See* AUTOMOTIVE ENGINE; BOILER; POWER.

In such simple cases the value is clear. However, in some others it can be difficult to calculate exactly. For example, the efficiency of the process for converting corn to alcohol for use as automobile fuel can be computed as the ratio of the heat value of the alcohol to the heat value of all the energy used to produce it, including the fuel for the tractor to plow and harvest the cornfield and even the fuel used to create the steel and fabricate the tractor. The question is then, how much to include in the overall efficiency determination. *See* SIMPLE MACHINE.

F. R. E. Crossley

Efflorescence

The spontaneous loss of water (as vapor) from hydrated crystalline solids. The thermodynamic requirement for efflorescence is that the partial

pressure of water vapor at the surface of the solid (its dissociation pressure) exceed the partial pressure of water vapor in the air.

A typical efflorescent substance is Glauber's salt, $Na_2SO_4 \cdot 10H_2O$. At 25°C (77°F) the dissociation pressure for the process in reaction (1) is 2586.4 pas-

$$Na_2SO_4 \cdot 10H_2O(s) \rightarrow Na_2SO_4(s) + 10H_2O(g) \quad (1)$$

cals (0.776 in. Hg), 81% of the saturation vapor pressure of pure water at this temperature. In a sufficiently humid atmosphere Glauber's salt also can deliquesce by the process shown in reaction (2).

$$Na_2SO_4 \cdot 10H_2O(s) + H_2O(g) \rightarrow Na_2SO_4 \text{ (saturated aqueous solution)} \quad (2)$$

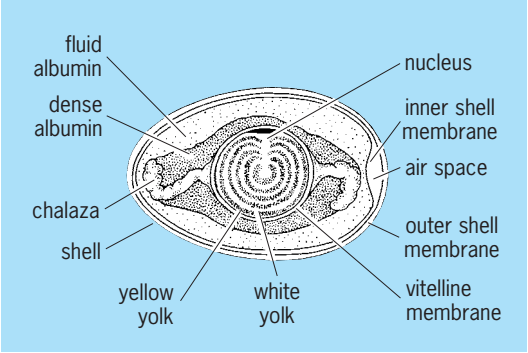
The vapor pressure of the saturated solution is 2919.27 Pa (0.876 in. Hg), 92% of the vapor pressure of pure water. Thus, Glauber's salt at 25°C (77°F) is stable in atmospheres having relative humidities of 81–92%; below 81% it effloresces; above 92% it deliquesces.

The spontaneous loss of water normally requires that the crystal structure be rearranged, and consequently, efflorescent salts usually go to microcrystalline powders when they lose their water of hydration. *See* DELIQUESCENCE; PHASE EQUILIBRIUM; VAPOR PRESSURE.

Robert L. Scott

Egg (fowl)

A single, large, living, female sex cell enclosed in a porous, calcareous shell through which gases may pass. Although they vary in size, shape, and color, the eggs of chickens, ducks, geese, and turkeys are essentially the same in structure and content (see **illus.**). Inward from the shell are the outer and inner shell membranes which are also permeable to gases. The membranes are constructed to prevent rapid evaporation of moisture from the egg but to allow free entry of oxygen, which is necessary for life. Air begins to penetrate the shell soon after the egg is laid, and it tends to accumulate in a space between the two membranes at the large end of the egg. *See* CELL (BIOLOGY).



Egg of a bird. (*After L. P. Sayles, ed., Biology of the Vertebrates, 3d ed., Macmillan, 1949*)

The inner shell membrane surrounds a mass of fluid albumin which, in turn, encloses a body of dense albumin; these two types of protoplasm constitute the so-called egg white. The central part of the egg is occupied by the yolk, which contains the vital egg nucleus and its associated parts. The yolk consists of alternating layers of yellow and white yolk. The yolk, enclosed by the vitelline membrane, is held in place by the chalaza which is anchored at each end of the egg and prevents undue mechanical disturbance. *See* CELL NUCLEUS; YOLK SAC.

When a sperm nucleus fuses with the egg nucleus, the process is called fertilization. Within a few hours or less, the fertilized egg begins a series of cell divisions and differentiations which result in the formation of the embryo. The embryo then undergoes further cell modification and eventually develops into the young of the species. In time the pressure created by the growth of the embryo causes the shell to rupture, and the young hatches. As the young emerges from the shell, it carries with it a part of the food and water originally in the yolk on which it can subsist for a few days. After this initial period, the young must have access to food and water. *See* CELL DIVISION; FERTILIZATION (ANIMAL); MITOSIS.

For a discussion of the steps involved in the development of various kinds of fertilized eggs into mature embryos *see* EMBRYOLOGY; OVUM; REPRODUCTION (ANIMAL); POULTRY PRODUCTION.

James F. Ferry

Egg water balance. The water vapor content of the environment is usually less than that found inside eggs, so water vapor is lost continually throughout incubation. Eggshell conductance, that is, the inverse resistance of the shell to gas movement, is directly related to egg mass, and inversely related to incubation time. The relationship between eggshell conductance and daily water loss, egg mass, and incubation time predicts that all bird eggs should lose about 15% of their initial mass as water vapor during incubation and that the vapor pressure gradient across the shell is similar for all bird species and relatively constant throughout incubation. The fractional water loss is apparently obligatory since eggs prevented from losing water show increased embryonic mortality and deformity. The obligatory nature of egg water loss may be related to the observation that the embryo inflates its lung just prior to hatching with the gas in the air space of the egg. The volume of the air space is related to the quantity of water loss since the egg shell is rigid.

Respiratory gas exchange. While the daily water loss from avian eggs during incubation appears to be relatively constant, such is not the case for the daily oxygen consumption (O_2 flux) and carbon dioxide production (CO_2 flux). The metabolic activity increases throughout incubation as the embryo develops and increases in mass, so both the O_2 and CO_2 exchange must similarly increase. Since the gas conductance of the shell is constant (all gases diffuse through the same pores), the pressure of O_2 inside the egg must fall and the pressure of CO_2 must rise if the embryo is to gain O_2 and lose metabolically

produced CO_2. The egg O_2 exchange of a variety of bird species measured just prior to hatching appears to increase in proportion to egg mass in much the same way that the eggshell gas conductance (as measured with water vapor) increases with increase in egg mass. These relationships allow one to predict that the gas pressures of O_2 and CO_2 (in the air space) are the same for all bird embryos just prior to hatching and are in fact similar to values one might expect to see in adult bird lungs. It would appear that one function of the eggshell is to prepare the embryo for transition to lung breathing at birth. *See* RESPIRATION.

Ralph A. Ackerman

Bibliography. R. W. Burley and D. V. Vadehra, *The Avian Egg: Chemistry and Biology*, 1989; B. M. Carlson, *Patten's Foundations of Embryology*, 6th ed., 2000.

Eggplant

A warm-season vegetable (*Solanum melongena*) of Asiatic orgin belonging to the plant order Polemoniales (formerly Tubiflorales). Eggplant is grown for its usually egg-shaped flesh fruit (see **illus.**). and is

Eggplant (*Solanum melongena*), cultivar Black Magic. (*Joseph Harris Co., Rochester, New York*)

eaten as a cooked vegetable. Cultural practices are similar to those used for tomatoes and peppers; however, eggplant is more sensitive to low temperatures. Popular purple-fruited varieties (cultivars) are Black Beauty and a number of hybrid varieties; fruits of other colors, including white, brown, yellow, and green, are used chiefly for ornamental purposes. Harvesting generally begins 70–80 days after planting. Florida and New Jersey are important eggplant-producing states. *See* PEPPER; SOLANALES; TOMATO.

H. John Carew

Ehrlichiosis

Ehrlichiosis is a tick-borne infection that often is asymptomatic but also can produce an illness ranging from a few mild symptoms to an overwhelming multisystem disease. Ehrlichiosis is included with those infections that are said to be emerging, either because they have been recognized only recently or because they were previously well known but now are occurring more frequently.

History. Prior to 1986, the bacteria of the genus *Ehrlichia* were of interest primarily to veterinarians. These microorganisms have a global distribution and are capable of producing a hemorrhagic disease in dogs and febrile illnesses of horses, cattle, sheep, and bison. A constant feature in all these circumstances is the predilection of the *Ehrlichia* to enter the cellular elements of the blood. The specific *Ehrlichia* species determined whether mononuclear cells, granulocytes, or platelets were invaded. Human illness associated with these organisms was first recognized in southern Japan in the 1950s. Because these illnesses seemed to be a local phenomenon, they received little attention elsewhere.

However, in 1986 a 51-year-old man, after planting some trees while vacationing in Arkansas, discovered some ticks on his neck and removed them uneventfully. Ten days later, he developed fever accompanied by a serious systemic illness that eluded specific diagnosis and required a prolonged hospitalization before his recovery. Careful microscopic examination of a slide preparation of his peripheral blood revealed distinctive inclusion bodies in his lymphocytes and monocytes. When reviewed by microbiologists and pathologists, these inclusions were identified as ehrlichiae. This illness was the first case of ehrlichiosis diagnosed in the United States. To date, over 400 cases of ehrlichiosis have been diagnosed in 30 states, and ehrlichiosis now is considered to be the most common tick-borne infection in this country.

Epidemiology. Human ehrlichiosis is caused by two distinct species: *E. chaffeensis* and an unnamed ehrlichial species. *Ehrlichia chaffeensis* infects primarily mononuclear blood cells; the disease produced by this species is referred to as human monocytic ehrlichiosis. The other ehrlichial species invades granulocytic blood cells, causing human granulocytic ehrlichiosis. The latter organism closely resembles *E. equi*, a species that infects horses.

Both of these ehrlichia species are transmitted to humans by the bite of infected ticks; however, the two species appear to have somewhat different geographic distributions. *Ehrlichia chaffeensis* occurs most commonly in the south-central and southeastern states, where it is associated primarily with the Lone Star tick (*Amblyomma americanum*); it is also transmitted by the common dog tick (*Dermacentor variabilis*). The agent of human granulocytic ehrlichiosis is found in the upper midwestern states of Wisconsin and Minnesota, as well as in several northeastern states. This agent seems to be transmitted principally by the deer tick (*Ixodes scapularis*).

Although ticks (the vector) are the mode of transmission of ehrlichial infections to humans, the ticks must acquire the ehrlichial organisms from animal sources (the reservoir hosts). The animal reservoirs for these organisms have not yet been confirmed, but white-tailed deer are a likely reservoir host of *E. chaffeensis* in the southeastern United States. *See* IXODIDES.

Clinical signs and pathogenesis. The forms of the disease caused by the two ehrlichial species are indistinguishable. Illness occurs most often during April–September, corresponding to the period when ticks are most active and humans are pursuing outdoor activities. Most individuals relate a history of tick exposure and have been engaged in work or recreation in wooded areas where they encountered ticks.

While obtaining its blood meal by biting an individual, the infected tick inoculates ehrlichiae into the skin. The microorganisms then spread throughout the body via the bloodstream. *Ehrlichia* must achieve an intracellular residence in order to survive and multiply. *Ehrlichia chaffeensis* seeks out tissue macrophages in the spleen, liver, and lymph nodes. It is suspected that the agent of human granulocytic ehrlichiosis has a tropism for granulocytic precursor cells in the bone marrow. Once the organism has entered the cell, it multiplies there for a period of time without producing symptoms (the incubation period). In ehrlichiosis, the incubation period can last from 1 to 3 weeks after exposure to the infected tick. Thereafter, individuals develop fever, chills, headache, and muscle pains. Gastrointestinal symptoms such as nausea, vomiting, and loss of appetite also are common. Laboratory abnormalities regularly include anemia, low white blood cell and platelet counts, and abnormal liver function. More severely ill individuals also may manifest abnormalities of the central nervous system, lungs, and kidneys. Elderly patients are more likely to become seriously ill than children and young adults. Fatality rates up to 4% have been documented. Because the clinical presentation is nonspecific, the diagnosis of ehrlichiosis may not be immediately apparent. Prolonged intervals between the onset of illness and the administration of appropriate therapy can lead to more severe disease symptoms and a greater risk of fatality. *See* CLINICAL MICROBIOLOGY; HEMORRHAGE; INFECTION.

Diagnosis and treatment. An important clue to the diagnosis of human granulocytic ehrlichiosis is the recognition of cytoplasmic vacuoles filled with ehrlichiae (morulae) in circulating neutrophils. Careful examination of stained smears of peripheral blood often yields such findings in human granulocytic ehrlichiosis. In contrast, similar detection of infection in circulating monocytes is a distinctly rare event in disease caused by *E. chaffeensis.* Laboratory diagnosis usually is made by detecting an increase in species-specific antibodies in serum specimens obtained during the acute and convalescent phases of the illness. However, such serologic testing is of no use in establishing the diagnosis before treatment is initiated. In an attempt to shorten this diagnostic lag, some research laboratories are offering to test blood specimens by using polymerase chain reaction (PCR) technology. Unfortunately, this method is not yet widely available. Therefore, therapy must be initiated on clinical suspicion.

Ehrlichiosis closely resembles another tick-borne illness, Rocky Mountain spotted fever, except that the rash, characteristic of spotted fever, is usually absent or modest. Hence, ehrlichiosis has been referred to as spotless fever. Fortunately, both diseases can be treated with tetracycline antibiotics. Most individuals respond to tetracycline therapy within 48–72 h. *See* RICKETTSIOSES.

Prevention. The avoidance of tick bites is fundamental to preventing ehrlichiosis. Sometimes this can best be accomplished by staying out of tick-infested areas. It has been documented that the regular use of insect repellant by individuals who spend time in grassy or wooded areas offers substantial protection against tick bites. It also is advisable to wear long-sleeved shirts and trousers that can be tucked into boots. Individuals should examine themselves for ticks periodically. Ticks attached to the skin can be removed with tweezers by grasping the tick as close to the point of attachment as possible and pulling gently and steadily.

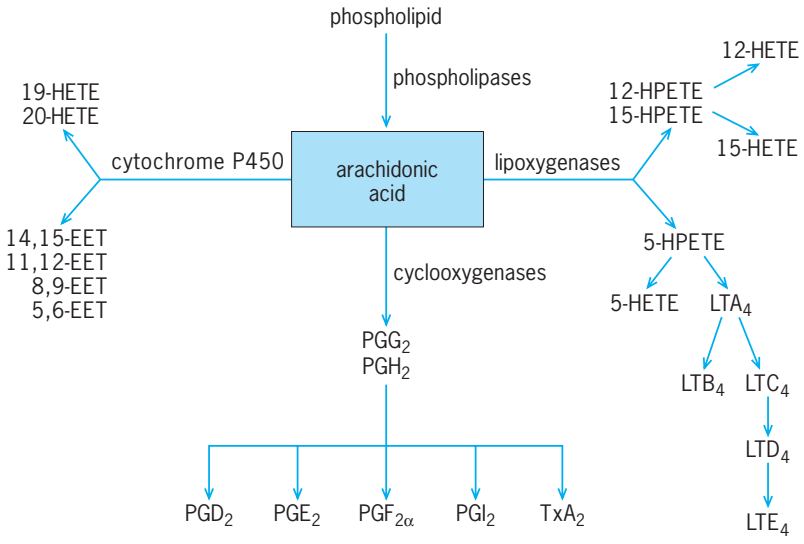
Steven M. Standaert; William Schaffner

Bibliography. J. E. Dawson et al., The interface between research and the diagnosis of an emerging tick-borne disease, human ehrlichiosis due to *Ehrlichia chaffeensis*, *Arch. Int. Med.*, 156:137–142, 1996; J. S. Dumler and J. S. Bakken, Ehrlichial diseases of humans: Emerging tick-borne infections, *Clin. Infect. Dis.*, 20:1102–1110, 1995; S. M. Standaert et al., Ehrlichiosis in a golf-oriented retirement community, *N. Engl. J. Med.*, 333:420–425, 1995; D. H. Walker and J. S. Dumler, Emergence of the ehrlichioses as human health problems, *Emerging Infect. Dis.*, 2:18–29, 1996.

Eicosanoids

A family of naturally occurring, biologically active substances derived from 20 carbon polyunsaturated fatty acids such as arachidonic acid. Members of the family include prostaglandins, thromboxanes, leukotrienes, and epoxyeicosatrienoic acids (EETs). Functioning as local hormones, they are synthesized and exert their biological actions in the same tissue. Many have physiological and pathological effects on the cardiovascular, pulmonary, reproductive, and digestive systems.

Biosynthesis and nomenclature. Eicosanoids are formed from polyunsaturated 20-carbon essential fatty acids. These fatty acids include 8,11,14-eicosatrienoic acid, 5,8,11,14-eicosatetraenoic acid (arachidonic acid), and 5,8,11,14,17-eicosapentaenoic acid. Arachidonic acid is the most abundant precursor of the eicosanoids in humans. It is esterified to phospholipids and other complex lipids in the cell membrane; thus, the cellular concentration of free arachidonic acid is low.



Enzymology of arachidonic acid metabolism. Structures and abbreviations for the compounds are given in Table 1.

Hormones such as histamine, bradykinin, and angiotensin II or physical or chemical stimuli increase the biosynthesis of eicosanoids by releasing arachidonic acid from the membrane lipids and making it available to the eicosanoid-synthesizing enzymes (see **illus.**). These stimuli activate phospholipase A_2 or phospholipase C either directly or by increasing intracellular calcium. Phospholipase A_2 hydrolyzes arachidonic acid from phospholipids, usually phosphatidylcholine or phosphatidylethanolamine. Phospholipase C converts phosphatidylinositol to a 1,2-diglyceride. The sequential action of diglyceride lipase and monoglyceride lipase on the diglyceride releases arachidonic acid. The free arachidonic acid is reesterified into cellular lipids or metabolized by cyclooxygenases, lipoxygenases, or cytochrome P450 monooxygenases. These enzymes add molecular oxygen or oxygens to the fatty acid.

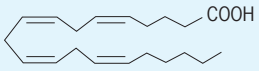
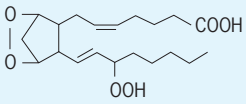
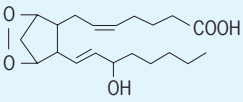
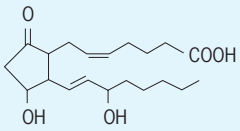
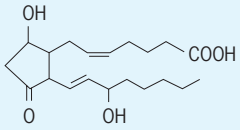
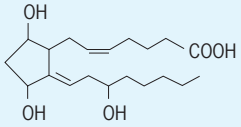
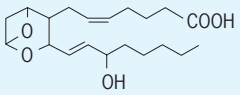
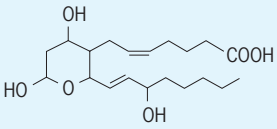
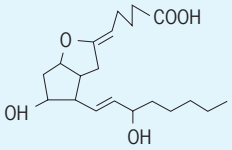
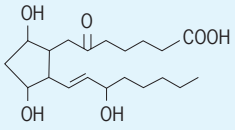
Metabolites of cyclooxygenases. The cyclooxygenases, also called prostaglandin endoperoxide synthases, metabolize unsaturated fatty acids to the prostaglandins and thromboxanes. There are two forms of the enzyme. Cyclooxygenase-1 is a constitutive enzyme found in most cells, whereas cyclooxygenase-2 is an inducible enzyme that becomes elevated in response to mediators of inflammation such as cytokines. Both enzymes perform the same two enzymatic reactions: endoperoxide synthesis and peroxide reduction. The endoperoxide synthase activity oxygenates and cyclizes the arachidonic acid, producing the cyclic endoperoxide prostaglandin G_2 (PGG_2) [**Table 1**]. The peroxidase activity reduces PGG_2 to PGH_2. Both PGG_2 and PGH_2 are unstable intermediates that are further metabolized by isomerases to PGE_2 and PGD_2, by prostacyclin synthase to PGI_2, and by thromboxane synthase to TXA_2. While most cells contain cyclooxygenase, cells vary in their complement of endoperoxide-metabolizing enzymes. As a result, cells make different prostaglandins. For example, vascular endothelial cells contain prostacyclin synthase and not thromboxane synthase. As a result, they synthesize PGI_2 but not TXA_2. In contrast, platelets contain only thromboxane synthase and thus make predominantly TXA_2. This selective synthesis regulates the prostaglandins produced and limits their sites of action. The cyclooxygenases are inhibited by nonsteroidal antiinflammatory drugs such as aspirin, indomethacin, and ibuprofen. By inhibiting cyclooxygenases, these drugs inhibit the synthesis of all of the prostaglandins and thromboxanes and block all of their biological actions. PGE_2 and PGD_2 are inactivated enzymatically in the lung and liver. TXA_2 and PGI_2 are chemically unstable and degrade nonenzymatically to the biologically inactive TXB_2 and 6-keto $PGF_{1\alpha}$, respectively. *See* ASPIRIN.

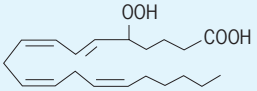
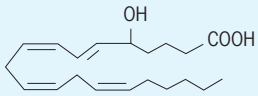
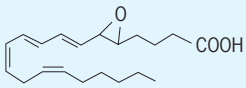
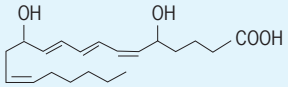
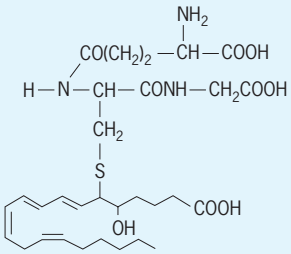
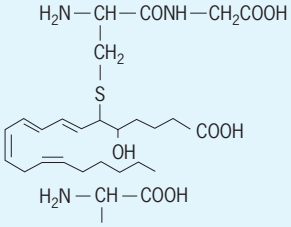
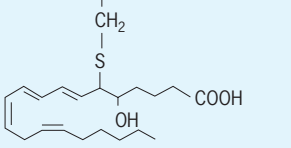
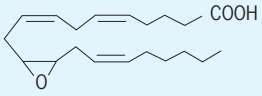
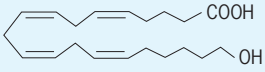
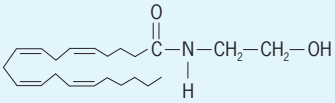
The cyclooxygenase metabolites fall into categories based on the substituents on the cyclopentane ring (Table 1). PGG_2 and PGH_2 are cyclic endoperoxides, PGE_2 and PGD_2 are hydroxy ketones, PGI_2 has a double ring structure, and $PGF_{2\alpha}$ is a diol. TXA_2 has a six-membered oxane ring. These categories of prostaglandins are further divided by the number of double bonds in the side chain, which is indicated by a subscript 1, 2, or 3. The prostaglandins produced from arachidonic acid have the subscript 2.

Metabolites of lipoxygenases. Lipoxygenases are a family of enzymes that catalyze the addition of a single molecule of oxygen into the unsaturated fatty acid to form a lipid hydroperoxide (Table 1). They are generally constitutive enzymes that depend on the availability of free arachidonic acid. This reaction involves hydrogen abstraction from an allylic carbon adjacent to two cis-double bonds. Since arachidonic acid has four cis-double bonds, several isomeric hydroperoxyeicosatetraenoic acids (HPETEs) may be synthesized, commonly 5-, 12-, and 15-HPETEs. The lipoxygenases place the hydroperoxy group on specific carbons. Cells differ in the lipoxygenases that they contain. For example, platelets contain only 12-lipoxygenase, which catalyzes the formation of only 12-HPETE, whereas leukocytes contain both 5-lipoxygenase and 12-lipoxygenase. 5-Lipoxygenase catalyzes the formation of only 5-HPETE. Like PGG_2 and PGH_2, HPETEs are unstable intermediates that may be further metabolized. They are reduced either enzymatically or nonenzymatically to the hydroxyeicosatetraenoic acid (HETE). Of particular importance is the 5-lipoxygenase pathway which leads to the production of the leukotrienes (Table 1). LTA_4 synthase catalyzes the rearrangement of 5-HPETE to the unstable epoxide, LTA_4. There are two possible enzymatic fates for LTA_4: hydrolysis by LTA hydrolase to LTB_4, or conjugation with glutathione by LTC synthase giving LTC_4. The glutathione group of LTC_4 may be sequentially metabolized by specific peptidases. Removal of the glutamic acid from the glutathione group gives LTD_4, and further removal of the glycine group gives LTE_4. LTB_4, LTC_4, LTD_4, and LTE_4 are all biologically active.

Metabolites of cytochrome P450. Cytochrome P450 is a family of greater than 100 isozymes (any two or more enzymes chemically distinct but functionally

TABLE 1. Structures of eicosanoids

Name	Abbreviation	Structure
Arachidonic acid (5,8,11,14-eicosatetraenoic acid)	AA	
Prostaglandin G_2	PGG_2	
Prostaglandin H_2	PGH_2	
Prostaglandin E_2	PGE_2	
Prostaglandin D_2	PGD_2	
Prostaglandin $F_{2\alpha}$	$PGF_{2\alpha}$	
Thromboxane A_2	TXA_2	
Thromboxane B_2	TXB_2	
Prostaglandin I_2	PGI_2	
6-Ketoprostaglandin $F_{1\alpha}$	6-ketoPGF$_{1\alpha}$	

Name	Abbreviation	Structure
5-Hydroperoxyeicosatetraenoic acid	5-HPETE	
5-Hydroxyeicosatetraenoic acid	5-HETE	
Leukotriene A_4	LTA_4	
Leukotriene B_4	LTB_4	
Leukotriene C_4	LTC_4	
Leukotriene D_4	LTD_4	
Leukotriene E_4	LTE_4	
11,12-Epoxyeicosatrienoic acid	11,12-EET	
20-Hydroxyeicosatetraenoic acid	20-HETE	
Anandamide		

the same) that metabolize endogenous compounds such as fatty acids and cholesterol and exogenous compounds such as drugs and xenobiotics. Arachidonic acid is metabolized by ω- and ω-1-hydroxylases or monooxygenases to 19-HETE and 20-HETE and by epoxygenases to EETs (Table 1). The ω-hydroxylations differ from the lipoxygenase reaction: the HETE is formed directly without the HPETE intermediate, and the ω-hydroxylase does not require the cis-double bonds. These hydroxylations are catalyzed by CYP1A and CYP4A isozymes. The epoxygenases add an oxygen across one of the four double bonds of arachidonic acid, giving 14,15-, 11,12-, 8,9-, or 5,6-EET. Epoxygenations result from CYP2B, CYP2C, and CYP2J isozymes. These enzymes are constitutively found in cells of the liver, kidney,

pituitary gland, and blood vessels. They are regulated by the availability of free arachidonic acid.

Other pathways. The free-radical peroxidation of arachidonic acid and arachidonic acid containing phospholipids gives rise to the isoprostanes. The isoprostanes associated with phospholipids may be released by phospholipases. These eicosanoids are comparable to the prostaglandins in structure, and some also have actions similar to the prostaglandins. Since their synthesis is independent of cyclooxygenase, their formation is not inhibited by nonsteroidal anti-inflammatory drugs.

Arachidonic acid and other fatty acids are coupled to ethanolamine by an amide bond to form *N*-acyl ethanolamides. For arachidonic acid, arachidonylethanolamide or anandamide is produced (Table 1). The synthetic pathway for anandamide and related *N*-acyl ethanolamides is unclear. Two pathways are proposed: hydrolysis of anandamide from *N*-arachidonylphosphatidylethanolamine by phospholipase D or direct condensation of arachidonic acid with ethanolamine. Anandamide is degraded by a fatty acid amidohydrolase, giving arachidonic acid and ethanolamine. The amidohydrolase acting in reverse may represent the synthetic pathway. The brain and heart are major sources of these compounds.

Physiological and pathological actions. The actions of the eicosanoids are diverse, affecting most organ systems (**Table 2**). As a result, when an eicosanoid is administered systemically in pharmacological doses, many of these actions are observed. However, physiologically and pathologically, eicosanoids are local hormones. They are synthesized and exert their action in a tissue, and only certain eicosanoids are made by the cells of a tissue. As a result, the action of an eicosanoid is usually discrete and related to the physiological demands of the tissue.

Mechanism of action. The prostaglandins, leukotrienes, TXA_2, and anandamide act by binding to and activating membrane receptor proteins. These receptors are of the seven-membrane spanning, serpentine receptor subgroup and are coupled to guanine nucleotide binding proteins. Distinct receptors have been identified for many eicosanoids (PGD_2, PGE_2, $PGF_{2\alpha}$, PGI_2, TXA_2, LTB_4, LTC_4, LTD_4/LTE_4) and anandamide. In many cases, subgroups of receptors have been identified. The existence of receptor subgroups explains how some eicosanoids such as PGE_2 have multiple actions. The prostaglandin receptors are termed prostanoid receptors, and they are named after the primary prostaglandin that activates the receptor. For example, the PGI_2 prostanoid receptor is termed the IP receptor. The other receptors are named in an analogous manner. Receptor subgroups are indicated by numbered subscript. The receptors are coupled to stimulation of adenylyl cyclase in the case of IP, DP, and EP_2 receptors. Adenylyl cyclase is inhibited by anandamide via a cannabinoid receptor and PGE_2 via the EP_3 receptor. The receptors for the leukotrienes, TXA_2, $PGF_{2\alpha}$, and PGE_2 (EP_1) are coupled to activation of phospholipase C and increases in intracellular calcium. Receptors have not been identified for 20-HETE or the EETs. Their

TABLE 2. Physiological actions of arachidonic acid metabolites

Eicosanoid	Biological effect	Physiologic or pathologic role
PGE_2	Vasodilation	Protects against intense vasoconstriction
	Decreases gastric acid secretion	Protects gastric mucosa
	Pain sensitization	Pain, inflammation
	Bronchodilation	
	Contracts gastrointestinal smooth muscle	Peristalsis, diarrhea
$PGF_{2\alpha}$	Contracts uterine smooth muscle	Parturition, dysmenorrhea
PGI_2	Contracts gastrointestinal smooth muscle	Peristalsis, diarrhea
	Vasodilation	Protects against intense vasoconstriction
	Inhibits platelet aggregation	Protects against thrombosis
	Decreases gastric acid secretion	Protects gastric mucosa
	Renin release	Blood pressure regulation
TXA_2	Vasoconstriction	Hemostasis, angina pectoris
	Causes platelet aggregation	Thrombosis, angina pectoris
	Contracts most smooth muscle	
	Bronchoconstriction	Bronchial asthma
PGD_2	Inhibits platelet aggregartion	
	Vasodilation	
	Pulmonary vasoconstriction	
LTC_4/LTD_4	Contracts gastrointestinal smooth muscle	
	Constricts peripheral airways	Bronchial asthma
	Increase mucous secretion	Bronchial asthma
	Leakage in microcirculation	Inflammation
LTB_4	Neutrophil and eosinophil chemotaxis	Inflammation
	Leakage in microcirculation	Inflammation
	Causes neutrophil aggregation	Inflammation
12-HETE/12-HPETE	Neutrophil chemotaxis	Inflammation
20-HETE	Vasoconstriction	Myogenic tone
EETs	Vasodilation	Regulates tissue blood flow
	Releases pituitary hormones	
Anandamide	Hypothermia	
	Antinociception	
	Catalepsy	
	Cerebral vasodilation	

mechanisms of action have not been clearly determined.

Cyclooxygenase metabolites. The prostaglandins and TXA_2 have several actions that are necessary for maintaining normal homeostasis and organ integrity. These effects are mediated by their constitutive synthesis through cyclooxygenase-1. PGE_2 and PGI_2 are potent vasodilators. These prostaglandins are synthesized in the blood vessel in response to vasoconstrictors such as angiotensin II and antagonize the effect of the vasoconstrictor. This action protects the organ from intense vasoconstriction and maintains nutritive blood flow. In the absence of these prostaglandins, injury to the organ occurs. These prostaglandins also protect the gastric mucosa. They inhibit gastric acid secretion and may exert other cellular protective effects. In their absence, gastric ulcers develop. TXA_2 is made principally by platelets. It causes vasoconstriction and promotes platelet aggregation. These actions prevent excessive blood loss when a blood vessel is damaged or severed. PGI_2 is made by the blood vessel wall and has the opposite effects. It inhibits platelet aggregation and prevents intravascular platelet aggregates from forming and obstructing blood flow. The balance between the proaggregatory or vasoconstrictor action of TXA_2 and the antiaggregatory or vasodilator action of PGI_2 is critical to cardiovascular homeostasis. For example, an increase in the effect of TXA_2 over PGI_2 has been implicated in ischemic heart disease. $PGF_{2\alpha}$ contracts uterine smooth muscle. With menstruation and destruction of the uterine lining, arachidonic acid is released and metabolized to $PGF_{2\alpha}$. Contraction of the uterine smooth muscle by $PGF_{2\alpha}$ contributes to the pain associated with menstruation. In pregnancy, an increase in prostaglandin synthesis is a major determinant of the onset of labor.

In inflammatory processes, immune cells migrate into the site of inflammation and release cytokines. These inflammatory mediators induce the expression of cyclooxygenase-2 in the immune cells and cells of the tissue. The increased synthesis of PGE_2 and PGI_2 mediates the vasodilation and pain sensitization associated with inflammation and arthritic conditions. Inhibitors of cyclooxygenase-1 and -2 such as aspirin and ibuprofen relieve the signs and symptoms of inflammation. However, they also block the constitutive, homeostatic functions of the prostaglandins by blocking cyclooxygenase-1, which can result in renal failure and gastric ulcers. Selective cyclooxygenase-2 inhibitors such as celecoxib are anti-inflammatory without the deleterious effects on homeostatic functions of the prostaglandins.

Lipoxygenase metabolites. The leukotrienes are synthesized by leukocytes, and they also serve as inflammatory mediators. The leukocytes are immune cells and become sequestered and activated at a site of inflammation. When activated, they release leukotrienes. LTB_4 is chemotactic and promotes leukocyte aggregation. Thus, it attracts other inflammatory cells to the site of inflammation and sequesters them. LTC_4, LTD_4, and LTB_4 increase capillary permeability, causing the leakage of fluid and protein from the vasculature into the tissue. This contributes to the swelling associated with inflammation. LTC_4 and LTD_4 also cause bronchiolar constriction and increase mucus formation in the airways. These actions are thought to contribute to bronchial asthma. Inhibitors of 5-lipoxygenase and leukotriene receptor antagonists are useful in treating inflammation and bronchial asthma.

Cytochrome P450 metabolites. The EETs and 20-HETE are involved in the regulation of vascular tone and organ blood flow. In the blood vessel, EETs are synthesized by the endothelial cells in response to vasodilator hormones such as acetylcholine and bradykinin. The EETs diffuse to the adjacent smooth muscle cells and exert their action. They open potassium channels and hyperpolarize the smooth muscle cell membrane. This results in vasodilation. Thus, EETs serve as mediators of vasodilation. 20-HETE is made by smooth muscle cells in response to increases in intravascular pressure. It inhibits potassium channel activity and depolarizes the cell membrane. Depolarization promotes the influx of calcium and vasoconstriction. 20-HETE is a mediator of myogenic tone in blood vessels. In the blood vessel wall, 20-HETE and the EETs are counterregulatory mechanisms that control blood flow.

Other pathways. Anandamide is a naturally occurring agonist of the cannabinoid receptor. It has biological actions identical to Δ^9-tetrahydrocannabinol, the active ingredient in marijuana, but is shorter-acting. Anandamide is synthesized and degraded by discrete regions of the brain and causes hypothermia (decreased body temperature), antinociception (decreased pain response), catalepsy (rigidity and decreased response to external stimuli), and cerebral vasodilation. It remains to be determined if anandamide functions as a chemical mediator in the brain. *See* ARTERIOSCLEROSIS; ARTHRITIS; ASTHMA; GOUT; MARIJUANA; PAIN; ULCER. William B. Campbell

Bibliography. C. Denzlinger, Biology and pathophysiology of leukotrienes, *Crit. Rev. Oncology/Hematology*, 23:167–223, 1996; C. J. Hillard and W. B. Campbell, Biochemistry and pharmacology of arachidonylethanolamide, a putative endogenous cannabinoid, *J. Lipid Res.*, 38:2382–2398, 1997; J. C. McGiff, Cytochrome P-450 metabolism of arachidonic acid, *Annu. Rev. Pharmacol.*, 31:339–369, 1991; W. L. Smith, L. J. Marnett, and D. L. DeWitt, Prostaglandin and thromboxane biosynthesis, *Pharmacol. Ther.*, 49:153–179, 1991; C. S. Williams and R. N. Dubois, Prostaglandin endoperoxide synthase: Why two isoforms?, *Amer. J. Physiol.*, 270:G393–G400, 1996.

Eigenfunction

One of the solutions of an eigenvalue equation. A parameter-dependent equation that possesses nonvanishing solutions only for particular values (eigenvalues) of the parameter is an eigenvalue equation, the associated solutions being the eigenfunctions (sometimes eigenvectors). In older usage the

terms characteristic equation and characteristic values (functions) are common. Eigenvalue equations appear in many contexts, including the solution of systems of linear algebraic equations (matrix equations), differential or partial differential equations, and integral equations. The importance of eigenfunctions and eigenvalues in applied mathematics results from the widespread applicability of linear equations as exact or approximate descriptions of physical systems. However, the most fundamental application of these concepts is in quantum mechanics where they enter into the definition and physical interpretation of the theory. Only linear eigenvalue equations will be discussed. *See* EIGENVALUE (QUANTUM MECHANICS); ENERGY LEVEL (QUANTUM MECHANICS); NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS.

Matrix equations. An orienting example will first be presented. A common problem, for instance in the theory of angular momentum, is to find the column vectors v which are solutions of the matrix, Eq. (1), where M is an N-by-N matrix with components M_{ij} which are assumed to be known, v is a column vector of length N with components v_i, and λ is the parameter. All quantities can in general be complex.

$$M \cdot v = \lambda v \tag{1}$$

For example, the nonvanishing solutions to Eq. (2)

$$\begin{pmatrix} -1 & 3 \\ 2 & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \lambda \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \tag{2}$$

may be sought. This equation can be rewritten as $M' \cdot v = \vec{0}$ for $M' = (M - \lambda 1_2)$. (Here, 1_2 is the two-dimensional unit matrix, and $\vec{0}$ is the zero vector.) The condition for such an equation to have nontrivial solutions for the vector, v, is that the determinant, $\det M'$, of the matrix M' is zero. Otherwise the inverse of M' would exist, implying that the only allowed solution is $v = (M')^{-1}\,\vec{0} = \vec{0}$. Thus, Eq. (3) is required.

$$\det \begin{pmatrix} -1-\lambda & 3 \\ 2 & -\lambda \end{pmatrix} = (1+\lambda)\lambda - 6 = 0 \tag{3}$$

The solutions are $\lambda = 2$ and $\lambda = -3$, the eigenvalues of the matrix equation. The two associated eigenvectors $v^{(1)}$ and $v^{(2)}$ are discovered by substituting these values into Eq. (2). For the eigenvalue $\lambda = 2$, the relation $v_1^{(1)} = v_2^{(1)}$ is found. This does not set the overall scale of $v^{(1)}$. Similarly, for $\lambda = -3$ the relation $v_1^{(2)} = -3v_2^{(2)}/2$ is found. The scale of the eigenvectors can be fixed by imposing the additional condition that the generalized dot product $(v^{(i)})^* \cdot v^{(i)} = 1$ for $i = 1, 2$.

The general N-by-N matrix equation can similarly be rewritten as Eq. (4). The condition for the existence of nonvanishing solutions for v is Eq. (5). This

$$(M - \lambda 1_N)v = \vec{0} \tag{4}$$

$$\det (M - \lambda 1_N) = 0 \tag{5}$$

determinant, when expanded, gives an Nth-order algebraic equation for λ of the form of Eq. (6). By the

$$\lambda^N + a_1\lambda^{N-1} + a_2\lambda^{N-2} + \cdots + a_N = 0 \tag{6}$$

fundamental theorem of algebra, this equation has N solutions λ_i, $i = 1, \ldots, N$, the eigenvalues of M, if λ is allowed to be complex. The solutions can be degenerate; in other words, a given value, λ, can appear more than once as a solution.

A particularly important special case is when the matrix M is self-adjoint (also known as hermitian), namely, when M is equal to its complex-conjugate transpose, as in Eq. (7). The eigenvalues of M are

$$M^\dagger \equiv M^{T*} = M \tag{7}$$

then guaranteed to be real, and the eigenvectors are orthonormal, as in Eq. (8), where $\delta^{ij} = 0$ when $i \neq j$

$$(v^{(i)})^* \cdot v^{(j)} = \delta^{ij} \tag{8}$$

and $\delta^{ij} = 1$ when $i = j$. For k degenerate eigenvalues, only a k-dimensional subspace is uniquely specified, but it is always possible to choose k such orthonormal eigenvectors, one for each eigenvalue. *See* ANGULAR MOMENTUM; MATRIX CALCULUS; MATRIX THEORY; RIGID-BODY DYNAMICS.

Linear operators. The general setting for linear eigenvalue equations is the notion of a linear operator $\hat{L}$ acting in a vector space V endowed with an inner product. A complex vector space is a set of elements (vectors), $\{|u\rangle, |v\rangle, \ldots\}$, that have the operations of addition and scalar multiplication defined, such that $\alpha|u\rangle + \beta|v\rangle$ is an element of V for any elements, $|u\rangle$ and $|v\rangle$, of V, and any complex numbers, α and β. Here, the notation introduced by P. Dirac in the context of quantum mechanics is employed. The operations of addition and scalar multiplication must further satisfy properties such as associativity, commutativity, and the distributive law.

A linear operator $\hat{L}$ is a rule that takes an arbitrary vector $|u\rangle$ in V into a, generally different, vector $|Lu\rangle$ in V, and that satisfies Eq. (9) for any vectors, $|u\rangle$

$$\hat{L}(\alpha|u\rangle + \beta|v\rangle) = \alpha|Lu\rangle + \beta|Lv\rangle \tag{9}$$

and $|v\rangle$, in V, and complex numbers α and β. The usual action of a matrix on a vector by matrix multiplication satisfies this linearity property. In this general setting, eigenvalue equations again take the form $\hat{L}|v\rangle = \lambda|v\rangle$. The total set of eigenvalues of a linear operator $\hat{L}$ acting on V is called the spectrum of $\hat{L}$. For differential operators this spectrum can contain continuous parts.

An inner product on the vector space V, denoted $\langle u|v\rangle$, is a complex-valued function of two arbitrary vectors, $|u\rangle$ and $|v\rangle$, in V which satisfies Eqs. (10)

$$\langle u|u\rangle \geq 0,\ \langle u|u\rangle = 0 \text{ only for } |u\rangle = |0\rangle \tag{10a}$$

$$\langle u|v\rangle = \langle v|u\rangle^* \tag{10b}$$

$$\langle u \mid \alpha v + \beta w\rangle = \alpha\langle u|v\rangle + \beta\langle u|w\rangle \tag{10c}$$

for any vectors, $|u\rangle$, $|v\rangle$, and $|w\rangle$, in V and complex

numbers α and β. Here, $|0\rangle$ is the zero vector in V. The inner product generalizes the usual dot product for three-dimensional vectors in physical space.

A basis for an N-dimensional complex vector space V is any set of N vectors $|e^{(i)}\rangle$ in V such that any vector $|u\rangle$ in V can be uniquely written as the sum in Eq. (11), where the N coefficients can be complex.

$$|u\rangle = \sum_{i=1}^{N} \alpha_i |e^{(i)}\rangle \tag{11}$$

For a general vector space, the number of basis elements can be (countably) infinite, with $N = \infty$, or the basis elements can even be labeled by a continuous parameter, in which case the dimension of the vector space is uncountably infinite and sums, such as in Eq. (11), need to be replaced by integrals. In both these latter cases, subtle questions of the convergence of various expressions arise.

Given any such basis and an inner product, a matrix representation L_{ij} of any abstract linear operator $\hat{L}$ can be constructed, by defining the components as in Eq. (12). The notion of the adjoint $\hat{L}^\dagger$ of an operator $\hat{L}$ can be generalized by the definition in Eq. (13)

$$L_{ij} = \langle e^{(i)} | L e^{(j)} \rangle \tag{12}$$

$$\langle u | L^\dagger v \rangle = \langle u | L u \rangle^* \tag{13}$$

for any $|u\rangle$ and $|v\rangle$ in V, and the matrix adjoint is then given by ${L^\dagger}_{ij} = L^*_{ji}$, in accordance with Eq. (7). A self-adjoint operator satisfies the property $\hat{L}^\dagger = \hat{L}$ for all elements in the vector space V. Just as in the case of a self-adjoint matrix, the eigenvalues of a such an operator are purely real, and the eigenvectors can always be chosen to be orthonormal.

Since the eigenvectors $\{|v^{(i)}\rangle\}$ of any self-adjoint operator $\hat{L}$ acting on V are a set of N orthonormal vectors in the N-dimensional space V, they form a basis for V and any vector $|w\rangle$ can be expanded as in Eq. (14), where the generally complex numbers

$$|w\rangle = \sum_{i=1}^{N} |v^{(i)}\rangle \langle v^{(i)} | w\rangle \tag{14}$$

$\langle v^{(i)} | w\rangle$ play the role of the α_i in Eq. (11). Since $|w\rangle$ is arbitrary, this equation implies the important consequence, Eq. (15), where $\hat{1}$ is the identity operator

$$\sum_{i=1}^{N} |v^{(i)}\rangle \langle v^{(i)}| = \hat{1} \tag{15}$$

on the space V (namely, the operator that takes any vector to itself). Equation (15) is known as the completeness relation for the eigenvectors of $\hat{L}$.

The relation (14), or equivalently (15), implies a very useful representation for any self-adjoint operator $\hat{\mathrm{M}}$. If $|w\rangle$ is expanded in the set of eigenvectors $|\tilde{v}^{(i)}\rangle$ associated with $\hat{\mathrm{M}}$, then, using the definition $\hat{M}|\tilde{v}^{(i)}\rangle = \lambda_i |\tilde{v}^{(i)}\rangle$, along with the fact that $|w\rangle$ is arbitrary, then gives Eq. (16). This is a simple,

$$\hat{M} = \sum_{i=1}^{N} \lambda_i |\tilde{v}^{(i)}\rangle \langle \tilde{v}^i| \tag{16}$$

finite-dimensional example of the spectral theorem, stating that the action of an operator may be represented purely in terms of its eigenvalue spectrum and eigenvectors. For matrices, this result says that any self-adjoint matrix can be put into a purely diagonal form by choosing the eigenvalues as a basis, the diagonal entries being the eigenvalues of the matrix. *See* HILBERT SPACE; LINEAR ALGEBRA; OPERATOR THEORY.

Differential eigenvalue equations. Eigenvalue problems also appear very frequently in the solution of linear differential equations. An example is the motion of a uniform vibrating string with fixed ends at $x = 0, L$. The partial differential equation that describes the displacement $y(x, t)$ of the string from its equilibrium position (taken to be $y = 0$) is Eq. (17).

$$\frac{\partial^2 y(x,t)}{\partial t^2} - c^2 \frac{\partial^2 y(x,t)}{\partial x^2} = 0 \tag{17}$$

where c is the wave velocity on the string. Along with this equation, appropriate initial and boundary conditions must be specified. The fixed end-point condition implies $y(0,t) = y(L,t) = 0$ for all t. A typical initial condition is that the string is held stationary with some particular displacement and then released.

The solution of the problem is greatly facilitated by the principle of superposition, which states that, for two solutions, $y^{(1)}(x,t)$ and $y^{(2)}(x,t)$, of a linear equation, the sum in expression (18) is also a solution

$$\alpha y^{(1)}(x,t) + \beta y^{(2)}(x,t) \tag{18}$$

for arbitrary constants α and β. The general strategy to solve such a problem is therefore to decompose $y(x, t)$ into sum of simple motions, the eigenfunctions of the problem, whose time development can straightforwardly be solved for, and then form a suitable linear combination such that the initial conditions are satisfied.

Explicitly, elementary solutions of the form $y(x, t) = b(x) \cos \omega (t - t_0)$ are studied. Substituting this into Eq. (17) results in Eq. (19). This is of the

$$\frac{d^2 b(x)}{dx^2} = -\frac{\omega^2}{c^2} b(x) \tag{19}$$

form of an eigenvalue equation with ω^2 as the parameter. The set of (differentiable) functions $\{f(x)\}$ on the interval $[0,L]$ which vanish at the end points forms a vector space V (of infinite dimension), and the differential operator d^2/dx^2 is a linear operator acting on V. An inner product on this space can be defined by Eq. (20), the adjoint operator $(d^2/dx^2)^\dagger$

$$\langle f | g \rangle \equiv \int_0^L dx (f(x))^* g(x) \tag{20}$$

can be defined with respect to this inner product by applying Eq. (13), and d^2/dx^2 can be shown to be a self-adjoint operator. The eigenvalues of this operator must therefore be real, and the associated eigenfunctions satisfy orthonormality and completeness relations.

The general solution to Eq. (19) is Eq. (21), with

$$b(x) = A\cos(\omega x/c) + B\sin(\omega x/c) \qquad (21)$$

A, B arbitrary constants. Imposing the boundary conditions at $x = 0$, L gives $A = 0$ and $B \sin(\omega L/c) = 0$. A nontrivial solution ($B \neq 0$) therefore fixes the frequency ω to be one of the discrete set given by Eq. (22), the eigenvalues of the problem. The associ-

$$\omega_n = \frac{\pi nc}{L} \quad \text{for} \quad n = 1, 2, \ldots \qquad (22)$$

ated eigenfunctions are $b_n(x) = \sin(\omega_n x/c)$. In the absence of fixed end-point boundary conditions (that is, for a string of infinite length) the possible values of ω would not be restricted, and the spectrum of eigenvalues would be continuous.

A linear combination, Eq. (23), which, by the prin-

$$y(x, t) = \sum_{n=1}^{\infty} a_n \sin(\omega_n x/c)\cos\ \omega(t - t_0) \qquad (23)$$

ciple of superposition, is still a solution to the original equation (17), is now constructed in such a way that the initial conditions are satisfied. The analog of the completeness relations, Eqs. (11) and (15), appropriate for this case imply that any continuous function $y_0(x)$ representing the initial displacement [satisfying $y_0(0) = y_0(L) = 0$] on the interval $[0,L]$ can be expanded in such a series (Fourier's theorem). The orthogonality property then leads to the formula (24) for the coefficients. Equation (24), together

$$a_n = \frac{2}{L}\int_0^L dx \sin(\pi nx/L) y_0(x) \qquad (24)$$

with Eqs. (22) and (23), constitutes a full solution of the problem. *See* DIFFERENTIAL EQUATION; FOURIER SERIES AND TRANSFORMS; INTEGRAL TRANSFORM; ORTHOGONAL POLYNOMIALS. John March-Russell

Bibliography. G. B. Arfken and H.-J. Weber (eds.), *Mathematical Methods for Physicists*, 4th ed., 1995; R. Geroch, *Mathematical Physics*, 1985; E. Merzbacher, *Quantum Mechanics*, 3d ed., 1997; R. Shankar, *Principles of Quantum Mechanics*, 2d ed., 1994.

Eigenvalue (quantum mechanics)

One of the values of the parameter in an eigenvalue equation for which the equation has a solution. A linear eigenvalue equation is a parameter-dependent equation of the form (1) that possesses nonvanish-

$$\hat{L}|v\rangle = \lambda|v\rangle \qquad (1)$$

ing solutions, $|v^{(n)}\rangle$, only for particular values, λ_n, of the parameter. These are the eigenvalues, the associated solutions being called the eigenfunctions (sometimes eigenvectors). Here $\hat{L}$ is a linear operator acting in a generalized vector space, V, with an inner product. The concepts of eigenvalue and eigenfunction are of fundamental importance in quantum mechanics. The relation between the mathematical formulation of the theory and the possible results of a physical measurement is couched in terms of the eigenvalues and eigenvectors of a privileged set, $\{\hat{L}\}$, of linear operators which represent all possible physical observables. *See* EIGENFUNCTION.

Wave functions and operators. The mathematical formulation and physical interpretation of quantum mechanics is encapsulated in a number of fundamental postulates. The first is that all possible information regarding the physical state of any system at a given time is encoded in a generalized vector of a vector space, V. In most quantum-mechanics applications these vectors are complex-valued functions, $\psi(x)$, of the coordinates, x, of the system, termed wave functions. Only this case is considered in the following. Physical observables, L, are represented by self-adjoint linear operators acting on this function space. For instance, the momentum of a particle moving in one dimension is represented by the differential operator in Eq. (2), where $\hbar$ is Planck's constant over 2π.

$$\hat{p} = -i\hbar\frac{d}{dx} \qquad (2)$$

Furthermore, any self-adjoint linear operator, $\hat{L}$, has a set of eigenfunctions, $\phi^{(n)}(x)$ (for simplicity, assumed indexed by a discrete label n), in terms of which any wave function, $\psi(x)$, in V may be expanded, as in Eq. (3). Here the coefficients a_n are

$$\psi(x) = \sum_n a_n \phi^{(n)}(x) \qquad (3)$$

complex numbers. The sense of Eq. (3) is that both sides are equivalent, at least as far as all physical measurements are concerned. This is known as the expansion postulate.

Measurement of observables. A particularly important aspect of the theory is that the only possible result of a physical measurement of an observable, L, is one of the eigenvalues, λ_n, of the operator $\hat{L}$ corresponding to L, as in Eq. (4). Since the result of a

$$\hat{L}\phi^{(n)}(x) = \lambda_n \phi^{(n)}(x) \qquad (4)$$

physical measurement is necessarily a real number, the set of operators allowed as representatives of observables must be restricted by the requirement that their eigenvalues are purely real. Self-adjoint operators have this property, and this is the reason why representatives of observables are restricted to this class. In addition, the probability of finding the result λ_n when the system is initially in an arbitrary state $\psi(x)$ is given by Eq. (5), where the brackets rep-

$$\text{Prob}(\lambda_n) = |\langle\phi^{(n)} \mid \psi\rangle|^2 = |a_n|^2 \qquad (5)$$

resent the inner product on the space V, given by Eq. (6). After measuring a particular value, λ_n, the

$$\langle\phi^{(n)} \mid \psi\rangle = \int dx(\phi^{(n)}(x))^* \psi(x) \qquad (6)$$

new state of the system is given purely by $\psi(x) = \phi^{(n)}(x)$. Only when the wave function of the system is a single eigenfunction, $\phi^{(n)}(x)$, of some observable,

$\hat{L}$, can the system be said to definitely have the property λ_n.

Eigenvalue spectra. Unlike the case of classical mechanics where dynamical quantities such as energy or angular momentum can take on continuously varying values, the spectrum of eigenvalues is very often discrete, or contains a discrete portion. The values of the associated dynamical quantities are then said to be quantized. For instance, the energy of an electron bound in an atom can be only one of a discrete set of energy eigenvalues. *See* ENERGY LEVEL (QUANTUM MECHANICS).

Interpretation of superposition. Although Eq. (3) used the fact that the superposition principle applies to quantum mechanics, its interpretation is quite different from that in classical mechanics. For example, if two functions describing the displacement of an elastic string are superposed, $a_1y_1(x) + a_2y_2(x)$, then this new string configuration has, in general, quite different physical properties from a string with either displacement $y_1(x)$ or $y_2(x)$. However, in quantum mechanics the superposition $a_1\phi^{(1)}(x) + a_2\phi^{(2)}(x)$ of two eigenfunctions has the interpretation that sometimes the system has exactly the physical properties of $\phi^{(1)}(x)$, other times that of $\phi^{(2)}(x)$, with probabilities $|a_1|^2$ and $|a_2|^2$ respectively.

Commutation of operators. If the energy operator (the hamiltonian), $\hat{H}$, of the system commutes with an operator $\hat{L}$, that is, $\hat{H}\hat{L} = \hat{L}\hat{H}$, then a wave function that is an eigenfunction $\phi^{(n)}(x)$ of $\hat{L}$ stays so as the system dynamically evolves. Thus, the associated eigenvalue, λ_n, is a constant of the motion, and can be thought of as labeling the state of the system. The term quantum number is then used for λ_n. Further, it is possible to find a wave function which is simultaneously the eigenvector of two different operators if and only if they commute. A fundamental property of quantum mechanics is that the maximal number of attributes that can be simultaneously, and precisely, known about any system is given by the eigenvalues of the largest possible commuting set of operators that also includes the hamiltonian $\hat{H}$, in other words, the maximal set of quantum numbers. For example, the operators representing momentum and position do not commute, and it is not possible to know both simultaneously with arbitrary accuracy.

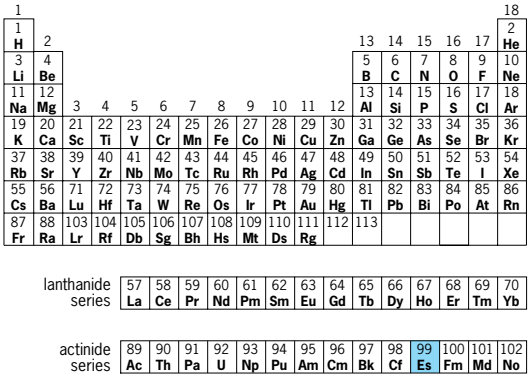
Although these postulates imply a picture of the physical world quite different from that in classical physics, the consequences of these postulates, together with the rest of quantum mechanics, have been verified with extraordinary precision. *See* NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS.

John March-Russell

Bibliography. S. Gasiorowicz, *Quantum Physics*, 2d ed., 1995; E. Merzbacher, *Quantum Mechanics*, 3d ed., 1997; R. Shankar, *Principles of Quantum Mechanics*, 2d ed., 1994.

Einsteinium

A chemical element, Es, atomic number 99, a member of the actinide series in the periodic table. It is not found in nature but is produced by artificial nuclear transmutation of lighter elements. All isotopes of einsteinium are radioactive, decaying with half-lives ranging from a few seconds to about 1 year. *See* ACTINIDE ELEMENTS; PERIODIC TABLE; RADIOACTIVITY.



Einsteinium is the heaviest actinide element to be isolated in weighable form. The metal is chemically reactive, is quite volatile, and melts at 860°C (1580°F); one crystal structure is known. *See* TRANSURANIUM ELEMENTS.

Stanley G. Thompson; Glenn T. Seaborg

Bibliography. S. Hofmann, *On Beyond Uranium: Journey to the End of the Periodic Table*, 2002; J. J. Katz and G. T. Seaborg, *The Chemistry of the Actinide Elements*, 2d ed., 1986; L. Morss and J. Fuger (eds.), *Transuranium Elements: A Half Century*, 1992; G. T. Seaborg, *Modern Alchemy: Selected Papers of Glenn T. Seaborg*, 1994; G. T. Seaborg and W. D. Loveland, *The Elements Beyond Uranium*, 1990.

El Niño

In general, an influx of warm water to the central and eastern equatorial Pacific Ocean off the coast of Peru and Ecuador, with a return period of 4–7 years. El Niño events come in various strengths: weak, moderate, strong, very strong, and extraordinary. The size of an event can be determined using various criteria: the amount of warming of sea-surface temperatures in the central and eastern Pacific from their average condition; the areal extent of that warm water anomaly; and the length of time that the warm water lingers before being replaced by colder-than-average sea-surface temperatures in this tropical Pacific region.

Under normal conditions the winds blow up the west coast of South America and then near the Equator turn westward toward Asia. The surface water is piled up in the western Pacific, and the sea level there is several tens of centimeters above average, while the sea level in the eastern Pacific is below average. As the water is pushed toward the west, cold water from the deeper part of the ocean along the Peruvian coast wells up to the surface to replace it. This cold water is rich with nutrients, making the coastal upwelling region along western South America among the most productive fisheries in the world.

Every 4–7 years those winds tend to die down and sometimes reverse, allowing the warm surface waters that piled up in the west to move back toward the eastern part of the Pacific Basin. With reduced westward winds, the surface water also heats up. Sea level drops in the western Pacific and increases in the eastern part of the basin. An El Niño condition can last for 12–18 months or longer, before the westward-flowing winds start to pick up again.

Limiting the impact. El Niño causes shifts in regional and local climate conditions around the globe. For many locations, forecasting El Niño provides useful information about how local weather conditions might be affected during the lifetime of the phenomenon. This lead time enables societies to prepare for the impacts of these extreme events.

The 1997–1998 and 1982–1983 events are considered the two biggest El Niños in the past 100 years. Both surprised the scientific community, but in different ways. The 1982–1983 event was the most damaging up to that time; no one had forecast it. The 1997–1998 event was larger and developed earlier and faster than expected. While El Niño's onset was not forecast, many of its worldwide impacts were, enabling societies to prepare for the onslaught. The 1982–1983 event emphasized to scientists and policy makers the importance of understanding and forecasting El Niño. The 1997–1998 event exemplified how many sectors of society (for example, agriculture, mining, construction, fisheries, hydroelectric power, and health) use El Niño forecasts to improve their decision-making processes.

El Niño is considered the second biggest climate-related influence on human activities, after the natural flow of the seasons. Although the phenomenon is at least thousands of years old, its impacts on global climate have only recently been recognized. Due to improved scientific understanding and forecasting of El Niño's interannual process, societies can prepare for and reduce its impacts considerably.

Michael H. Glantz

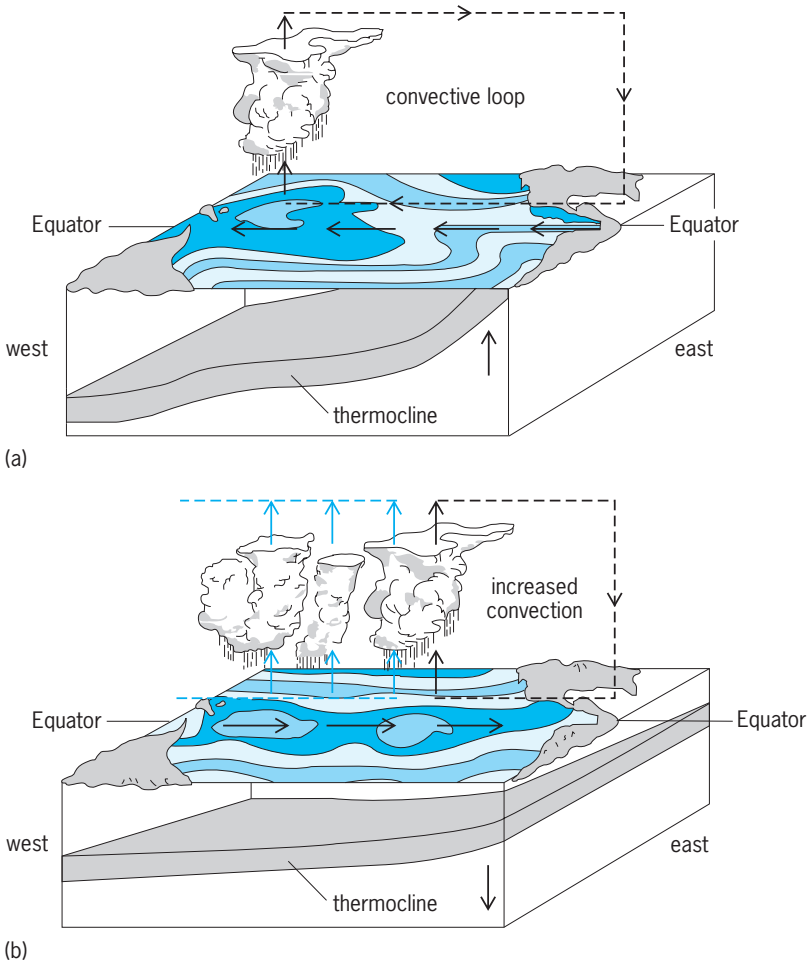


Fig. 1. Conditions of the tropical atmosphere–ocean system over the equatorial Pacific. (*a*) Normal, in which the thermocline is highly tilted and approaches the surface in the east Pacific, the eastern sea-surface temperature is cool, the western sea-surface temperature is warm, and rainfall remains in the far west. (*b*) El Niño (warm phase of ENSO) in which the thermocline flattens, the east and central Pacific warms, and the rainfall moves into the central Pacific.

Tropical Pacific Ocean. The tropical Pacific Ocean is 15,000 km (9300 mi) wide. Over the very warm water at its western end, there is heavy rainfall, low atmospheric surface pressure, and large-scale atmospheric ascent. At the eastern end the water is usually cool, there is little rainfall, atmospheric surface pressure is high, and the air is descending (**Fig. 1***a*). This east-west asymmetry of sea-surface temperature in the Pacific is quite remarkable, for the Earth rotating about its axis sees the Sun with no east-west asymmetry.

The proximate cause of this east-west difference in the sea-surface temperature is the direction of the surface wind, which in the tropics normally blows from east to west. If there were no wind, the sea-surface temperature would be uniformly warm in the east-west direction, and it would rapidly drop at deeper levels in the ocean. The region of rapid temperature change with depth is called the thermocline (Fig. 1*a*), and in the absence of winds would remain flat from east to west.

The thermocline can be thought of as marking the depth of the cold water. When the thermocline is close to the surface, cold water is near the surface; when the thermocline is deep, cold water is far from the surface. The normal westward wind tilts the thermocline, and deepens it in the west and shallows it in the east (Fig. 1*a*). The westward wind tends to bring cold water to the surface in the east (where the cold water is near the surface), and to bring warm water to the surface in the west (where the cold water is far from the surface). The west-to-east decrease of sea-surface temperature is therefore due to the cold water upwelled in the east by the westward winds. In the absence of westward winds, the temperature of the tropical Pacific would be uniform from east to west and would have about the same temperature that the western Pacific has now.

Superimposed on this east-to-west temperature difference is an annual cycle that has east Pacific sea-surface temperatures coolest in August and warmest in April, in contrast to midlatitudes, which have January and July as the extreme months. In general, the warm pool in the west moves northwest

during northern summer and southeast during northern winter. Throughout this complicated annual cycle, the basic pattern is maintained, with warm water in the west and its concomitant heavy rainfall, low pressure, and upward atmospheric motion. *See* PACIFIC OCEAN; SEAWATER; TROPICAL METEOROLOGY; UPWELLING.

Temperature, pressure, and precipitation cycles. Every 4–7 years, the temperature pattern of the tropical Pacific changes dramatically. It is as if all the processes that normally maintain the climatic conditions no longer occur and the tropical Pacific becomes uniformly warm, which would be expected in the absence of winds. This sporadic failure of the normal climate and its transition to a warmer Pacific has acquired the name El Niño (Fig. 1*b*). Occasionally the opposite also occurs, where the eastern Pacific becomes cooler than normal, rainfall decreases still more, atmospheric surface pressure increases, and the westward winds become stronger. This irregular cyclic swing of warm and cold phases in the tropical Pacific is called the El Niño Southern Oscillation (ENSO). The Southern Oscillation is the oscillation of the difference of surface pressure between the east and the west as the sea-surface temperature changes. When the eastern tropical Pacific is warm (El Niño or warm-phase conditions), the pressure in the east decreases and the pressure in the west increases, so the anomalous pressure difference (the Southern Oscillation Index) is negative. Precipitation actively participates in the cycle since regions of persistent precipitation lie over the warmest waters. *See* CLIMATOLOGY; MARITIME METEOROLOGY; PRECIPITATION (METEOROLOGY).

Observing ENSO. Prior to 1979, observations of equatorial Pacific climate were limited to surface measurements from islands and merchant ships; beginning in 1979, satellite estimates of surface and atmospheric-column quantities became available. Aside from a small number of research cruises, the interior of the ocean remained a mystery. As a result of the international Tropical Ocean Global Atmosphere (TOGA) program (1985–1995), an observing system specifically designed to reveal and monitor those features then believed most relevant to ENSO was deployed, and is maintained to this day (**Fig. 2**).

The centerpiece of the system is a set of 70 buoys moored to the bottom of the ocean by anchors (usually discarded railroad wheels) and connected to a surface torus by 4–5 km (2.5–3 mi) of line. The torus contains a suite of instruments that measure near-surface quantities (winds, humidity, and sea-surface temperature). Hung on the line beneath the surface and extending downward to about 500 m (1600 ft) is a set of instruments that measure temperature and pressure (with some measuring ocean currents) and relay the information to computers on the torus. Once a day the information is relayed to satellites and then becomes freely available by the Global Telecommunication System, whose users are the weather services of the world, and by the Internet to other users and to the public. Additional observations by ships, surface buoys drifting with the currents, and tide

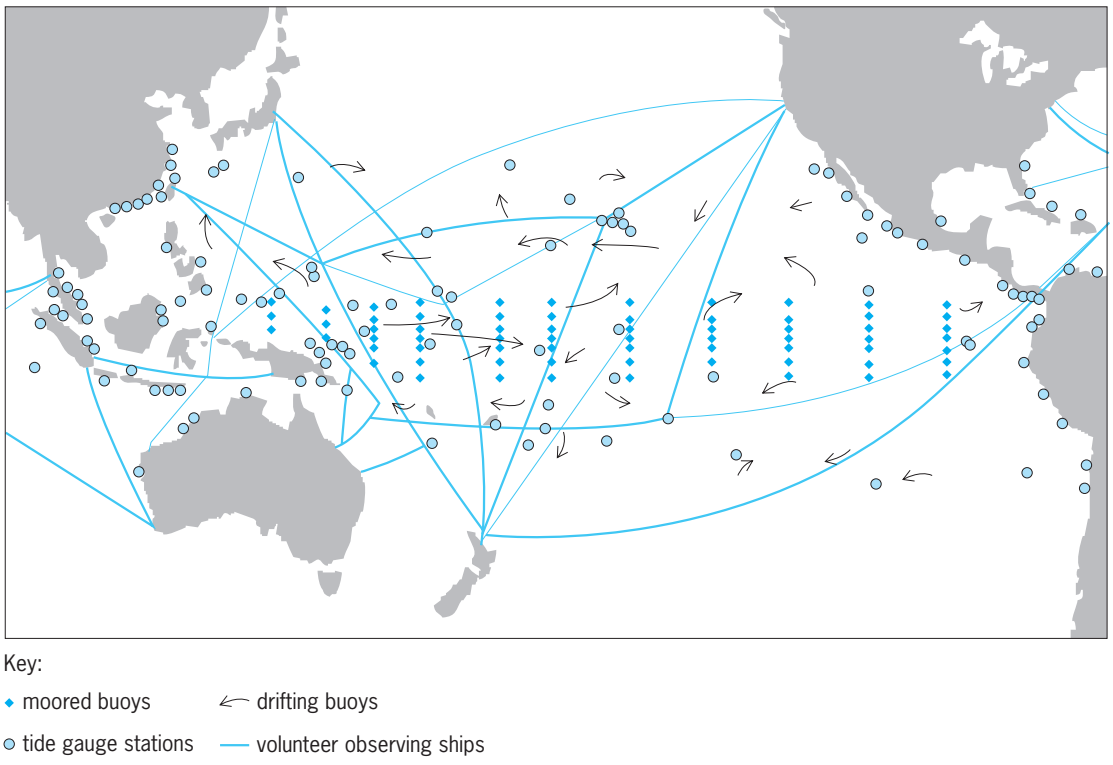


Fig. 2. ENSO Observing System. Diamonds are the moored buoys, arrows are drifting buoys, points are tide gauges, and lines are ships.

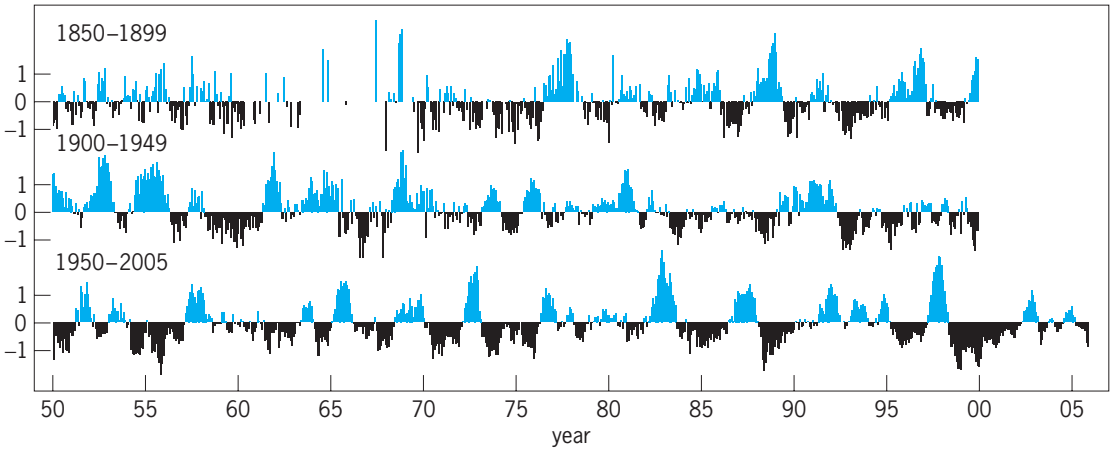


Fig. 3. Time series of an index of ENSO consisting of the monthly average temperature anomaly in the equatorial Pacific from 6°S to 6°N, 90°W to 180°W. The global mean sea-surface temperature deviation has been subtracted for each month to remove the effect of the gradual warming of the Earth's temperature, which is approximately 0.8°C (1.44°F) over the record. There are no observations for many of the months in the earlier periods, especially the 1860s, which accounts for their more ragged appearance.

gauges at islands complete the system. The ENSO observing system has been in full operation since 1995 and provides an unprecedented view of the evolution of ENSO. In particular, it was available to observe what is arguably the largest warm phase of ENSO in the last 150 years, the event of 1997–1998. *See* INSTRUMENTED BUOYS; SATELLITE METEOROLOGY.

History of ENSO. Instrumental records of surface atmospheric pressure and sea-surface temperatures document ENSO variability back to the 1870s and, with less confidence, to the 1850s. The monthly average central and eastern equatorial Pacific sea-surface temperature exhibits persistent, large-magnitude (>1°C or 1.8°F) deviations from the annual cycle (**Fig. 3**) that are the temporal signature of ENSO. The 1982–1983 and 1997–1998 warm phases of ENSO were characterized by temperature deviations in excess of 2°C (3.6°F). It also can be seen that the frequency of the warm and cold episodes varies throughout the record and, in particular, from decade to decade. The years since the 1997–1998 warm episodes have been, on average, colder than the long-term mean.

ENSO-related precipitation and temperature variations produce changes and hence proxy records of ENSO variability in many geophysical and biological systems. Occasional measurements of South American glacier masses corroborate some of the major ENSO episodes of the past 100 years. For the period back to 1500, records of sailing transit times, newspaper and literary accounts of rainfall and flooding, descriptions of mortality of marine organisms and birds, and other sources have been used to chronicle large-magnitude ENSO warm episodes. Centennial to millennial ENSO variability can be inferred from lake deposit and tree growth bands, and coral, shellfish, and glacier isotope concentrations.

The latter proxy records have been used to document ENSO back to 130,000 years ago (hereafter 130 Ka), and include epochs in which the solar forcing was different from today. ENSO variability appears to have been diminished during 6 Ka, a period when precession of the Earth's orbit enhanced the boreal summer equatorial solar forcing (output) and diminished the boreal winter solar forcing. ENSO episodes tend to develop during the summer, and the enhanced forcing is thought to have inhibited ENSO occurrences by differentially warming the western equatorial Pacific, which leads to increased precipitation in the west, and westward surface winds, stronger upwelling, and colder sea surface temperatures in the east. The global glacier extent in 6 Ka was not too different from today, at least in comparison with the epochs centered on 20 Ka (last glacial maxima) and 140 Ka, during which the ice volumes were five times the current values. The coral record suggests that ENSO variability was diminished relative to today during these icy epochs. Changes in the eccentricity of the Earth's orbit significantly changed the solar forcing during these periods, and the mechanisms by which these slow insolation changes influence ENSO are a subject of current research. *See* CLIMATE HISTORY; INSOLATION.

Teleconnections. The core of the ENSO is in the tropical Pacific, but the phenomenon influences the climate throughout the tropics and the extratropical latitudes of the winter hemisphere. Such planetary-scale atmospheric patterns are referred to as teleconnections and, in the case of ENSO, they are due to changes in the distribution of tropical Pacific precipitation.

Typical teleconnections during the boreal winter of a warm ENSO cycle are shown in **Fig. 4**. Within the tropics, warm ENSO is associated with enhanced precipitation in the central and eastern equatorial Pacific, torrential precipitation in the northern coastal desert of Peru, and diminished precipitation (and sometimes droughts) over Indonesia, northern Australia, northeast Brazil, and southern Africa. The expansion of the region of persistent precipitation into the tropical Pacific also leads to warming of the entire tropical troposphere by about

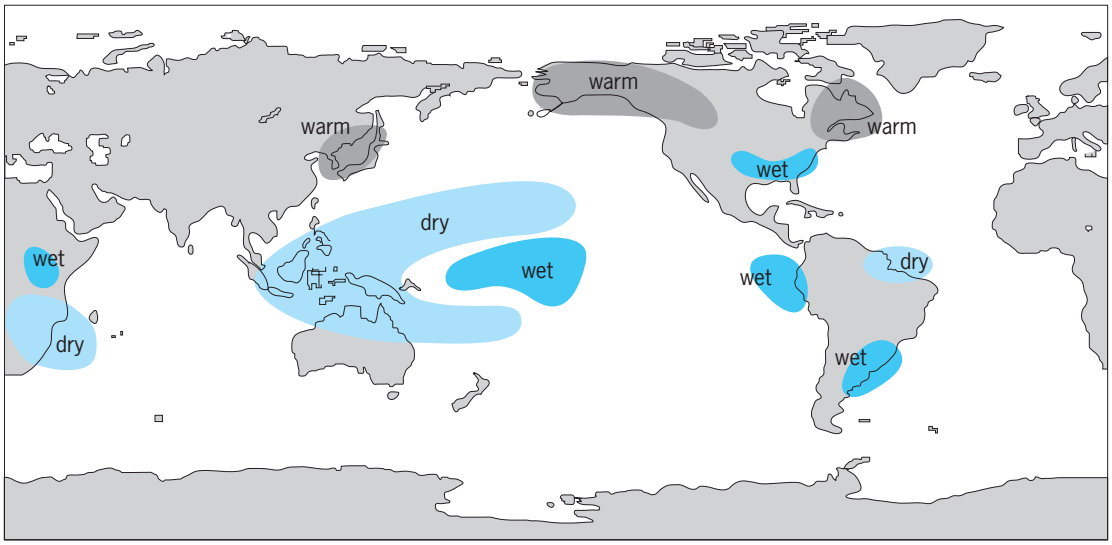


Fig. 4. Teleconnection pattern of warm phases of ENSO to the globe.

1°C (1.8°F). Precipitation deviations of opposite sign are observed during cold ENSO cycles, with the exception of the Peru coast, where it only rains only during warm ENSO. Enhanced atmospheric subsidence during warm ENSO warms the tropical Indian Ocean (**Fig. 5***b*, upper panel) and inhibits the normal February through May rains in northeast Brazil as well as the precipitation that maintains the equatorial Andean glaciers. Warm ENSO is associated with more hurricanes in the eastern Pacific and less in the Caribbean/Atlantic, and vice versa during cold ENSO cycles.

ENSO influences the higher latitudes primarily through changes in the preferred paths of storms, called storm tracks, in the winter hemisphere. Winter storms (weather) and their preferred locations are associated with the mean position of the jet streams, and these storms move heat north and south and produce precipitation in the middle latitudes. The rearrangement of storm tracks by ENSO produces deviations in the seasonal mean climate and in the distribution of extreme daily weather events over North America (Fig. 4), with diminished storminess over Alaska and Canada, and enhanced precipitation over the southern United Sates. Similar relationships exist in the Southern Hemisphere during austral winter with, for example, a southward shift of precipitation from the Andes' Altiplano at 15°S to the Chilean coast during ENSO warm episodes.

The year-to-year, decadal, and secular variability of the tropical Pacific climate all influence the teleconnections and climate impacts on the Northern Hemisphere. Tropical Pacific climate variability on both the year-to-year and longer time scales influence surface temperature and precipitation over the contiguous United States in the regions shown in Fig. 4. The extratropical temperature and precipitation deviations are different for warm and cold ENSO phases and reflect a nonlinear relationship between the storm tracks and the planetary-scale atmospheric circulation in the extratropics. Although the teleconnections from the Pacific are primarily into the winter hemisphere, there is observational and modeling evidence that cool surface-temperature deviations in the tropical Pacific can contribute to persistent summertime droughts in the central United States.

Decadal ENSO. Tropical Pacific climate variability is dominated by the year-to-year swings of the ENSO cycle that are shown in Fig. 3. There are also small-amplitude changes in sea-surface temperatures over broad regions of the Pacific, including the tropics, on time scales longer than the year-to-year variability. Figure 5*a* presents a pair of time series that documents tropical Pacific temperature variability.

Year-to-year ENSO temperature deviations (Fig. 5*a* upper) are characterized by deviations in excess of 0.5°C (0.9°F), centered on the Equator in the central and eastern Pacific, with equatorial upwelling playing a dominant role in producing these temperature changes. Deviations of comparable magnitude and like sign are found along the coasts of Ecuador and Peru, and are due to changes in upwelling. Smaller-magnitude (<0.15°C or 0.27°F) warm deviations occur in the tropical Indian Ocean, and cold deviations (<−0.15°C) are found in the western equatorial Pacific, North Pacific (30°S–50°N), and South Pacific (20°S–50°S). Longer, decadal time-scale surface-temperature deviations (Fig. 5*b* lower) are superficially similar to those for the year-to-year variability, but the extratropical and tropical Pacific deviations are comparable in magnitude (>0.2°C or 0.36°F) on longer time scales, and the central and eastern equatorial temperature deviations extend over a broader latitudinal scale than for the year-to-year variability. For comparison, the secular warming of the oceans, which has been removed from these analyses, is approximately 0.8°C (1.44°F) over the same period. The decadal variability is associated with cooler eastern Pacific conditions in the mid-1940s through late 1970s and warmer conditions in the 1980s and 1990s (Fig. 5*a* lower). This decadal oscillation is related to the Pacific Decadal Oscillation,

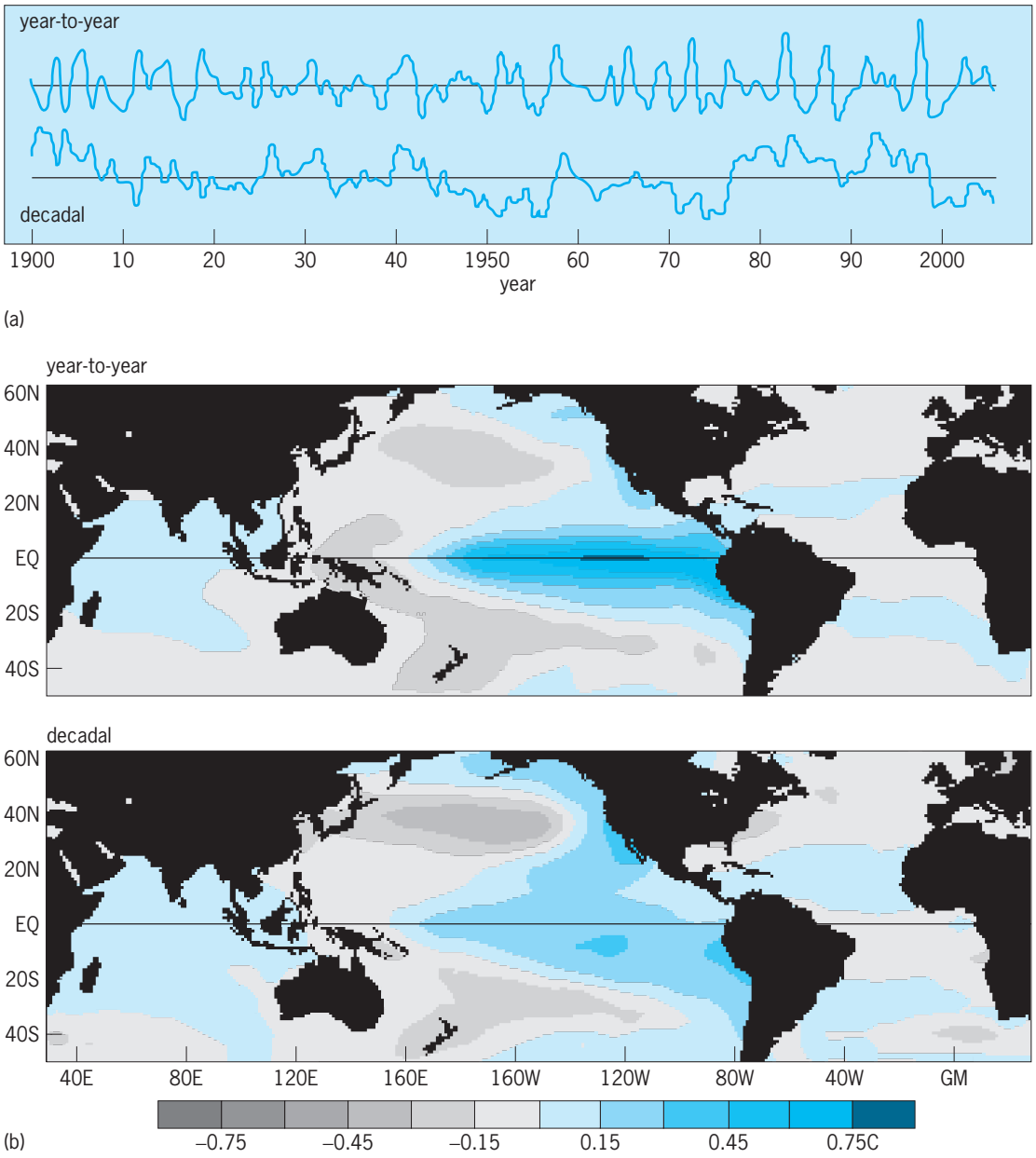


Fig. 5. Time series of (*a*) tropical Pacific ENSO and ENSO-like variations and (*b*) the characteristic global patterns that go with these series.

which has effects similar to those of ENSO on the United States.

The existence of decadal (and longer time scale, say global warming) modulations of ENSO call into question the definition of ENSO anomalies with respect to a single unchanging background climate.

Theories of ENSO. It is impossible to understand ENSO by considering the atmosphere or the ocean alone. The normal tropical Pacific sea-surface temperature (warm in the west, cold in the east) can be understood in terms of the wind blowing westward over the surface of the ocean. Similarly, the normal atmosphere over the tropical Pacific (heavy rainfall in the west, westward winds, low pressure, and ascending air) can be understood in terms of the sea-surface temperature. But to predict the future evolution of ENSO, the winds would have to be known in order to predict the sea-surface temperature, and the sea-surface temperature would have to be known in order to predict the winds.

The only way to simulate the consistent evolution of the atmosphere and ocean is to make coupled numerical models of them. A coupled model is one in which the atmospheric component sees and drives the oceanic component at the same time as the oceanic component sees and drives the atmospheric component. The way to predict the future evolution of ENSO is to start with the atmosphere–ocean system in its current state and allow the coupled system to evolve in a mutually consistent manner. *See* CLIMATE MODELING.

The first such coupled model was constructed in the mid-1980s by S. E. Zebiak and M. A. Cane. It was a simplified model in that only the upper layers

of the ocean were modeled and, more importantly, the annual cycle of sea-surface temperature and surface winds was specified. In this sense, the model was an anomaly model, calculating anomalies with respect to the specified annual cycle. Its importance lay in the fact that it was able to simulate ENSO in ways thought to be relevant. This advance led to a relatively simple explanation of ENSO and made possible the first successful long-term prediction of sea-surface temperature in the eastern Pacific by means of a numerical model of the coupled atmosphere–ocean system. *See* CLIMATIC PREDICTION.

The mechanism of ENSO that proved to be operative in the simplified Zebiak-Cane model relied on instability of the coupled system: when water in the eastern Pacific was warm, eastward winds were forced at the western edge of the warm water. These eastward winds further warm the water by a number of different mechanisms: by advecting warm water from the warmer West Pacific; by deepening the thermocline and therefore quelling the upwelled cold water; and by weakening the westward winds and therefore weakening the upwelling itself. As the instability sets on and the warm water grows, planetary (Rossby) waves are excited that contain the seeds of the reversal of the instability. They propagate westward, are reflected at the western edge of the Pacific, and return to counter the growing warming in the eastern Pacific. The combined sequence of events is called the delayed oscillator mechanism. The period and amplitude of the ENSO cycle depends on the strength of the coupling, the size of the Pacific, and the delay of the Rossby waves. This mechanism is not complete, however. It explains the regular ENSO cycle but not its irregularity as evidenced in Fig. 3. This irregularity has variously been explained by the addition of random noise to the cycle, the nonlinear interaction of the ENSO cycle with the annual cycle, the nonlinear interaction of the ENSO cycle with other growing instabilities, and by many other possibilities. *See* OCEAN WAVES.

A completely different class of theories is that the interaction between the atmosphere and the ocean is not unstable. How can a stable system exhibit ENSO? Although the theory is mathematically recondite, there is a class of dynamics (nonnormal) that can exhibit a growth of disturbances followed by decay. Since all ENSO phases first grow and then decay, this has to be allowed as an explanation of ENSO. A stable coupled atmosphere–ocean system perturbed by random noise has been shown to have properties and predictive power currently indistinguishable from those of unstable theories.

Modeling ENSO. Since ENSO is a coupled atmosphere–ocean phenomenon, the system must be composed of a numerical model of the atmosphere suitably coupled to a numerical model of the ocean. If we want also to model the effects of ENSO on the rest of the world, the models must be global and include land, ice and snow, and so on. The model must be run for a very long time, and the statistics of the annual cycle and ENSO are recorded and compared to historical data such as in Fig. 3. This has been done to very good effect for simplified models where the annual cycle is specified and only anomalies from the annual cycle are calculated. But for fully coupled numerical models where the only input is from the Sun and all else is calculated, the simulation of ENSO has not been so good. The problem is that, due to systematic biases, the simulation of the annual cycle in the tropical Pacific is generally poorly done, and some of the higher-frequency phenomena (such as the intraseasonal oscillations) are also poorly done. These issues generally are agreed to be the dominant problem with the simulation of ENSO. This makes it difficult to examine the changes of ENSO with global warming, but it does not preclude effective predictions since the predictions are initialized by observations. While the systematic biases in the model may tend to pull the prediction back to unrealistic states over time, there will generally be a time in which the predictions are pretty good. It would still be of great value for both short-term predictions (a season or so in advance) and global-warming simulations to improve the models in the manner indicated.

Todd Mitchell; Edward S. Sarachik

Bibliography. K. Achuta Rao and K. R. Sperber, Simulation of the El Niño Southern Oscillation, *Clim. Dyn.*, 19:191–209, 2002; A. G. Barnston et al., Long-lead seasonal forecasts—where do we stand?, *Bull. Amer. Meteorol. Soc.*, 75:2097–2114, 1994; A. G. Barnston et al., Multimodel ensembling in seasonal climate forecasting at IRI, *Bull. Amer. Meteorol. Soc.*, 84:1783–1796, 2003; M. A. Cane, S. E. Zebiak, and S. C. Dolan, Experimental forecasts of El Niño, *Nature*, 321:827–832, 1986; H. R. Diaz and U. Markgraf (eds.), *El Niño: Historical and Paleoclimatic Aspects of the Southern Oscillation*, 1993; L. Goddard et al., Current approaches to seasonal-to-interannual climate predictions. *Int. J. Climatol.*, 21:1111–1152, 2001; T. N. Palmer et al., Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER), *Bull. Amer. Meteorol. Soc.*, 85:853–872, 2004; S. G. H. Philander, *El Niño, La Niña, and the Southern Oscillation*, 1990; C. F. Ropelewski and M. S. Halpert, North American precipitation and temperature patterns associated with the El Nino/Southern Oscillation (ENSO), *Mon. Weath. Rev.*, 114:2352–2362, 1986; J. M. Wallace et al., On the structure and evolution of ENSO-related climate variability in the tropical Pacific: Lessons from TOGA, *J. Geophys. Res. Oceans*, 103:14,241–14,260, 1998; S. E. Zebiak and M. A. Cane, A model El Niño/Southern Oscillation, *Mon. Weath. Rev.*, 115:2262–2278, 1987; Y. Zhang, J. M. Wallace, and D. S. Battisti, ENSO-like interdecadal variability 1900–93, *J. Clim.*, 10:1004–1020, 1997.

Elasmobranchii

The subclass within the Chondrichthyes (cartilaginous fishes) that includes the sharks (Euselachii) and the skates and rays (Batoidei). The other subclass within Chondrichthyes, according to traditional

classifications, is the Holocephali (chimaera, or ratfishes). It is probable that both groups arose independently during the Silurian or Early Devonian from a group of extinct armored fishes, the Placodermi. The elasmobranchs are distinguished by separate gill openings, amphistylic or hyostylic jaw suspension, and sensory ampullae (of Lorenzini) in the head region. Characters shared with the holocephalans include a variably calcified cartilaginous endoskeleton, placoid scales, urea-retention mechanism, clasper organs in the male for internal fertilization, and the absence of an air (swim) bladder. *See* CHIMAERIFORMES; CHONDRICHTHYES; PLACODERMI; SCALE (ZOOLOGY); SWIM BLADDER.

Evolutionary development. The history of the elasmobranchii can best be understood in terms of three successive evolutionary levels: cladodont, hybodont, and modern. Because of their incomplete fossil record, it is impossible to work out the detailed phylogenetic relationships of the extinct and living groups.

Cladodonts. The cladodont sharks first appeared in the Devonian; except for the specialized pleuracanths (**Fig. 1***a*), which persisted into the Triassic, they had disappeared by the end of the Pennsylvanian. The cladodonts all possess amphistylic jaw suspension. Their teeth have a pointed central cusp, two or more lateral cusps, and a flattened disklike base. The notochord was persistent and was not replaced by centra.

The basal part of the pectoral fin skeleton is variable in design but never tribasal. The radial cartilages of the pectoral and pelvic fins are not divided and extend more or less to the fin margin. The pelvic plates are separated. Claspers are usually present. The dorsal fins are preceded by spines. The caudal fin is heterocercal, nearly equilobate, and the radials of the hypochordal lobe are unsegmented. The body scales are usually multilobed, each lobe having a separate pulp cavity. Some cladodont sharks possess relatively large semicircular canals; in others these are the same size as in modern sharks. *See* COPULATORY ORGAN; EAR (VERTEBRATE).

Cladodonts were primarily pelagic predators, capable of seizing and tearing prey. The relatively stiff fins suggest that their maneuverability was more limited than in modern sharks. Representative cladodonts include the Devonian genera *Cladodus* and *Cladoselache* (Fig. 1*b*). *See* DEVONIAN.

Hybodonts. The hybodonts first appeared in the Mississippian Period. Although their skeleton resembles that of the cladodonts, they were more advanced in having only a tribasal pectoral fin skeleton, divided and shortened paired fin radials, clearly differentiated anal fin, and ribs and hemal arches along the entire length of the unreduced notochord. The acquisition of the tribasal pectoral fin, present in all later sharks, and a more flexible hypochordal lobe in the caudal fin apparently enabled these sharks to maintain their equilibrium more effectively than the cladodonts.

Some hybodonts had cladodont-like teeth and presumably fed on swimming prey, but much of the adaptive radiation at this level involved the development of crushing or grinding dentitions related mainly to bottom feeding. The late Palaeozoic genus *Ctenacanthus* is a good example of a conservative hybodont; the Mesozoic *Hybodus* (Fig. 1*c*) has a more advanced fin structure and a crushing dentition. *Ctenacanthus* lived from the Mississippian into the Permian, and *Hybodus* from the Triassic through the Cretaceous.

Modern elasmobranchs. Modern sharks arose during the Jurassic Period long before the hybodonts had become extinct in the Late Cretaceous. The transition from the hybodont to the modern level is poorly documented by fossils. However, there are three groups of aberrant (part hybodont, part modern) living sharks. Heterodontidae (Port Jackson sharks), Chlamydoselachidae (frilled sharks), and Hexanchidae (six-gill and seven-gill sharks). All except the Port Jackson sharks are rare deep-water forms, and each group has a somewhat different combination of hybodont and modern shark characters. These archaic sharks suggest the kind of evolutionary experimentation that took place during the rise of the modern sharks. *See* JURASSIC; SELACHII.

The skates and rays (batoids) are mostly bottom-dwelling elasmobranchs that possess many features distinct from the sharks. Present evidence suggests that the batoids arose during the Jurassic from some benthic hybodont group.

Modern features. Morphologically the modern shark level may be defined by the following features:

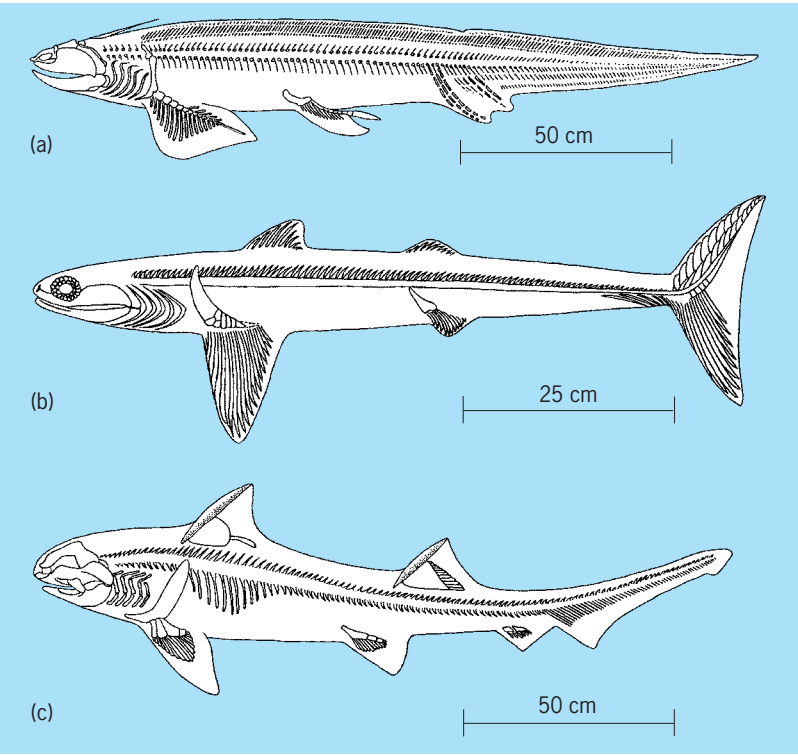


Fig. 1. Lateral views of fossil sharks. (*a*) *Pleurocanthus*, a specialized cladodont. (*b*) *Cladoselache*, a cladodont. (*c*) *Hybodus*, a hybodont.

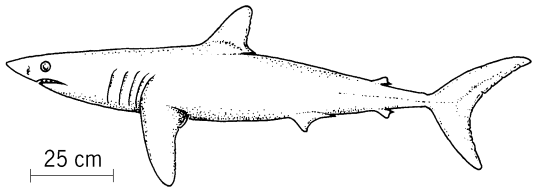


Fig. 2. Mackeral shark (*Isurus*). (*After H. B. Bigelow and W. C. Schroeder, Fishes of the Western North Atlantic, pt. 1, 1948*)

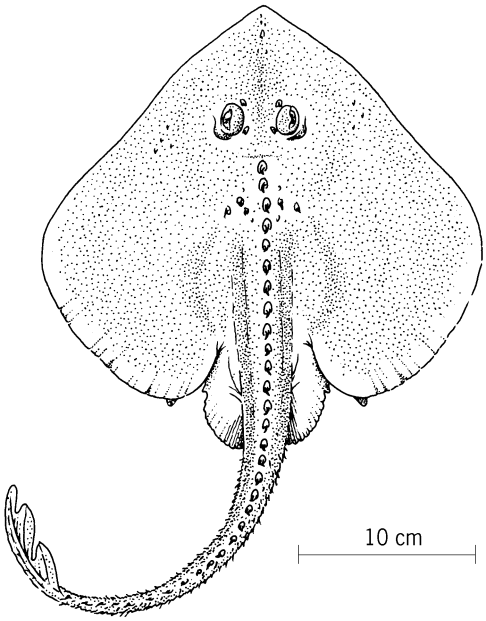


Fig. 3. Skate (*Raja*). (*After H. B. Bigelow and W. C. Schroeder, Fishes of the Western North Atlantic, pt. 2, 1953*)

hyostylic jaw suspension, shortened jaws, jaw protrusion mechanism, complete replacement of the notochord by calcified centra, fusion of the pelvic plates, and presence of only single-lobed placoid scales. Although the rostrum was short and blunt in the cladodonts, it became elongated in some hybodonts and, with a few exceptions, it projects well beyond the mouth in all modern sharks (**Fig. 2**).

Skates and rays differ from the sharks in having a depressed body, modification in the mode of gill ventilation, freer hyostylic jaw suspension, elaboration of the pavement dentition for crushing or grinding, enlargement of the pectoral fins, and disappearance of the anal fin. With the pectoral fins assuming the main role in locomotion, the posterior part of the body, including the caudal fin, became reduced to whiplike proportions (**Fig. 3**). *See* BATOIDEI; SELACHII.

Bobb Schaeffer

Bibliography. H. B. Bigelow and W. C. Schroeder, *Fishes of the Western North Atlantic*, pt. 1, 1948, pt. 2, 1953; C. E. Bond, *Biology of Fishes*, 2d ed., 1997; E. S. Herald, *Living Fishes of the World*, 1961; F. H. Pough et al., *Vertebrate Life*, 5th ed., 1998; B. Schaeffer, Comments on elasmobranch evolution, in P. W. Gilbert et al. (eds.), *Sharks, Skates and Rays*, 1967.

Elasticity

The property whereby a solid material changes its shape and size under the action of opposing forces, but recovers its original configuration when the forces are removed. The theory of elasticity deals with the relations between the forces acting on a body and the resulting changes in configuration, and is important in many branches of science and technology, for instance, in the design of structures, in the theory of vibration and sound, and in the study of the forces between atoms in crystal lattices.

Elastic constants. The forces acting on a body are expressed as stresses and measured as force per unit area. Thus if a bar $ABCD$ of square cross section (**Fig. 1***a*) is fixed at one end and subjected to a force F uniformly distributed over the other end DC, the stress is $F/(DC)^2$. This stress causes the bar to become longer and thinner and to assume the shape $A'B'C'D'$. The strain is measured by the ratio (change in length)/(original length), that is, by $(B'C' - BC)/(BC)$. According to Hooke's law, stress is proportional to strain, and the ratio of stress to strain is therefore a constant, in this case Young's modulus, denoted by E, so that $E = F(BC)/(DC)^2\ (B'C' - BC)$. *See* HOOKE'S LAW; STRESS AND STRAIN; YOUNG'S MODULUS.

Poisson's ratio σ is defined as the ratio of lateral strain to longitudinal strain, so that $\sigma = BC(DC - D'C')/DC\ (B'C' - BC)$. The bar of Fig. 1*a* is in a state of tension, and the stress is tensile; if the force F were reversed in direction, the stress would be compressive. Stresses of this type are called direct or normal stresses; a second type of stress, known as tangential or shear stress, is illustrated in Fig. 1*b*. In this case, the configuration $ABCD$ becomes $ABC'D'$, with the shear forces F acting in the directions AB and CD. The shear strain is measured by the angle θ, and if the body is originally a cube, the shear stress is $F/(DC)^2$. The ratio of stress to strain, $F/(DC)^2\theta$, is the shear or rigidity modulus G, which measures the

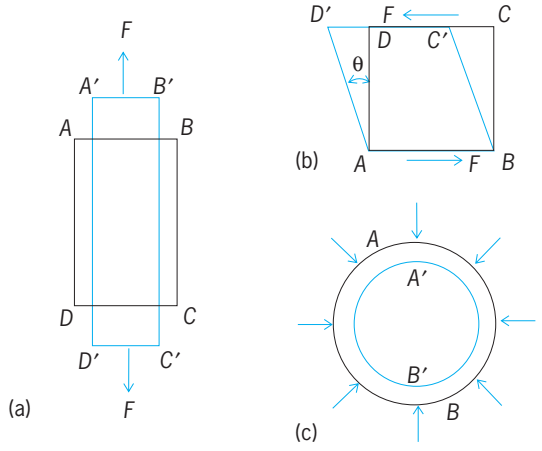


Fig. 1. Stresses on a bar. (*a*) Direct or normal stress. (*b*) Tangential or shear stress. (*c*) Change in volume with no change in shape. (All deformations are exaggerated.)

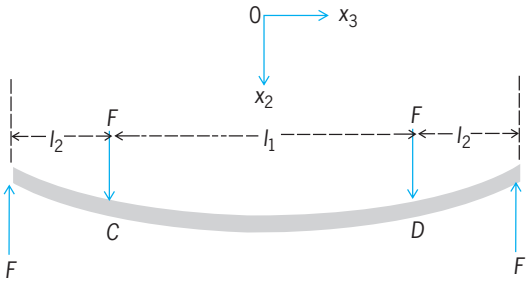


Fig. 2. Flexure and torsion in a bar.

resistance of the material to change in shape without change in volume.

A further elastic constant, the bulk modulus k, measures the resistance to change in volume without changes in shape, and is illustrated in Fig. 1*c*. The original configuration is represented by the circle AB, and under a hydrostatic (uniform) pressure P, the circle AB becomes the circle $A'B'$. The bulk modulus is then $k = P\upsilon/\Delta\upsilon$, where $\Delta\upsilon/\upsilon$ is the volumetric strain. The reciprocal of the bulk modulus is the compressibility.

Determination of values. The elastic constants may be determined directly in the way suggested by their definitions; for instance, Young's modulus can be determined by measuring the relative extension of a rod or wire subjected to a known tensile stress. Less direct methods are, however, usually more convenient and accurate. Prominent among these are the dynamic methods involving frequency of vibration and velocity of sound propagation. The elastic constants can be expressed in terms of frequency of (or velocity in) regularly shaped specimens, together with the dimensions and density, and by measuring these quantities, the elastic constants can be found. *See* ULTRASONICS.

The elastic constants can also be determined from the flexure and torsion of bars. As an illustration, consider a bar AB (**Fig. 2**) of breadth b (in the x_1 direction) and depth d (in the x_2 direction) supported by forces F at the ends, and loaded symmetrically by forces F at points C and D. Over the portion CD there is a uniform bending moment $M = Fl_2$, and the theory of bending shows that the portion CD is bent into the arc of a circle such that Eq. (1) applies, where R

$$R = \frac{EI}{M} \tag{1}$$

is the radius of curvature, E is Young's modulus, and I is the moment of inertia of cross section, equal to $bd^3/12$ for a rectangular cross section. The longitudinal stress at the lower face of the bar is tensile, and at the upper face, compressive. The middle plane of the bar is free of stress, and is the neutral axis. The stress at a distance x_2 from the neutral axis is shown in Eq. (2).

$$T = \frac{Ex_2}{R} = \frac{Mx_2}{I} \tag{2}$$

It is thus possible to determine E from Eq. (1) by measuring I, R, and M; conversely, if E is known, the stress may be determined from Eq. (2). *See* LOADS, TRANSVERSE.

Practical limitations. In practice, stress is only proportional to strain, and the strain is only completely recoverable within certain limits, called the elastic limits of the material. The stress below which the strain is completely recoverable is sometimes called the limit of perfect elasticity, and the stress up to which Hooke's law is obeyed is sometimes called the proportional limit or limit of linear elasticity. Above the elastic limits, the material is subject to time-dependent effects, and as the stress is further increased, the ultimate strength of the material is approached. *See* PLASTICITY; STRENGTH OF MATERIALS.

Theory of elasticity. In classical elasticity theory, it is assumed that the strains are always small; Hooke's law is therefore obeyed; the strains are completely recoverable and, moreover, are superposable, so that the strain produced by the joint action of two or more stresses is the sum of the strains produced by them individually.

In order to develop the theory, it is necessary to specify the stresses and strains more closely. **Figure 3** shows the stress components T_{ij} (where i, j may take the values 1, 2, or 3) acting on the faces of a cube, parallel to the coordinate axes x_1, x_2, x_3. The first suffix indicates the direction of the stress component and the second the direction of the normal to the plane under consideration. Stresses of the type T_{11} are normal stresses, and of the type T_{12}, shear stresses. The conditions for zero rotation of the cube are $T_{12} = T_{21}$, $T_{13} = T_{31}$, $T_{23} = T_{32}$, and there are therefore six independent stress components.

In addition to the stresses T_{ij}, body forces proportional to volume (for instance, forces due to the weight of the body) may also be acting. If the stresses T_{ij} vary with position, application of Newton's second law leads to Eq. (3) for the x_1 direction where ρ

$$\frac{\partial T_{11}}{\partial x_1} + \frac{\partial T_{12}}{\partial x_2} + \frac{\partial T_{13}}{\partial x_3} + X_1 = \rho f_1 \tag{3}$$

is the density, f_1 is the acceleration, and X_1 the body force component per unit volume along x_1, together with two similar equations for the x_2 and x_3 directions. If $f_1 = f_2 = f_3 = 0$, these equations become the equations of equilibrium, and if, further, $X_1 = X_2 = X_3 = 0$, they become the equations of equilibrium in the absence of body forces. The preceding

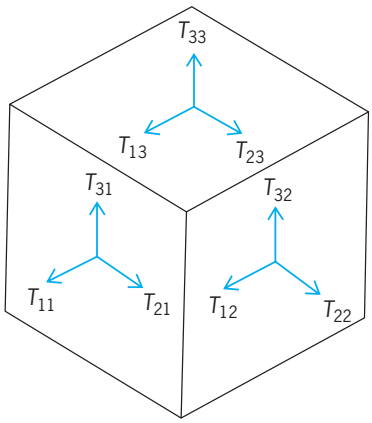


Fig. 3. Stress components acting on the faces of a cube.

equations are important in many branches of elastic theory and, for example, provide a starting point in the study of vibrating bodies and of the twisting of cylinders and prisms with cross sections of various shapes. *See* TORSION.

The components of strain are specified in a similar way to the stresses. There are six independent strain components: S_{11}, S_{22}, S_{33}, S_{23}, S_{13}, and S_{12}. If, as a result of strain, the coordinates of a point x_1, x_2, x_3 become x_1+u_1, x_2+u_2, x_3+u_3, the quantities u_1, u_2, and u_3 are the components of the displacement vector, and the strain components are as displayed in Eq. (4), so that, for example the relations shown by Eq. (5) would hold true.

$$S_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) \tag{4}$$

$$S_{11} = \frac{\partial u_1}{\partial x_1} \qquad S_{12} = \frac{1}{2}\left(\frac{\partial u_1}{\partial x_2} + \frac{\partial u_2}{\partial x_1}\right) \tag{5}$$

By eliminating the displacements from these equations, the so-called compatibility equations are obtained with three of the type shown in Eq. (6) and three of the type shown in Eq. (7).

$$\frac{\partial^2 S_{22}}{\partial {x_3}^2} + \frac{\partial^2 S_{33}}{\partial {x_2}^2} = 2\frac{\partial^2 S_{23}}{\partial x_2 \partial x_3} \tag{6}$$

$$\frac{\partial^2 S_{11}}{\partial x_2 \partial x_3} = \frac{\partial}{\partial x_1}\left(-\frac{\partial S_{23}}{\partial x_1} + \frac{\partial S_{13}}{\partial x_2} + \frac{\partial S_{12}}{\partial x_3}\right) \tag{7}$$

The stresses and strains have so far been denoted by two suffixes. This is essential if the methods of tensor analysis are to be applied to elasticity problems, but for many purposes a single suffix notation is adequate. The change from a two- to a one-suffix notation for the stresses is simply $T_{11}=T_1$, $T_{22}=T_2$, $T_{33}=T_3$, $T_{23}=T_4$, $T_{13}=T_5$, $T_{12}=T_6$. The change of notation for the strains is $S_{11}=S_1$, $S_{22}=S_2$, $S_{33}=S_3$, $2S_{23}=S_4$, $2S_{13}=S_5$, $2S_{12}=S_6$; the factor 2 is required to make the strains S_4, S_5, and S_6 conform with the usual definition of shear strain (Fig. 1*b*).

Hooke's law generalized. Hooke's law may be generalized to the statement that each stress component is proportional to each strain component, equivalent to the six equations (8*a*) which may be written more concisely as Eq. (8*b*), where the summation extends

$$\begin{aligned} T_1 &= c_{11}S_1 + c_{12}S_2 + c_{13}S_3 + c_{14}S_4 + c_{15}S_5 + c_{16}S_6 \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ T_6 &= c_{61}S_1 + c_{62}S_2 + c_{63}S_3 + c_{64}S_4 + c_{65}S_5 + c_{66}S_6 \end{aligned} \tag{8a}$$

$$T_q = \sum_r c_{qr}S_r \tag{8b}$$

over $r = 1, 2, 3, 4, 5,$ and 6. The elastic constants c_{qr} are termed the stiffnesses; there are altogether 36 of them but they are subject to the reciprocal relations $c_{qr}=c_{rq}$ imposed by thermodynamic requirements, and the number is thus reduced to 21.

Additional relations can be derived from the three assumptions that the interatomic forces act along the lines joining the centers of atoms in the lattice, that the atoms are situated at centers of symmetry, and that the lattice is initially at zero stress. These relations, called Cauchy relations, are $c_{23}=c_{44}$, $c_{13}=c_{55}$, $c_{12}=c_{66}$, $c_{14}=c_{56}$, $c_{25}=c_{46}$, $c_{45}=c_{36}$ and, if true, would reduce the number of stiffnesses to 15. Experiment shows, however, that they are not true in general; nevertheless, their investigation provides an indication of the extent to which these three assumptions hold in any particular case.

The generalized Hooke's law can also be written to express the strains in terms of the stresses given in Eq. (9), in which the quantities s_{qr} are the elas-

$$S_q = \sum_r s_{qr}T_r \qquad (r = 1, 2, 3, 4, 5, 6) \tag{9}$$

tic compliances. If the six simultaneous equations of Eq. (8) are solved for the strains, the compliances are obtained in terms of the stiffnesses as in Eq. (10), where Δc is the determinant shown in Eq. (11) and

$$s_{qr} = \frac{\Delta c_{qr}}{\Delta c} \tag{10}$$

$$\Delta c = \begin{vmatrix} c_{11} & c_{12} & c_{13} & c_{14} & c_{15} & c_{16} \\ c_{12} & c_{22} & c_{23} & c_{24} & c_{25} & c_{26} \\ c_{13} & c_{23} & c_{33} & c_{34} & c_{35} & c_{36} \\ c_{14} & c_{24} & c_{34} & c_{44} & c_{45} & c_{46} \\ c_{15} & c_{25} & c_{35} & c_{45} & c_{55} & c_{56} \\ c_{16} & c_{26} & c_{36} & c_{46} & c_{56} & c_{66} \end{vmatrix} \tag{11}$$

Δc_{qr} is the cofactor obtained by deleting the row and column containing c_{qr} from the determinant Δc.

The 21 stiffnesses (or compliances) of the generalized Hooke's law describe the elastic behavior of a material belonging to the triclinic crystal system. The existence of symmetry elements reduces the number of independent elastic constants in the other crystal systems to the following numbers: monoclinic, 13; orthorhombic, 9; tetragonal, 7 or 6; trigonal, 7 or 6; hexagonal, 5; and cubic, 3. Materials belonging to all of these systems are anisotropic, and the elastic properties depend upon direction within the material. If the properties are independent of direction, the material is isotropic and its elastic behavior is completely described by two independent stiffnesses (or compliances). *See* CRYSTALLOGRAPHY.

The stress-strain relations, referred to the principal axes in the orthorhombic, hexagonal, cubic, and isotropic systems, are given in the **table**. The equations involving the compliances are completely analogous, with S and T interchanged, and s_{qr} written for c_{qr}, except where $T_q = {}^1/_2\,(c_{11} - c_{12})\,S_q$, in which case $S_q = 2(s_{11} - s_{12})\,T_q$.

Rochelle salt is an example of an orthorhombic crystal; materials which, although not crystalline, possess the same symmetry and matrix of elastic constants as orthorhombic crystals are said to be orthotropic. Wood and plywood are materials of this description, and orthotropic elastic theory has also been applied to laminated plastics and reinforced concrete.

Single-crystal zinc, cobalt, magnesium, and ice are hexagonal materials; they are transversely isotropic because the properties are independent of direction

Stress-strain relations

Stress	Strain			
	Orthorhombic	Hexagonal	Cubic	Isotropic
$T_1 =$	$c_{11}S_1 + c_{12}S_2 + c_{13}S_3$	$c_{11}S_1 + c_{12}S_2 + c_{13}S_3$	$c_{11}S_1 + c_{12}S_2 + c_{12}S_3$	$c_{11}S_1 + c_{12}S_2 + c_{12}S_3$
$T_2 =$	$c_{12}S_1 + c_{22}S_2 + c_{23}S_3$	$c_{12}S_1 + c_{11}S_2 + c_{13}S_3$	$c_{12}S_1 + c_{11}S_2 + c_{12}S_3$	$c_{12}S_1 + c_{11}S_2 + c_{12}S_3$
$T_3 =$	$c_{13}S_1 + c_{23}S_2 + c_{33}S_3$	$c_{13}S_1 + c_{13}S_2 + c_{33}S_3$	$c_{12}S_1 + c_{12}S_2 + c_{11}S_3$	$c_{12}S_1 + c_{12}S_2 + c_{11}S_3$
$T_4 =$	$c_{44}S_4$	$c_{44}S_4$	$c_{44}S_4$	$(c_{11} - c_{12})S_4/2$
$T_5 =$	$c_{55}S_5$	$c_{44}S_5$	$c_{44}S_5$	$(c_{11} - c_{12})S_5/2$
$T_6 =$	$c_{66}S_6$	$(c_{11} - c_{12})S_6/2$	$c_{44}S_6$	$(c_{11} - c_{12})S_6/2$

in all of the planes normal to the hexagonal axis.

Single-crystal copper, gold, silver, nickel, and the alkali halides (for example, sodium chloride) are important cubic materials. The stress-strain equations are derived from those of the orthorhombic system by superimposing the condition that the three principal directions are all equivalent. This does not mean that the properties are independent of direction; for example, the compliance s_{11}' in an arbitrary direction is given by Eq. (12), where a_1, a_2, a_3 are the

$$s'_{11} = s_{11} - 2(s_{11} - s_{12} - s_{44}/2) \cdot \left(a_1^2 a_2^2 + a_2^2 a_3^2 + a_3^2 a_1^2\right) \quad (12)$$

cosines of the angles between the arbitrary direction and the cubic axes. This equation shows that s'_{11} depends on orientation unless $s_{44}/2 = s_{11} - s_{12}$. *See* PHOTOELASTICITY.

R. F. S. Hearmon

Bibliography. A. P. Boresi, R. J. Schmidt, and O. M. Sidebottom, *Advanced Mechanics of Material*, 5th ed., 1993; S. F. Borg, *Fundamentals of Engineering: Elasticity*, 1990; J. E. Marsden and T. J. Hughes, *Mathematical Foundations of Elasticity,* 1983, reprint 1994; S. P. Timoshenko and J. N. Goodier, *Theory of Elasticity,* 3d ed., 1970.

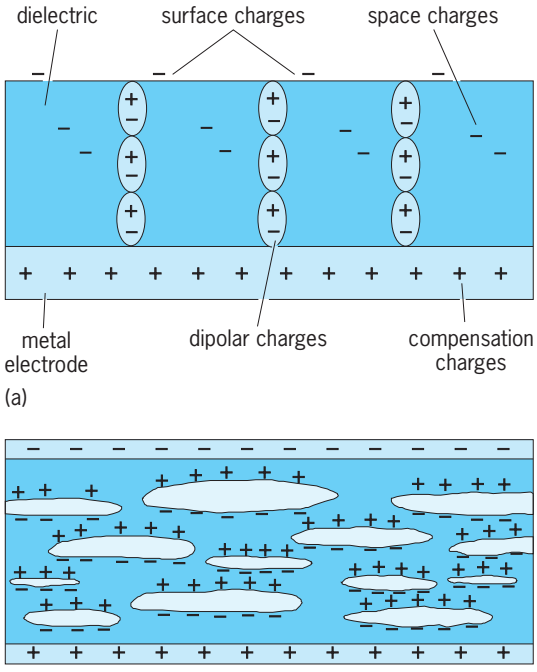
Electret

A solid dielectric with a quasipermanent electric moment. Electrets may be classified as real-charge electrets and dipolar-charge electrets. Real-charge electrets of present importance are either solid dielectrics with charges of one polarity at or near one side of the dielectric and charges of opposite polarity at or near the other side, or cellular dielectrics with charges of opposite polarity on the two sides of the voids. The dipolar-charge electrets are dielectrics with aligned dipolar charges. Some dielectrics are capable of storing both real and dipolar charges. Examples of charge arrangements in electrets are shown in the **illustration**. *See* POLARIZATION OF DIELECTRICS.

Modern electrets used in research and in applications are often films ranging in thickness from the submicrometer scale for silicon-based materials to sheets of 5–50-μm (0.2–2-mil) thickness (foil electrets) for suitable polymer materials. They are frequently metallized on one or both sides, depending on the intended use. In many applications of the solid real-charge electrets, the external fields of the charged dielectric are used and the films are metallized on only one side to allow the electric field to emanate from the nonmetalized side. In applications of the cellular real-charge electrets and the dipolar electrets, the piezoelectric effect generated by such samples is used and the films are metallized on both surfaces. *See* PIEZOELECTRICITY.

Solid real-charge electrets. In these electrets, the intrinsic conductivity and carrier mobility must be small enough that excess charges can persist for long periods of time. Suitable materials are the fluorocarbons polytetrafluoroethylene (PTFE) and its copolymer fluoroethylenepropylene (FEP), as well as polypropylene (PP), polycarbonate (PC), polyimide (PI), polyethylene terephthalate (PETP), and inorganic materials such as silicon dioxide and aluminum oxide. PETP can also assume a considerable dipole polarization. *See* POLYMER.

Charging of real-charge electrets, metallized on one side, can be achieved by exposure of the nonmetallized side to a corona discharge; by application



Schematic cross sections of electrets. (*a*) Solid electret disk metallized on one side with real charges (surface and space charges) and dipolar charges. (*b*) Cellular electret metallized on both sides with charges of opposite polarities on the upper and lower surfaces of the voids.

of an electric field at elevated temperatures through an electrode opposing the nonmetallized side of the dielectric; or by contacting the nonmetallized side with a biased, wet electrode. Another charging technique, allowing greater control than other methods, is electron injection with a partially penetrating electron beam. In all cases, carriers of one polarity are deposited onto or injected into the insulator and trapped, and a compensation charge of equal magnitude but opposite sign flows into the sample electrode. Thus, the charge arrangement exhibits an electric moment. Another method, applied to photoconductors metallized on both sides with transparent electrodes, consists in the application of an electric field during irradiation with ultraviolet or visible light. Carrier generation by the light and charge separation by the field yield the desired electric moment, resulting in a photoelectret. *See* CORONA DISCHARGE; PHOTOCONDUCTIVITY.

PTFE and FEP are electronegative and therefore hold negative charges better than positive ones. Real-charge densities of up to 5 millicoulombs/m^2 (3.2 microcoulombs/$in.^2$) can be stored in these polymers. Electrets are stable over long periods of time. At the customarily used charge densities of about 0.2 millicoulombs/m^2 (130 nanocoulombs/$in.^2$), corresponding to surface potentials of approximately 250 V in a 25-μm-thick (1-mil) sample, the time constant of the decay of the surface potential of PTFE and FEP is about 200 years if storage is at room temperature and low relative humidity under shielded conditions. This time constant drops to about 10 years at 50°C (122°F) and 90% relative humidity.

Silicon dioxide (SiO_2) electrets are preferably made of a thermally grown silicon dioxide layer, usually less than 1 μm (0.04 mil) thick, on a silicon substrate. Negative charges deposited onto the free oxide surface show better stability than negative volume charges or positive charges. Surface-charging techniques, such as the corona method, are therefore utilized. After charging, a compensation charge of opposite polarity is located at the silicon–silicon dioxide interface. Charge densities are generally chosen in the range 10–20 mC/m^2 (6–12 μC/$in.^2$), corresponding to surface potentials of 300–600 V on a 1-μm-thick layer. Time constants of the charge decay exceed those for PTFE and FEP.

Cellular real-charge electrets. These newer electrets with piezoelectric properties consist of a cellular dielectric, typically a polymer such as PP, with voids of the order of a few micrometers height and lateral dimensions of several tens of micrometers. Such voids can be generated by biaxially stretching a polymer film filled with small inorganic particles. During the stretching, microcracks form around the particles and develop into flat voids. These can be enlarged in the thickness direction by a pressure treatment, consisting of an increase of the pressure of the surrounding gas such that gas penetrates into the voided structure. A subsequent pressure release inflates the material. Thereafter, the sample is metalized on one of its surfaces.

Charging of such a sample is achieved by exposure of its unmetallized surface to a corona discharge. The depositied charges together with their countercharges on the metallization generate a high internal field that will cause breakdowns in the voids and thus deposition of charges of opposite sign on the internal void surfaces. Finally, the sample is metallized on the other surface. The material now exhibits a large piezoelectric d_{33} coefficient that is very stable at room temperature. Problematic is the thermal depolarization at elevated temperatures found in the PP samples. Other materials with higher depolarization temperatures are now under study. The samples also show hysteresis behavior as a function of the applied field and are therefore often referred to as ferroelectrets.

Dipolar-charge electrets. In these electrets, the relaxation time of the molecular reorientation has to be large enough that the dipole alignment persists for long periods of time. Most prominent among such materials are poly(vinylidenefluoride) [PVDF] and vinylidenefluoride-trifluoroethylene copolymers [VDF-TrFE], which exhibit piezoelectric and pyroelectric properties linked to the dipole polarization. Inorganic piezoelectric crystals and ceramics also belong to this category but are not discussed here. *See* PIEZOELECTRICITY; PYROELECTRICITY.

The preferred technique for the polarization of dipolar electrets is also the above corona method applied to one-sided metallized samples. Again, the injected charges set up a large field within the dielectric that aligns the dipoles. Another method consists in the direct application of a high field to a two-sided metallized sample at an elevated temperature. After the polarization process, the samples exhibit sizable d_{31} and d_{33} coefficients.

In the case of PVDF, polarization densities of up to 0.1 C/m^2 (65 μC/$in.^2$) can be achieved by corona charging at room temperature with fields of up to 400 MV/m (10 MV/in.) on samples prestretched three to five times their original length. The stretching makes the material, which is semicrystalline, recrystallize in a polar form and also reorients the lamellar crystals so that the dipoles are more precisely aligned by the applied field.

PVDF has a relatively stable polarization at room temperature. At 100°C (212°F), the polarization decays initially (within minutes) to half its original value and stabilizes thereafter. Similar stability is achieved for the piezoelectric activity of this material.

Comparison of piezoelectric properties. Typical d_{33} coefficients for cellular PP and solid PVDF are compared in the **table** with values for the conventional piezoelectric materials quartz and lead zirconate

Comparison of piezoelectric properties

Piezoelectric material	d_{31}, pC/N	d_{33}, pC/N
Quartz	0.73 (d_{14})	2.3 (d_{11})
PZT-5	170	370
PVDF	25	20
Cellular PP (expanded)	1	600

titanate (PZT-5) . Since the electret materials are polymers, they have certain other properties such as mechanical softness and flexibility that distinguish them from the conventional piezoelectrics. Particularly noteworthy is the huge d_{33} coefficient of cellular PP. The electrets have already found numerous commercial applications.

Applications. Important commercial applications of real-charge electrets are in electroacoustic and electromechanical transducers, air filters, electret dosimeters, and adhesive posters. Of particular significance are electret microphones, which are made in quantities of more than 2×10^9 annually and account for 90% of the microphones produced. Also of interest are biological applications based on the blood compatibility of charged polymers or on their favorable influence on wound or fracture healing. Commercial applications of dipolar electrets are in piezoelectric transducers for airborne and waterborne sound and in pyroelectric detectors. *See* AIR FILTER; DIELECTRIC MATERIALS; DOSIMETER; ELECTRET TRANSDUCER; ELECTRICAL INSULATION; MICROPHONE.

Gerhard M. Sessler

Bibliography. S. Bauer, R. Gerhard-Multhaupt, and G. M. Sessler, Ferroelectrets: Soft electroactive foams for transducers, *Phys. Today*, 57(2): 37–43, 2004; Institute of Electrical and Electronics Engineers, *IEEE 12th International Symposium on Electrets*, 2005; V. Kestelman, L. Pinchuk, and V. Goldade, *Electrets in Engineering*, Kluwer Academic, 2000; G. M. Sessler and R. Gerhard-Multhaupt (eds.), *Electrets*, vols. 1 and 2, 3d ed., Laplacian Press, 1999.

Electret transducer

A device for the conversion of acoustical or mechanical energy into electrical energy, and vice versa, which utilizes a quasipermanently charged or polarized dielectric material (electret). Examples are certain microphones, headphones, hydrophones, and ultrasonic devices. Depending on the electret material and the usage, either the external electric field of real-charge electrets or the piezoelectric activity of dipolar electrets is used. In the simplest implementation, a transducer based on real-charge electrets consists of a metal backplate (first electrode) covered by a mechanically tensioned diaphragm. The diaphragm is a film electret carrying a metal coating (second electrode) on the side facing away from the backplate. Provisions are made to maintain a shallow air gap between electret and backplate. The air gap is occupied by an electrostatic field originating from the electret charges. Upon acoustical or mechanical deflection of the diaphragm, such a device generates an electrical output signal between its two electrodes; similarly, application of an electrical signal results in diaphragm deflections. Such electret devices are therefore self-biased electrostatic or condenser transducers. They thus exhibit all the advantages of this transducer class, such as wide dynamic range and flat response over a frequency range of several decades, without requiring the external bias necessary in conventional transducers of this kind. On the other hand, a transducer based on the piezoelectric activity of a dipolar electret may consist of just the electret film covered by two electrodes. Compression or bending of the film causes again an electrical output signal between the electrodes, while the application of a voltage to the electrodes results in mechanical action. *See* ELECTRET; ELECTROSTATICS; TRANSDUCER.

Film electret. Suitable polymer films are electrically charged or polarized to produce an external electric field or piezoelectric activity, respectively. Materials capable of quasipermanently trapping charge carriers (that is, storing charges over periods of years or decades) include polytetrafluoroethylenepropylene (PTFE), its copolymer polyfluoroethylenepropylene (FEP or Teflon®), and polypropylene (PP), while substances capable of carrying a quasipermanent dipole polarization include poly(vinylidenefluoride) [PVDF] and some of its copolymers. A polymer of considerable interest is cellular PP, which can carry charges on the surfaces of its internal voids and therefore exhibits a very large piezoelectric activity. The cellular PP loses its charges at elevated temperatures and is therefore suitable only for room-temperature use. Before charging or poling, all the materials are either metallized on one side or backed up with a metal electrode.

Electret microphones. The most widely used electret transducer is the electret microphone. This is a condenser microphone consisting of two electrodes separated by an electret and an air gap. The electret generates an electric field in the air gap. If one of the electrodes is designed as a diaphragm that vibrates in response to an incident sound wave, an output signal proportional to the sound pressure is generated.

Such microphones are designed as transducers with a backplate electret or a diaphragm electret. The backplate-electret microphones, which are the more common transducers, consist of a (6–12-μm) (0.25–0.5-mil) Teflon layer adhering to a metallic or metal-coated backplate and charged to 200–400 microcoulombs/m^2 (130–260 nanocoulombs/in.^{2}), corresponding to an external bias of about 100 V. The surface of this electret is placed next to a metallic or metallized diaphragm, leaving a shallow air layer of about 20 μm (0.8 mil) thickness. The stiffness of the air layer can be decreased (and thus the sensitivity of the microphone can be improved) by connecting the air layer to a larger cavity by means of small holes through the backplate. The electrical output of the microphone is taken between the backplate, which is insulated from the outer case, and the metal part of the diaphragm. The output is fed into a high-impedance preamplifier. The diaphragm-electret microphones typically consist of a 12–25-μm-thick (0.5–1-mil) Teflon or PTFE film, metallized on one side and charged on the other side to charge densities similar to the above-mentioned values. A backplate and an intermediate air gap complete the system. These transducers are simpler and less expensive than the backplate-electret microphones but mechanically not as stable. This is attributed to stress relaxation of the tensioned Teflon diaphragm, which

causes small changes in sensitivity.

At frequencies below the resonance frequency, the acoustical properties of electret microphones are largely governed by the restoring force on the diaphragm. Since the mechanical tension of the diaphragm is generally kept at a relatively low value (about 10–50 newtons/m or 0.7–3.5 lb/ft), the restoring force is determined by the compressibility of the air layer. Controlling the restoring force by the air layer is advantageous because changes in tension due to stress relaxation thus have only a minor effect on the sensitivity, which is largely independent of transducer area.

Under open-circuit conditions, a displacement d of the diaphragm of an electret microphone causes a frequency-independent output voltage given by Eq. (1), where E is the (constant) electric field in

$$v = Ed = \frac{\sigma D_1 d}{\varepsilon_0(D_1 + \varepsilon D_2)} \tag{1}$$

the air layer of the transducer; σ is the charge density of the electret projected onto its surface; D_1 and D_2 are the thicknesses of electret film and air layer, respectively; ε_0 is the permittivity of free space; and ε is the relative dielectric constant of the electret material. As in conventional electrostatic transducers, the displacement is proportional to the applied pressure in a wide frequency band extending from a lower cutoff given by a pressure-equalization leak in the back cavity to an upper cutoff determined by the resonance frequency. The voltage response for constant sound pressure is frequency-independent in this range.

Typical electret microphones designed for the audio-frequency range have constant sensitivities of 10–50 mV/pascal (1–5 mV/microbar or 70–350 V/lb · in.$^{-2}$) in the frequency range 20 to 15,000 Hz. Nonlinear distortion is less than 1% for sound-pressure levels below 140 dB, and the impulse response is excellent, owing to the flat amplitude and phase characteristics. Other properties of electret microphones are their low sensitivity to vibration, owing to the small diaphragm mass, and their insensitivity to magnetic fields. Compared with conventional electrostatic transducers, electret microphones have the following advantages: they do not require a dc bias; they have three times higher capacitance per unit area, resulting in a better signal-to-noise ratio; and they are not subject to destructive arcing between foil and backplate in humid atmospheres and under conditions of water condensation.

Various electret microphones for operation at infrasonic and ultrasonic frequencies, covering the range from 0.001 Hz to 200 MHz, have been designed by properly positioning the upper and lower cutoff frequencies discussed above. Furthermore, transducers with directional characteristics such as cardioid, bidirectional, toroidal, and second-order unidirectional have been developed. *See* MICROPHONE.

Electret headphones. Another group of electret transducers made of real-charge electrets are headphones. The above microphone designs can also be utilized in this case. These transducers produce a certain amount of nonlinear distortion due to the partially quadratic dependence of the force F per unit area, exerted on the diaphragm, on the applied ac voltage v, as can be seen from Eq. (2). High-voltage

$$F = \frac{-\frac{1}{2}(D_1\sigma + \varepsilon\varepsilon_0 v)^2}{\varepsilon_0(D_1 + \varepsilon D_2)^2} \tag{2}$$

sensitivities have been achieved by making the air-gap thickness small. The distortion can be eliminated by the use of push-pull transducers and monocharge electrets.

For a push-pull transducer, a diaphragm consisting of two electrets, each metallized on one side and with these metal layers in contact, is sandwiched between two perforated metal electrodes, forming a symmetrical system. Application of a signal voltage $\bar{v}$ in antiphase to the two electrodes causes the electrodes to exert forces F_1 and F_2 on the diaphragm. These forces are given by Eq. (2) if $\pm\bar{v}$ is substituted for v. The net force $F = F_1 - F_2$ on the diaphragm is determined by Eq. (3), which is linear in $\bar{v}$, indicating the absence of nonlinear distortion. The linearity, however, is maintained only if both air-layer thicknesses remain at the same value D_2. This is possible only if D_2 is relatively large, which makes such systems less sensitive than the above-described single-ended transducers.

$$F = \frac{-2\bar{v}\sigma\varepsilon D_1}{(D_1 + \varepsilon D_2)^2} \tag{3}$$

This transducer can be simplified and improved by using an asymmetric system in which a single electret coated on one side with a low-conductivity layer is used. The diaphragm holds an excess charge because of the presence of an induction charge on the front electrode. Thus, application of an ac signal causes a net force between diaphragm and plates that is linear if the resistivity of the coating (or the resistor R) prevents charge equalization during a period of lowest-frequency signal applied.

An interesting modification of the above push-pull transducers can be achieved by using nonmetallized monocharge electrets, which differ from customary electrets in that they carry only a single-polarity charge and no compensation charge. It is possible to form these electrets by electron injection or by the use of wet-contact charging methods. Since a compensating charge is absent, monocharge electrets have a strong external field that is independent of electrode distance. Thus, if such electrets are used in transducer configurations, large air gaps of the order of 1 mm (0.04 in.) can be used without loss in field strength. This in turn results in highly efficient transducers that can operate at large displacement amplitudes. *See* EARPHONES.

Electromechanical transducers. Apart from their use in electroacoustic transducers, such as microphones and earphones, real-charge electrets have been applied to electromechanical transducers. Examples are touch or key transducers, impact-sensitive line transducers, relay switches, and optical display panels.

Relatively close in their design principles to the

electroacoustic transducers described above are the first two of these groups. The touch or key transducers depend on the manual deflection of an electret relative to an electrode in a microphonelike structure. Large-voltage signals can be generated with such devices. Line transducers consist essentially of a coaxial cable with polarized dielectric in which the center conductor and the shield serve as electrodes. Mechanical excitation resulting in a deformation of the shield at any point along the length of such a cable produces an electrical output signal.

Very different design principles are used in the relay switches and the optical display panels. The relay switches utilize the external fields of electrets mounted in bistable arrangements to open or close contacts. Such devices have power requirements 1000 times smaller than equivalent electromagnetic systems; thus considerable power conservation over present magnetic components is possible. A similar design is used in the optical display panels where illuminated channels are opened and closed by hinged and opaque electrets moved by electrostatic forces. Although relatively high voltages are required, the necessary currents are small and flow only during the switching operation. *See* ELECTRONIC DISPLAY; RELAY.

Hydrophones based on PVDF. Many types of hydrophones based on the piezoelectric polymer PVDF and its copolymers have been described. Among these are hydrophones for the sonic and near-ultrasonic frequency ranges based on the use of thick sheets or cylinders of the piezoelectric material, as well as hydrophones for the higher ultrasonic range, generally designed as membrane or needle hydrophones.

The lower-frequency hydrophones can be relatively simple devices consisting of an electroded film of the piezoelectric polymer that is encapsulated for waterproofing. If the dimensions of the hydrophone are small compared to the wavelength of the sound, equal sound pressure in all directions prevails and the output signal of the transducer is given by the hydrostatic piezoelectric constant. More sophisticated designs include stiffening plates to avoid resonances due to extensional modes of the polymer film, two-layer constructions to improve shielding, and cylindrical designs. Arrays or transducers with shaped elements allow directional characteristics to be obtained. Hydrophones based on these principles generally operate in the frequency range from less than 1 kHz to several MHz.

The membrane hydrophones for the higher ultrasonic range, up to 100 MHz, also consist of one or two peripherally clamped thin sheets of piezoelectric polymer with a diameter of several centimeters, poled in a tiny central section of less than 1 mm (0.04 in.) in diameter, which is the actual sensing area. These hydrophones have the advantage of being more or less transparent to sound, thus allowing sound pressure to be measured without disturbing the sound field by generating unwanted reflections. In the needle hydrophones, a metallic needle serves as the support and inner electrode for a surrounding layer of the piezoelectric polymer, which is covered on its outer side with another electrode. The outside diameter of the needle is generally less than 1 mm (0.04 in.) to 0.3 mm (0.012 in.), thus permitting again measurements without perturbation of the sound field. The frequency range of needle hydrophones can extend up to 80 MHz. *See* HYDROPHONE.

Piezoelectric cellular transducers. The very large piezoelectric thickness coefficient of charged cellular PP can be used to sense sound, force, impact, vibration, or motion. For example, cellular PP has been used in microphones of very simple design. Such transducers consist essentially of only the PP film and suitable housing and shielding. If stacked layers of film are used, the microphone sensitivity increases proportionally to the number of layers. Using this technique, microphones consisting of six layers have been built that show sensitivities of 15 mV/Pa, comparable to the sensitivity of conventioanl electret microphones. Other features of these transducers are relatively flat frequency response, low noise level, and low distortion. The major advantage is that a minute and critical air gap, as used in electret microphones, is not needed. Cellular PP can also be used for sound generation. Since the generated sound pressure increases with frequency, sound sources are particularly suitable for the ultrasonic range, below the resonance frequency of the PP film that is in the region above 100 kHz.

Important electromechanical transducers based on cellular PP are piezoelectric floor mats, keyboards, control panels, pickups for musical instruments, and biomedical sensors. The floor mats are step- or impact-sensitive and can detect the motion of patients in hospitals, of intruders in industrial plants or military installations, or of vehicles on highways. Because of the inexpensive material, large floor areas may be covered. Push-buttons of cellular PP for keyboards and control panels have the advantage of high sensitivity and mechanical flexibility. The biomedical sensors are used to monitor, for example, heartbeat, breathing, and forces on limbs in humans and animals.

Applications. Because of their favorable properties, simplicity, and low cost, electret transducers have been used in many applications as research tools and in the commercial market.

Among the research applications are microphones for use in acousto-optic spectroscopy, applied to the detection of air pollution and to the study of reaction kinetics of gases and optical absorption of solids. Because of the favorable noise performance of electret microphones, the detection threshold for air pollutants has been lowered by more than an order of magnitude. Other applications of electret microphones have been in aeronautics and shock-tube studies, in which the low vibration sensitivity of these transducers is crucial. The wide frequency range of electret transducers, discussed above, made possible their application in infrasonic atmospheric studies and in ultrasonic investigations of liquids and solids. A broad field of application is in noise

monitoring and noise abatement. Examples are all kinds of mobile and stationary noise measurements as well as active noise reduction. In addition, audio and ultrasonic arrays of electret microphones have been used in near-field acoustic holography. Yet other uses of electromechanical electret transducers are in such diverse areas as microforce detection, vibration analysis, and leak detection in space stations. For example, microforce sensors based on PVDF are used in the micromanipulation and microassembly of MEMS (micro-electro-mechanical systems) structures. In all these applications, the simplicity and reliability of electret transducers are of importance. *See* ACOUSTICAL HOLOGRAPHY; ACTIVE SOUND CONTROL; INFRASOUND; MICRO-ELECTRO-MECHANICAL SYSTEMS (MEMS); NOISE MEASUREMENT; PHOTOACOUSTIC SPECTROSCOPY; SHOCK TUBE; ULTRASONICS.

Of all commercial applications of electret devices, the electret microphone for amateur, professional, studio, and telephone use is most prominent. About 2×10^9 such microphones are produced annually, accounting for about 90% of the entire output of microphones. Almost half of these are used in stationary and mobile telephones. Other uses of electret microphones are in high-fidelity work, such as audio studios, and in MP-3 players, sound-level meters, noise dosimeters, camcorder, and toys. Miniature microphones are used in hearing aids, in telephone operator headsets, and in communication headsets, often as noise-canceling (first-order gradient) transducers. The success of the electret microphone is primarily due to its acoustic quality and low cost. It is noteworthy that, owing to their low vibration sensitivity, electret microphones were the first transducers to be built directly into widely used tape recorders. In addition to electret microphones, there are single-backplate and push-pull electret headphones, also with monocharge electrets. The latter are among the highest-quality headsets available. *See* HEARING AID; MAGNETIC RECORDING; SOUND RECORDING; TELEPHONE.

Apart from their use in electroacoustics, electret devices have found many commercial applications in the underwater, ultrasonic, and electromechanical fields. The PVDF hydrophones in the varieties discussed above for the sonic and ultrasonic frequency ranges have been commercially available for a long time and, because of their many advantages, enjoy an undisputed position as the preferred devices in the field. The same holds for the ultrasonic devices for airborne sound based on space-charge and piezoelectric polymers. In the electromechanical field, transducers based on PVDF and on cellular piezoelectric PP have been introduced commercially. In particular, step-sensitive floor mats, touch-sensitive keyboards and control panels, and biomedical force sensors are available.

Gerhard M. Sessler

Bibliography. Institute of Electrical and Electronics Engineers, *IEEE 12th International Symposium on Electrets*, 2005; G. M. Sessler and R. Gerhard-Multhaupt (eds.), *Electrets*, vols. 1 and 2, 3d ed., Laplacian Press, 1999.

Electric charge

A basic property of elementary particles of matter. One does not define charge but takes it as a basic experimental quantity and defines other quantities in terms of it. The early Greek philosophers were aware that rubbing amber with fur produced properties in each that were not possessed before the rubbing. For example, the amber attracted the fur after rubbing, but not before. These new properties were later said to be due to "charge." The amber was assigned a negative charge and the fur was assigned a positive charge.

According to modern atomic theory, the nucleus of an atom has a positive charge because of its protons, and in the normal atom there are enough extranuclear electrons to balance the nuclear charge so that the normal atom as a whole is neutral. Generally, when the word charge is used in electricity, it means the unbalanced charge (excess or deficiency of electrons), so that physically there are enough "nonnormal" atoms to account for the positive charge on a "positively charged body" or enough unneutralized electrons to account for the negative charge on a "negatively charged body."

The rubbing process mentioned "rubs" electrons off the fur onto the amber, thus giving the amber a surplus of electrons, and it leaves the fur with a deficiency of electrons. *See* ELECTROSTATICS.

In line with the previously mentioned usage, the total charge q on a body is the total unbalanced charge possessed by the body. For example, if a sphere has a negative charge of 1×10^{-10} coulomb, it has 6.24×10^8 electrons more than are needed to neutralize its atoms. The coulomb is the unit of charge in the SI system of units. *See* COULOMB'S LAW; ELECTRICAL UNITS AND STANDARDS.

The surface charge density σ on a body is the charge per unit surface area of the charged body. Generally, the charge on the surface is not uniformly distributed, so a small area ΔA which has a magnitude of charge Δq on it must be considered. Then σ at a point on the surface is defined by the equation below.

$$\sigma = \lim_{\Delta A \to 0} \frac{\Delta q}{\Delta A}$$

The subject of electrostatics concerns itself with properties of charges at rest, while circuit analysis, electromagnetism, and most of electronics concern themselves with the properties of charges in motion. *See* CAPACITANCE; ELECTRIC CURRENT; ELECTRICITY.

Ralph P. Winch

Electric contact

A part, in an electrical switching device, made of conducting material, for the purpose of closing, opening, or changing the conductive path of an electrical circuit. To open or close a circuit, an electric contact is made to come in contact with or separate from its mating part. Devices embodying contacts for these

purposes are electric switches, relays, contactors, and circuit breakers. Contacts may be actuated directly or through a linkage that is driven either manually, mechanically, electromagnetically, hydraulically, or pneumatically.

An electric contact is also an essential part of a variable resistor. In such an application the contact or wiper provides a moving connection to the resistive element. For such service the contact material is chosen for low abrasion combined with absence of a high-resistance film such as would be formed by an adherent oxide. *See* POTENTIOMETER; RHEOSTAT.

Requirements for contacts differ markedly depending on applications. For contacts in relays and low-power applications, reliability in completing a circuit may be of the utmost importance. Surface films and contaminants are not tolerable. Noble metals such as gold and platinum are sometimes sealed hermetically for this purpose. For the purpose of carrying high continuous currents, electrical resistance must be kept to a minimum. Contacts with high silver content are usually used. High contact forces, up to hundreds of pounds, may be applied when the contacts are in the closed position. For contacts that must interrupt high currents, refractory materials such as tungsten and molybdenum are often used. These materials also possess antiwelding characteristics, which are essential for contacts that must close in on short-circuit currents and later open under normal operating force.

On many occasions all the aforementioned properties are desired in one material. Special alloys such as silver cadmium oxide or refractory materials impregnated with silver have been developed for this purpose.

A related electrical part is the brush on rotating machines. Pairs of brushes provide continuity between the external circuit and slip rings or commutator segments on the rotating portion of the machine. Similar arrangements of brushes and slip rings connect to continuously rotating radar antenna mounts. Graphite, because of its inherent lubricating nature, is a usual constituent of such sliding contacts for high-current operation; corrosion-resistant metals or metal alloys are used for low-current applications. *See* CIRCUIT (ELECTRICITY); ELECTRIC ROTATING MACHINERY.

Thomas H. Lee

Bibliography. ARS Electronics, *Receiving Tube Characteristics and Socket Connection Guide*, 1981; CES Industries Inc. Staff, *Contactor Sensor Operation*, 1982; R. Holm and E. Holm, *Electric Contacts*, 4th ed., 1967; Institute of Electrical and Electronics Engineers, *1998 IEEE 44th Holm Conference on Electrical Contacts*, 1998.

Electric current

The net transfer of electric charge per unit time. It is usually measured in amperes. The passage of electric current involves a transfer of energy. Except in the case of superconductivity, a current always heats the medium through which it passes. For a discussion of the heating effect of a current *see* JOULE'S LAW.

On the other hand, a stream of electrons or ions in a vacuum, which also may be regarded as an electric current, produces no local heating. Measurable currents range in magnitude from the nearly instantaneous 10^5 or so amperes in lightning strokes to values of the order of 10^{-16} A, which occur in research applications.

All matter may be classified as conducting, semiconducting, or insulating, depending upon the ease with which electric current is transmitted through it. Most metals, electrolytic solutions, and highly ionized gases are conductors. Transition elements, such as silicon and germanium, are semiconductors, while most other substances are insulators.

Electric current may be direct or alternating. Direct current (dc) is necessarily unidirectional but may be either steady or varying in magnitude. By convention it is assumed to flow in the direction of motion of positive charges, opposite to the actual flow of electrons. Alternating current (ac) periodically reverses in direction.

Conduction current. This is defined as the transfer of charge by the actual motion of charged particles in a medium. In metals the current is carried by free electrons which migrate through the spaces between the atoms under the influence of an applied electric field. Although the propagation of energy is a very rapid process, the drift rate of the individual electrons in metals is only of the order of a few centimeters per second. In a superconducting metal or alloy the free electrons continue to flow in the absence of an electric field after once having been started. In electrolytic solutions and ionized gases the current is carried by both positive and negative ions. In semiconductors the carriers are the limited number of electrons which are free to move, and the "holes" which act as positive charges.

Displacement current. When alternating current traverses a capacitor, there is no physical flow of charge through the dielectric (insulating material, gas, or vacuum), but the effect on the rest of the circuit is as if there were a continuous flow. Energy can pass through the capacitor by means of the so-called displacement current, and this displacement current generates a surrounding magnetic flux, just as an orthodox current would. James Clerk Maxwell introduced the concept of displacement current in order to make complete his theory of electromagnetic waves. *See* ALTERNATING CURRENT; CAPACITANCE; CONDUCTION (ELECTRICITY); DIELECTRIC MATERIALS; DIRECT CURRENT; DISPLACEMENT CURRENT; ELECTRIC INSULATOR; ELECTRICAL RESISTANCE; FREE-ELECTRON THEORY OF METALS; SEMICONDUCTOR; SUPERCONDUCTIVITY.

John W. Stewart

Electric distribution systems

Systems that comprise those parts of an electric power system between the subtransmission system and the consumers' service switches. It includes distribution substations; primary distribution feeders;

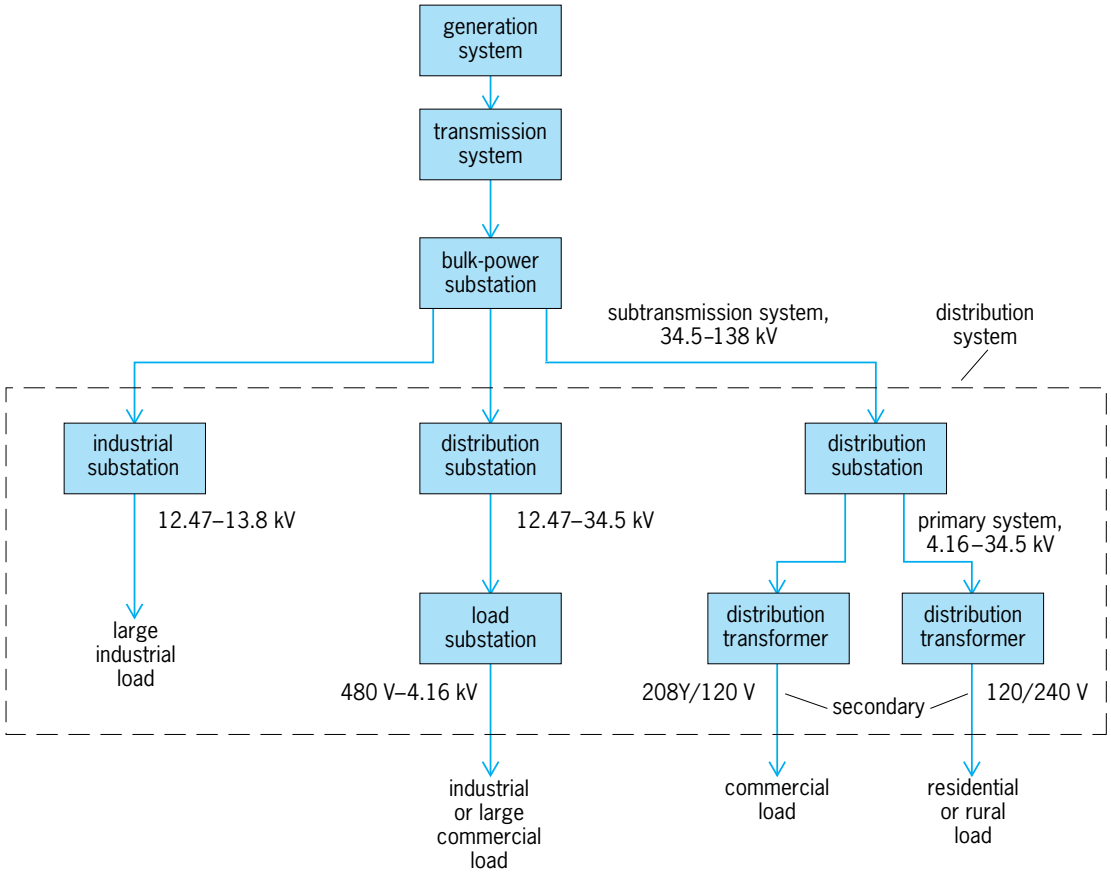


Fig. 1. Overview of the power system from generation to consumer's switch.

distribution transformers; secondary circuits, including the services to the consumer; and appropriate protective and control devices. Sometimes, the subtransmission system is also included in the definition.

The subtransmission circuits of a typical distribution system (**Fig. 1**) deliver electric power from bulk power sources to the distribution substations. The subtransmission voltage is usually between 34.5 and 138 kV. The distribution substation, which is made up of power transformers together with the necessary voltage-regulating apparatus, bus-bars, and switchgear, reduces the subtransmission voltage to a lower primary system voltage for local distribution. The three-phase primary feeder, which usually operates at voltages from 4.16 to 34.5 kV, distributes electric power from the low-voltage bus of the substation to its load center, where it branches into three-phase subfeeders and three-phase and occasionally single-phase laterals. Most of the three-phase distribution system lines consist of three-phase conductors and a common or neutral conductor, making a total of four wires. Single-phase branches (made up of two wires) supplied from the three-phase mains provide power to residences, small stores, and farms. Loads are connected in parallel to common power-supply circuits. *See* ALTERNATING CURRENT.

Subtransmission. The subtransmission system is that part of the power system which delivers power from bulk power sources, such as large transmission substations. The subtransmission may be made up of overhead open-wire construction on wood poles or underground cables. Although the voltages of these circuits can occasionally range from 12.47 to 345 kV, the majority are at the 69-, 115-, and 138-kV voltage levels. There is a trend toward the use of the higher voltages as a result of the increasing use of higher transmission voltages. The subtransmission-system designs vary from simple radial systems to a subtransmission network. The major considerations affecting the design are cost and reliability. *See* CONDUCTOR (ELECTRICITY); ELECTRIC POWER TRANSMISSION.

Substation. The distribution substation is an assemblage of equipment for the purpose of switching, regulating, and changing the supply voltage from subtransmission to primary distribution. More important substations are designed so that the failure of a piece of equipment in the substation or one of the subtransmission lines to the substation will not cause an interruption of power to the customer. The location of a substation is dictated by the voltage levels, voltage regulation considerations, subtransmission costs, substation costs, and the costs of primary feeders, mains, and distribution transformers. It is also restricted by esthetic considerations. A typical substation may have power transformers, circuit breakers, disconnecting switches, station buses and insulators, current-limiting reactors, shunt reactors, current transformers, potential transformers, capacitor voltage transformers, coupling capacitors,

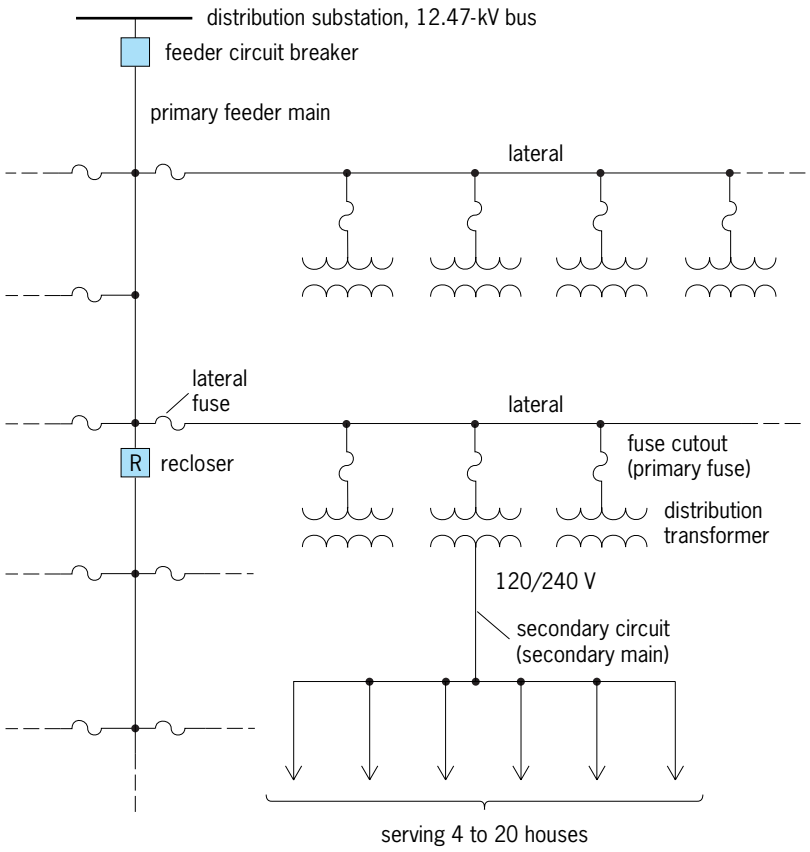


Fig. 2. Typical power distribution feeder.

protective relays, lightning arresters, station batteries, and other equipment. *See* ELECTRIC POWER SUBSTATION.

Primary system. The part of the electric utility system which is between the distribution substation and distribution transformers is called the primary system. It is made up of circuits known as primary feeders or distribution feeders. A feeder includes the main or main feeder, which usually is a three-phase four-wire circuit, and branches or laterals, which usually are single-phase or three-phase circuits, tapped off the main (**Fig. 2**). A feeder is usually sectionalized by means of reclosing devices at various locations so as to remove as little as possible of a faulted circuit and to hinder service to as few consumers as possible. This goal is achieved through the coordination of all the fuses and reclosers.

The primary system leaving the substation is most often in the 11,000–15,000-V range. A specific voltage used is 12,470-V line-to-line and 7200-V line-to-neutral (conventionally written 12,470Y/7200 V). Some utilities still use a lower voltage, such as 4160Y/2400 V. However, the use of primary distribution circuits in the 25- and 35-kV classes is increasing; all are four-wire systems. Single-phase loads are connected line-to-neutral on the four-wire systems. Congested and heavy-load locations in metropolitan areas as well as new residential areas are customarily served by underground primary feeders made up of radial three-conductor cables.

The primary system design is affected by the cost, importance of the load, and the required level of service continuity and reliability. The simplest and least expensive, and thus the most common, is the radial design. The low reliability of this design is improved by a modification called the loop-type primary design, in which the feeder loops through the feeder load area and returns to the substation bus. The most expensive and reliable design is the primary-network type, in which a system of interconnected feeders is supplied by a number of substations. Such a system supplies a load from several directions and consequently provides the maximum reliability and quality of service.

Secondary system. This is the part of the electric power system between the primary system and the consumer's property. The secondary distribution system includes distribution transformers, secondary circuits (secondary mains), consumer services (or service drops), and meters to measure consumer energy consumption. Generally, the secondary systems are designed in single phase for areas of residential customers and in three phase for areas of industrial or commercial customers with high load densities. Secondary voltages are provided by distribution transformers that are connected to the primary system, and they usually are associated with utilization voltages. Residential and most rural customers are supplied by 120/240-V single-phase three-wire systems. Commercial and small industrial customers are supplied by 208Y/120-V or 480Y/277-V three-phase four-wire systems. The secondary voltage also supplies multiple street lights.

Spot networks. These are used in downtown areas for high-rise buildings with extremely high load densities, and sometimes also for areas of industrial or commercial customers. Often, large shopping centers are supplied by spot networks. In such a network, all transformers and protecting apparatus are placed at the same location.

Good voltage. Essentially, good voltage means that the average voltage level is correct, that variations are within acceptable voltage limits, and that sudden and momentary changes in voltage level do not cause objectionable light flicker. Of course, the utilization voltage of the customer changes with changing load on the system as well as the location of the customer on the system. However, voltage variation is usually less than 5% at the customer's meter.

Good continuity. Service continuity is the provision of uninterrupted electric power to the customer. Good continuity means that such service is almost always provided.

Back-up systems. To increase the service reliability for critical loads, such as hospitals, computer centers, and crucial industrial loads, back-up systems, such as emergency generators or batteries, with automatic switching devices are provided. Computer installations are backed up by uninterruptible power supplies (UPS). *See* UNINTERRUPTIBLE POWER SUPPLY.

Overhead construction. Most existing distribution systems in residential, industrial, and rural areas are

of overhead construction. The distribution transformers are mounted near the tops of poles, and bare primary and secondary conductors are strung from pole to pole.

Underground construction. Almost all residential developments now being built are served by underground residential distribution (URD) systems using underground cables for esthetic reasons. The URD system is free from service outages and accompanying repair expenses caused by lightning, rain, ice, sleet, snow, and wind storms. Also, the probability of hazard caused by building fires and by motor vehicles and other foreign objects coming into contact with the system elements is far less than what might be expected on overhead systems. Other costs, such as the cost of tree trimming, are totally eliminated by using the URD systems. In general, underground distribution systems cost more than comparable overhead systems. However, greater public interest in the esthetic appearance of residential communities and reductions in the cost of underground equipment and installations have made the use of URD systems popular. Residential underground systems normally do not have the refinements and operating advantages found in conventional downtown commercial underground systems. Extensive duct banks, transformer vaults, access holes, and submersible equipment are not required in residential systems.

Cables. In the past the most frequently used underground cable was stranded copper with an oil-impregnated paper insulation wrapped in a lead sheath. It is expensive to buy, install, and splice, but it has a long service life. It is available in single- and three-phase cable and in voltages up to 46 kV. Practical residential underground distribution was made possible by the development of direct-buried underground cable that used a relatively inexpensive plastic insulation, typically cross-linked polyethylene (XLPE), and practical elbow connectors. The insulated cable has stranded conductors surrounding a bare concentric neutral. Other insulating plastics include high-molecular-weight polyethylene (HMWP), sometimes called high-density track-resistant polyethylene (HDPE), and ethylene propylene rubber (EPR). The use of such plastics as the protective sheath for cables has made practical their direct burial in the ground without the need for ducts and access holes. Splicing procedures for such cables are also simple and less costly. Long lengths of cable capable of being plowed directly into the ground or placed in narrow and shallow trenches naturally reduce installation and maintenance costs. Further savings may be realized if other facilities, such as telephone, cable television, gas, and water pipes, are installed simultaneously.

System design. In general, the heavy three-phase feeders are located overhead along the periphery of a residential development, and the laterals to the pad-mount transformers are buried about 40 in. (100 cm) deep. The lateral cable conductors come up from underground into the high-voltage compartment of the distribution transformer. The secondary service lines then run to the individual dwellings at a depth of about 24 in. (60 cm), and come up into the dwelling meter through a conduit. The service conductors run along easements and do not cross adjacent property lines.

Faults. The frequency of faults is lower on underground systems than on overhead systems. Also, faults are not so likely to "bunch up" due to storm conditions. However, faults are much more difficult and time-consuming to find, isolate, and repair on underground systems. Service-restoration requirements have dictated primary-system designs which operate as a normally open loop. In the event of a cable fault, such a design facilitates the location and isolation of the failure and faster service restoration to all customers on the unfaulted segment of the primary loop. A well-designed URD system should permit ease of sectionalizing in order to isolate a faulted portion, and should be flexible enough to accept load growth with a minimum number of changes. The looped primary and single-phase banked secondary system provide these requirements at the lowest cost. The banked secondary system takes better advantage of the diversity among loads and allows the use of transformers with a smaller kilovolt-ampere capacity and possibly fewer transformers. It also offers better average voltage conditions at the loads and less voltage dip due to motor starting for a given transformer rating, conductor size, and spacing between transformers than the simple-radial system offers. *See* VOLT-AMPERE.

Transformers. The transformers used for the URD system are hermetically sealed against moisture, including their bushings and terminals. The terminals may contain one or more insulated disconnecting elbows, enabling the disconnection of the transformer and the sectionalizing of the primary circuits. The transformers may be completely buried or installed on ground-level pads or semiburied pads. Pad-mounted transformers are the predominant type of transformer in URD systems. They are usually installed on concrete slabs, or pads. They may utilize "dead-front" configuration with separable insulated connectors, or elbows. Many other combinations of pad-mount construction and accessory equipment are available. Residential subsurface transformers are used much less frequently. They are installed in relatively tight-fitting vaults with the cover grating of the vault at ground level. Direct-buried transformers are not popular because of the problem of location of accessory equipment and its operation under adverse weather conditions such as snow and ice. *See* TRANSFORMER.

Dispersed storage and generation. Electric distribution systems are undergoing major changes to accommodate the development of alternative sources for generating electric energy and enhanced opportunities for small power producers and cogenerators. Usually, these generators are small (typically ranging from 10 kW to 10 MW and connectable to either side of the meter) and can be economically connected only to the distribution system. If properly planned and operated, dispersed storage and generation may provide benefits to distribution

systems by reducing capacity requirements, improving reliability, and reducing losses. Dispersed-storage-and-generation technologies include hydroelectric, diesel generators, wind-electric systems, solarelectric systems, batteries, storage space and water heaters, storage air conditioners, hydroelectric pumped storage, photovoltaics, and fuel cells. *See* ELECTRIC POWER SYSTEMS; ENERGY SOURCES.

Turan Gönen

Bibliography. American National Standards Institute, *Preferred Voltage Ratings for A-C Systems and Equipment: Guide for (EEI R-6-1949) (IEC38)*, ANSI C84.1, 1989; M. T. Bishop, A. G. Jones, and W. F. Israel, Overcurent protection alternatives for underground distribution systems, *IEEE Trans. Power Delivery*, 10(1):69–77, January 1995; T. Gönen, *Electric Power Distribution System Engineering*, 1986; A. J. Pansini, *Guide to Electrical Power Distribution Systems*, 1992; W. L. Weeks, *Transmission and Distribution of Electrical Energy*, Harper and Row, New York, 1981.

Electric field

A condition in space in the vicinity of an electrically charged body such that the forces due to the charge are detectable. An electric field (or electrostatic field) exists in a region if an electric charge at rest in the region experiences a force of electrical origin. Since an electric charge experiences a force if it is in the vicinity of a charged body, there is an electric field surrounding any charged body.

Field strength. The electric field intensity (or field strength) **E** at a point in an electric field has a magnitude given by the quotient obtained when the force acting on a test charge q' placed at that point is divided by the magnitude of the test charge q'. Thus, it is force per unit charge. A test charge q' is one whose magnitude is small enough so it negligibly alters the field in which it is placed. The direction of **E** at the point is the direction of the force **F** on a positive test charge placed at the point. Thus, **E** is a vector point function, since it has a definite magnitude and direction at every point in the field, and its defining equation is (1).

$$\mathbf{E} = \frac{\mathbf{F}}{q'} \tag{1}$$

Ralph P. Winch

Principle of superposition. The field at a distance of r meters from a point charge of q coulombs in vacuum is given by Eq. (2), where **E** is the field

$$\mathbf{E} = \frac{q}{4\pi\epsilon_0 r^2} \tag{2}$$

strength in volts per meter and $\epsilon_0 = 8.85 \times 10^{-12}$ farad/meter is the permittivity of free space. **E** is a vector, directed along the radius vector from the point charge; positive **E** is directed away from a positive charge or toward a negative charge. For an assembly of charges, the resultant field is, by the principle of superposition, the vector sum of the field components due to the individual charges. This summation may be performed directly, but in practical cases Gauss' theorem often affords a more powerful and convenient method. *See* GAUSS' THEOREM.

A. E. Bailey

Electric displacement. Electric flux density or electric displacement **D** in a dielectric (insulating) material is related to **E** by either of the equivalent equations (3), where **P** is the polarization of the

$$\mathbf{D} = \epsilon_0\mathbf{E} + \mathbf{P} \qquad \mathbf{D} = \epsilon\mathbf{E} \tag{3}$$

medium, and ϵ is the permittivity of the dielectric which is related to ϵ_0 by the equation $\epsilon = \epsilon_r\epsilon_0$, ϵ_r being the relative permittivity of the dielectric. In empty space, $\mathbf{D} = \epsilon_0\mathbf{E}$. The units of **D** are coulombs per square meter.

In addition to electrostatic fields produced by separations of electric charges, an electric field is also produced by a changing magnetic field. The relationship between the **E** produced and the rate of change of magnetic flux density $d\mathbf{B}/dt$ which produces it is given by Faraday's law of induced electromotive forces (emfs) in Eq. (4), where $d\mathbf{s}$ is a vector element

$$\oint \mathbf{E} \cdot d\mathbf{s} = -\int_A \frac{d\mathbf{B}}{dt} \cdot d\mathbf{A} \tag{4}$$

of path length directed along the path of integration in the general sense of **E**. Thus $\oint \mathbf{E} \cdot d\mathbf{s}$ is the emf induced in this closed path of integration. The area of the surface bounded by the path of integration is A, and the direction of $d\mathbf{A}$, an infinitesimal vector element of this area, is the direction of the thumb of the right hand when the fingers encircle the path of integration in the general sense of **E**. The right side of Eq. (4) is seen to be the negative of the time rate of change of the magnetic flux linking the path of integration chosen for the left side.

In an electrostatic field, $\oint \mathbf{E} \cdot d\mathbf{s}$ is always zero. *See* ELECTRIC CHARGE; ELECTROMAGNETIC INDUCTION; POTENTIALS.

Ralph P. Winch

Bibliography. D. Halliday, R. Resnick, and J. Walker, *Fundamentals of Physics*, 7th ed., 2005; E. M. Purcell, *Electricity and Magnetism*, 2d ed., 1985; H. D. Young and R. A. Freedman, *Sears and Zemansky's University Physics*, 11th ed., 2003.

Electric filter

A transmission network used to selectively modify the components of a signal according to their frequencies. In most cases a filter is used to enhance signals of desired frequencies while suppressing signals of undesired frequencies. An ideal filter would pass only desired frequencies while completely suppressing all unwanted frequencies, without any dispersion in time of the frequencies. Unfortunately, ideal filters are impossible to achieve.

Applications. Electric filters are used in most electronic communication systems. Whether communication is over wire, free space, or optical fiber, multiple channels of information can be multiplexed on different frequency bands. Unwanted signals

and noise are introduced along the communications path. The main function of electric filters is to separate the desired signal or channel from all others and from any noise or interference. For example, an AM radio receiver may have a low-pass filter after the antenna to separate the AM frequency band from higher frequency bands and, elsewhere in the radio, a band-pass filter to select the desired station out of the AM band. *See* ELECTRICAL COMMUNICATIONS; MULTIPLEXING AND MULTIPLE ACCESS; RADIO RECEIVER.

Although electronic filters are commonly thought of as devices for conferring selectivity to communication paths, they are used in almost every part of electronic equipment, such as the damping element in phase-locked loops, cleanup devices for frequency sources, and pulse expansion and compression devices for radar. One of the simplest and most common filters is the bypass capacitor used to restrict high-frequency electronic noise. *See* ELECTRICAL NOISE; PHASE-LOCKED LOOPS; RADAR.

Transfer function. The performance of a filter is, as is any transmission network, characterized by a transfer function. A transfer function is the ratio of the output signal to the input signal. In many filter types, including those composed of lumped elements, the transfer function can be expressed as a ratio of two polynomials, shown in Eq. (1). Here H is the transfer function, a_i and b_i are the coefficients of the two polynomials, and $s = j\omega$, ω being the frequency in radians and $j = \sqrt{-1}$. The zeros of the denominator of the transfer function are called the poles or natural frequencies of the filter. The zeros of the numerator are referred to as the transmission zeros or occasionally finite attenuation poles. Attenuation is the reciprocal of the transfer function or $1/H(s)$. Both the poles and the zeros of the filter can be complex and, for stability considerations, the poles must lie in the left half of the complex s-plane. *See* COMPLEX NUMBERS AND COMPLEX VARIABLES.

$$H(s) = \frac{b_0 + b_1 s + b_2 s^2 + \cdots + b_n s^n}{1 + a_1 s + a_2 s^2 + \cdots + a_n s^n} \qquad (1)$$

The characteristics of a filter can be analyzed from its transfer function. The transfer function can be expressed in a polar form, shown in Eq. (2), where

$$H(j\omega) = |H(j\omega)| e^{j\theta(\omega)} \qquad (2)$$

$|H(j\omega)|$ is the magnitude and $\theta(\omega)$ is the phase function. The square of the magnitude is equal to the ratio of the output power to maximum available power. The power gain of the filter, expressed in decibels, is given in Eq. (3). Loss would be defined as the negative of gain. For example, the gain function of a band-pass filter approximates the ideal amplitude response, wherein the gain is constant within the passband and frequencies outside the passband are completely suppressed (**Fig. 1***a*). The phase function of a typical band-pass filter decreases as the frequency is increased through the passband (Fig. 1*b*). Another important parameter is group delay (Fig. 1*c*), defined in Eq. (4). Group delay is used to characterize how different frequencies are dispersed in time. Digital data are often transmitted by using shifts in frequency or phase. If the delay variation is too great,

$$\alpha(\omega) = 20 \log_{10} |H(j\omega)| \quad \text{dB} \qquad (3)$$

$$\tau(\omega) = \frac{d\theta(\omega)}{d(\omega)} \qquad (4)$$

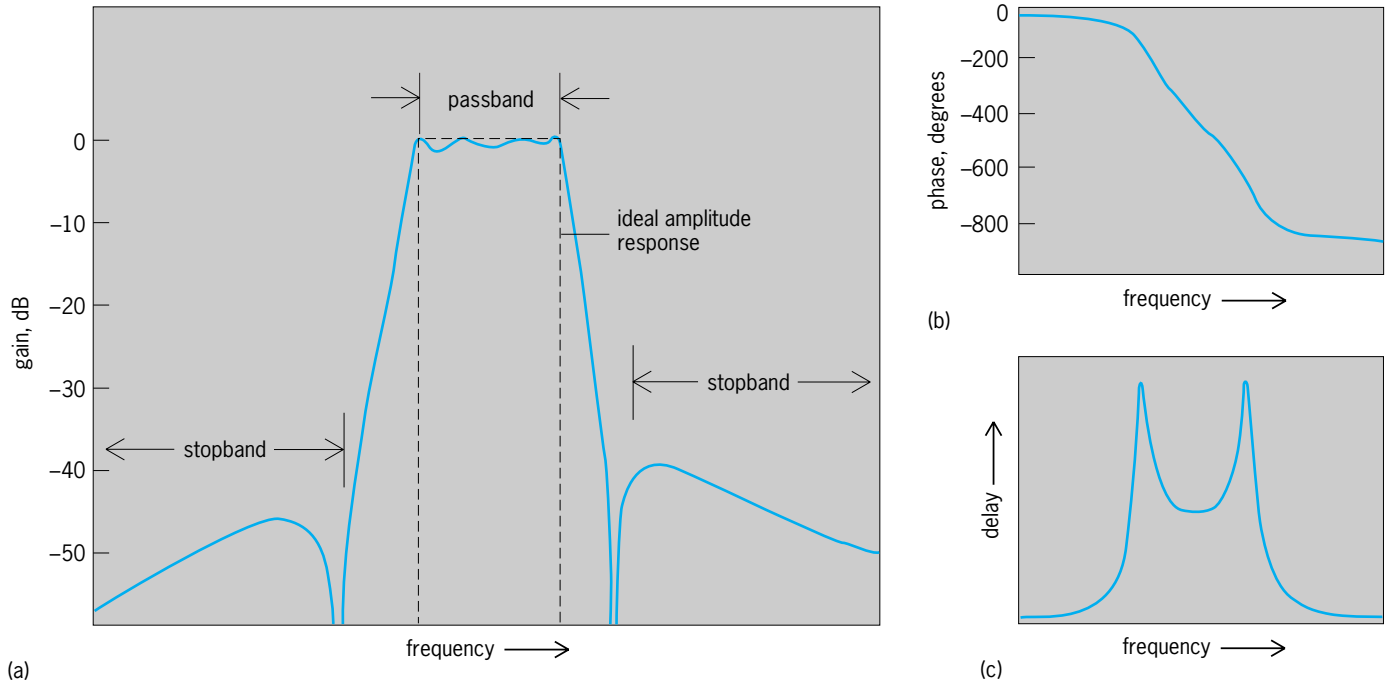


Fig. 1. Characteristics of a typical band-pass filter. (*a*) Gain characteristics, compared with those of an ideal filter. (*b*) Phase plot. (*c*) Delay plot.

these shifts will be detected incorrectly. *See* DECIBEL; GAIN; RESPONSE.

Characteristics. Filters are characterized in multiple ways. The expression low-pass, Butterworth, *LC* describes a filter. The descriptior low-pass indicates the relation of the passed to the rejected frequencies. Butterworth describes the type of polynomials in the transfer function. *LC* indicates the construction method. This filter is made of inductors (*L*'s) and capacitors (*C*'s).

Filter types. Filters are classified by the relationship of the frequencies that are selectively passed, referred to as the passband, to those which are attenuated, referred to as the stopband. An ideal low-pass filter passes all frequencies below a specified cutoff frequency and rejects those above. A high-pass filter does the opposite. An ideal band-pass filter will pass a band of frequencies while rejecting all others (Fig. 1*a*); a band-reject filter will reject a band of frequencies and pass all others.

An all-pass filter passes all frequencies but does, however, modify the time delay characteristics. It normally corrects delay distortions caused by other sections of a communication path.

All the above classifications are based on frequency-domain considerations. In addition, there are two terms that apply to the time-response characteristics of a filter. A finite impulse response (FIR) filter, when exposed to a change in input, will settle to a steady state within a finite amount of time. An infinite impulse response (IIR) filter will continue oscillating in a decaying manner forever.

A further consideration in classifying a filter is whether the frequency response is constant in time or varies. If it varies with time, as a function of the input signal, it is known as an adaptive filter. This type of filter finds use in speech and image enhancement and echo cancellation.

Approximations. Realizable filters are only approximations to the ideal. An ideal filter has instantaneous cutoff from the passband to the stopband. It passes all the desired frequencies with equal amplitude and delay at each frequency. Actual filters have a band of frequencies where the filter transitions from the passband to the stopband. The passband amplitude may not be flat but can have ripple (Fig. 1*a*). In the stopband of a filter that includes attenuation poles, the amplitude response will flare back up between those attenuation poles. The filter can also be dispersive, where different frequencies have different time delays.

The transfer function is the ratio of two polynomials. At least one of these polynomials would require an infinite number of terms to meet the instantaneous cutoff requirements of the ideal filter. However, there are several classes of polynomial that do approximate the ideal filter. These functions are used to synthesize actual filters.

The Chebyshev approximation to the ideal filter is a polynomial function optimized so that for a given stopband rejection the maximum variation in the passband is a minimum. The numerator of the transfer function of a Chebyshev filter would be 1, and it would have equal ripple in the passband; that is, all the peaks in the amplitude would be of the same height and all the dips of the same depth. An elliptic-integral filter, also known as a Cauer filter, has both equal ripple in the passband and equal minima of attenuation in the stopband; that is, the heights of the flarebacks between the attenuation poles are all the same. Both the numerator and denominator polynomials of the transfer function are of order greater than zero. The passband has the same criteria as the Chebyshev and, in addition, the minimum attenuation in the stopband is maximized. For polynomials of a given order, Chebyshev and Cauer filters provide the least ripple for the maximum stopband selectivity. Higher ripple can be traded for an increase in selectivity. However, an increase in delay variation and a degradation in stability will also occur. Increased stability and lower delay variation is obtained by allowing the ripple to decrease. By decreasing it in a Chebyshev design to zero, where the derivatives of $H(s)$ at center frequency become zero, a Butterworth filter is obtained. The Butterworth polynomial is the most stable. Changes in its coefficients or the element values used to realize Butterworth response cause the least changes in ripple of all the polynomial types.

In many applications the time-domain characteristics are as important as those of the frequency domain. In radar the key information is provided by pulse peaks. In many digital systems the information is encoded in the phase of a signal. If all the frequencies passing through a filter have the same time delay, the signal shape will remain unchanged and the encoded information will be recoverable. A filter with a gaussian-function amplitude response will have a linear phase and a constant delay characteristic. The normalized transfer function for a low-pass filter is givenby Eq. (5). This function is not a polynomial

$$|H(j\omega)| = e^{-\omega^2/2} \tag{5}$$

and not physically realizable. Approximations that try to match the amplitude or delay response of this function are known collectively as gaussian filters. The Bessel filters, also known as Thomson filters, are members of this group. They have maximally flat delay in the sense that the derivatives of delay at center frequency vanish. Other members of this group try to match the gaussian function only in the passband portion of their response. This characteristic allows increased selectivity in the stopband. Others try to provide constant delay in the Chebyshev sense by having equal delay ripple. *See* DISTRIBUTION (PROBABILITY).

Filter realization. Because filters are used over wide frequency and bandwidth ranges and with such varying performance criteria, many methods have been devised for creating a filter function (see **table**).

A term used to measure the quality of components in passive filters is Q. The quantity Q is defined as the ratio of energy lost in one cycle to total energy stored. It is a measure of how much loss an element has. An inductor with an inductance of L henrys and

Application ranges of various filter types

Filter type	Frequency range	Bandwidth (% of center frequency)
LC	0–5 GHz	>0.5
Crystal	5 kHz–300 GHz	0.01–5
Ceramic	300 kHz–50 MHz	0.5–20
Mechanical	1 kHz–700 kHz	0.02–10
Surface-acoustic-wave (SAW) resonator	30 kHz–2 GHz	0.01–0.1
SAW transversal	30 kHz–2 GHz	0.2–50
Active	<100 kHz	>0.5
Switched capacitor	<40 kHz	>0.5
Digital	<10 MHz	>0.1
Microwave	0.5 GHz–200 GHz	>0.5
Ceramic dielectric resonator (CDR)	0.7 GHz–5 GHz	>0.1

series parasitic resistance of R ohms will have a Q equal to $\omega L/R$, where ω is the angular frequency. Filter elements with low Q values increase the loss of a filter and limit the minimum bandwidth that may be achieved. *See* Q (ELECTRICITY).

LC and RC filters. These filters are based on lumped element inductors (L's), capacitors (C's), and resistors (R's). The transfer function for a typical *RC* filter (**Fig. 2***a*), which contains one pole at $-1/RC$ on the real axis and no zeros, is given in Eq. (6). Because the

$$H(s) = \frac{1}{1 + sCR} \tag{6}$$

poles of *RC* filters can be located only on the negative real axis, only low-pass responses can be achieved. To achieve complex poles and zeros, inductors must be used (Fig. 2*b*). Inductors have relatively low Q's, and at low frequencies are bulky. The Q's of typical inductors range from ten to a few thousand, and the larger the Q, the larger the inductor. Many of the other filter types were created to avoid the problems with inductors.

Most *LC* filters are designed as ladder structures with resistive termination. A ladder structure consists of alternating series and shunt arms (**Fig. 3***a*). The poles and zeros of the transfer function are provided by resonant circuits within the arms. The ability to control these resonances gives the ladder structure stability in both the passband and stopband. An alternate to a ladder would be a multipath structure. To achieve high rejection in the stopband requires that the amplitude and phase be closely controlled at the points where the paths join in order that the signals from the separate circuits will cancel. Therefore components must have tighter tolerances than those needed to realize a ladder design. Therefore, ladder structures tend to predominate in *LC* filter design. In crystal-filter design, however, because of the closeness of the impedance pole and zero of crystals, multipath (mainly lattice) structures are common. *See* RESONANCE (ALTERNATING-CURRENT CIRCUITS).

In a ladder structure a transfer-function pole occurs if a series arm has an impedance zero or a shunt arm has an impedance pole. Transfer-function zeros occur where a series impedance arm has a pole, or a shunt arm a zero. A capacitor connected in series or parallel with an inductor would be the simplest circuit for creating either an impedance zero or a pole. An impedance structure that contains both a finite-frequency pole and a zero consists of a capacitor (C_1), in series with an inductor capacitor, (L_1) and a second capacitor (C_2), connected in parallel (Fig. 3*b*). This circuit also contains a pole at zero frequency and a zero at infinite frequency.

An *LC* structure is created by choosing a transfer function, of the form shown in Eq. (1), to meet the filter's specification. This function may be a Chebyshev or other standard type or one created by the designer that is optimum for the specification. By finding the zeros of the numerator and denominator polynomials of the transfer function, the numbers of transfer-function poles and zeros at zero, finite, and infinite frequency are determined. At this point an *LC* circuit structure can be synthesized. By factoring out poles or zeros sequentially from the transfer function, the type of arm and component values required can be determined. As the process continues, the ladder gains arms as the transfer function describing the residual portion of the circuit is reduced in complexity.

Acoustic filters. This class of filters includes crystal, ceramic, mechanical, and surface-acoustic-wave (SAW) filters. These devices convert electrical energy

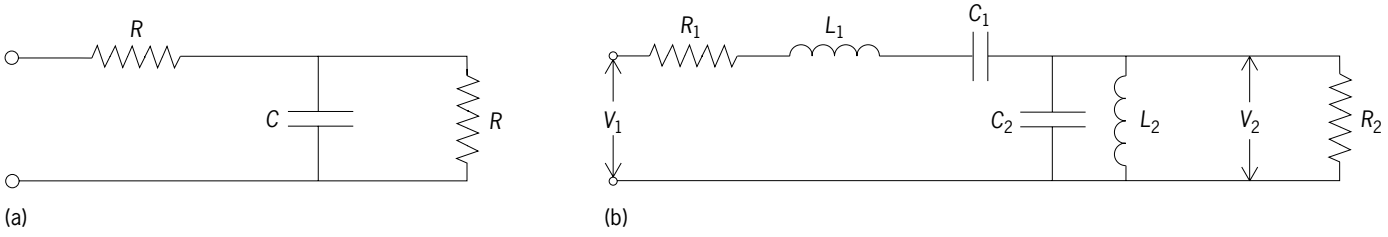


Fig. 2. Passive, lumped-element filters. (*a*) RC low-pass filter. (*b*) *LC* band-pass filter.

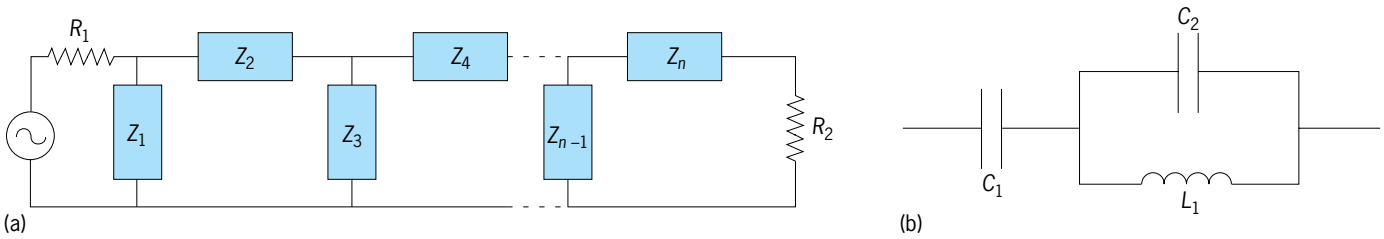


Fig. 3. Ladder structure. (*a*) Overall structure with alternating series and shunt arms. (*b*) LC circuit whose impedance contains both a finite-frequency pole and a zero. Such circuits are used in the series and shunt arms of the ladder structure.

to mechanical vibrations, process the signal acoustically, and then convert the energy back to an electrical form. The equations describing a mechanically vibrating resonator, where energy is cycled between kinetic motion and stress, match those of an inductor and capacitor (*LC*) attached in parallel, where energy is cycled between the electric field of the capacitor and the magnetic field of the inductor. However, the mechanical resonators have much higher Q's and better stability than the LC circuit. With the addition of a transducer to convert electrical energy to acoustic, the *LC* circuits can be replaced with mechanical resonators. *See* RESONANCE (ACOUSTICS AND MECHANICS); TRANSDUCER.

SAW filters can be built in both transversal structures and resonators structures. Crystal, ceramic, and mechanical filters are realized only as resonator filters. Acoustic resonator filters use mechanical resonators coupled together either electrically or acoustically. Each resonator provides either a pole or a zero in the transfer function. In addition, the piezoelectric coupling of mechanical to electrical energy provides a second resonance that may be used to create additional poles and zeros. *See* HELMHOLTZ RESONATOR; PIEZOELECTRICITY.

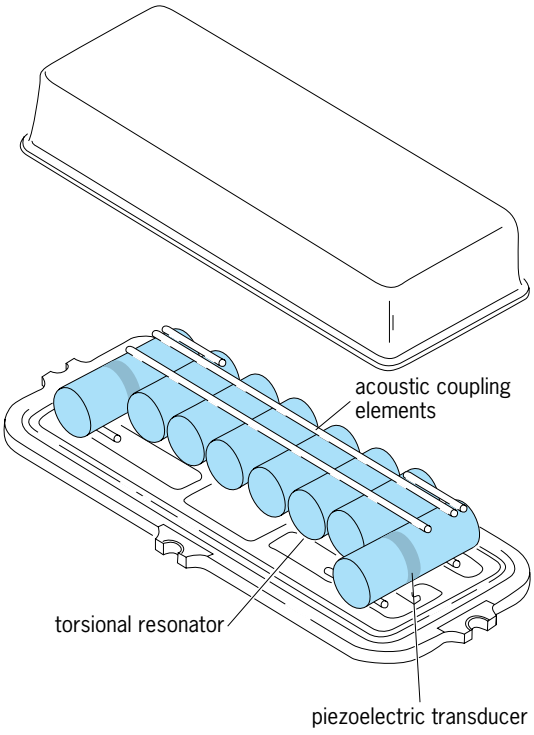


Fig. 4. Mechanical filter.

The difference between crystal and ceramic filters is mostly in the material that they use for their resonators. In crystal filters the resonators are made from quartz. It is a piezoelectric material with excellent temperature stability and high Q. Ceramic filters use other piezoelectric materials, mainly a lead-zirconate-titanate (PZT) ceramic, for their resonators. These materials have much higher coupling coefficients than quartz and therefore can achieve much wider bandwidths, but have lower Q's and poorer temperature stability.

Mechanical filters (**Fig. 4**) use metal for their resonant elements. Coupling between elements is done mechanically. Piezoelectric or other transducer types are used at the input and output to convert electrical energy to acoustic energy. Mechanical filters have temperature stabilities and Q's between those of quartz and ceramic filters.

SAW resonators trap surface acoustic waves between two reflecting structures. Interdigital fingers are used to couple electrical energy piezoelectrically into and out of the acoustic wave. Transversal SAW filters use a transmitting and receiving set of piezoelectric interdigital fingers. One or both of the finger sets are weighted to create a desired impulse response. The spacings between interdigital fingers act as time delays for the acoustic wave. These time delays and weighting of the fingers are analogous to processes used in digital filters. Because of this analogy, many of the algorithms that are used for transversal digital filters have been applied to SAW filters. *See* SURFACE-ACOUSTIC-WAVE DEVICES.

Other filter realizations. Many techniques are used to create filters. Inductors are replaced with transistor networks in active filters, discussed below, to reduce size and cost. By using an analog-to-digital converter the transfer function can be created mathematically by a digital processor. At high frequencies, transmission lines and waveguide structures replace lumped elements. *See* DIGITAL FILTER; INTEGRATED-CIRCUIT FILTER; MICROWAVE FILTER; SWITCHED CAPACITOR CIRCUIT.

Donald P. Havens

Active filters. An active filter comprises resistors, capacitors, and active elements such as operational amplifiers. It is also referred to as an active-*RC* filter.

Advantages and disadvantages. Active filters can realize the same filter characteristics as passive ones comprising resistor, capacitor, and inductor elements.

They have, however, several advantages over their passive counterparts:

1. Active filters can provide gain, and are frequently used to simultaneously match filtering (frequency-determining) and gain specifications.

2. They are readily implemented in integrated-circuit technology, whereas the inductor element of passive filters is not readily realized. As a result, the active filter is inexpensive, and is attractive for its small size and weight. In addition, it is readily included with other signal-processing functions on a single integrated circuit. *See* INTEGRATED CIRCUITS.

3. The design of active filters is considerably simpler than that of passive ones. In addition, it is easy to provide for variability, which can be used to change filter characteristics by electrical input signals.

The active filter also has some disadvantages:

1. Since the active filter contains electronic components, it requires a power supply, which adds to the complexity of the realization. The electronic components also place restrictions on the level of the signals that can be applied to the filter and on the noise component that the filter may add to the filtered signal. *See* ELECTRONIC POWER SUPPLY.

2. The mathematical process by which the active filter produces filtering characteristics in general requires the use of internal feedback. When this feedback is positive, the resulting filter may be very sensitive to lack of precision in component values, and the effects of aging and environmental conditions. *See* FEEDBACK CIRCUIT.

Single-amplifier filters. One important classification of active filters is according to the number of gain elements (amplifiers) that are used in their design. Historically, the first practical filter used a single amplifier of positive low gain, in the range of 1–5. The amplifier was realized by an operational amplifier and two feedback resistors, R_A and R_B. The resulting configuration could be used for low-pass, high-pass, and band-pass applications. Other filtering functions could be obtained by additional circuit complexity. The circuit is usually referred to as a Sallen-Key filter (**Fig. 5***a*). Since the Sallen-Key filter uses positive feedback, the sensitivity problem mentioned above usually restricts the use of this filter to low-Q applications.

A more frequently utilized single-amplifier filter uses an inverting operational amplifier directly as the active element. The filter is called an infinite-gain, single-amplifier filter (Fig. 5*b*). Similar filter configurations may be used to realize high-pass and band-pass filter characteristics. Although the design equations for the infinite-gain, single-amplifier filter are more complex than those for the Sallen-Key one, the use of negative feedback which is implicit in this realization provides for a far more practical realization with considerably lower sensitivities.

The single-amplifier, second-order filter realizations described above can also be used to realize higher-order filter functions. [The order of a filter is the number n in Eq. (1), giving its transfer function.] To do this, separate filter sections are used to realize second-order factors, $T_i(s)$ of the overall transfer function, and the resulting sections are connected in a cascade configuration in which the output of one filter section is used as the input of the next one (Fig. 5*c*). There are $n/2$ sections labeled T_i, and since each is of second order, an nth-degree function results.

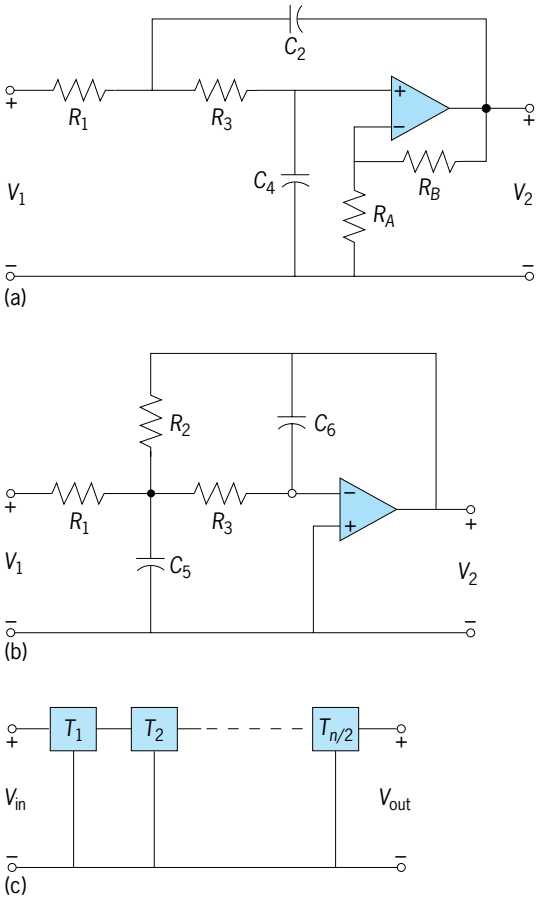


Fig. 5. Single-amplifier filters. (*a*) Low-pass Sallen-Key filter. (*b*) Low-pass, infinite-gain, single-amplifier filter. (*c*) Cascade of second-order filter realizations.

Multiple-amplifier filters. Another type of second-order filter, which can be used either by itself or as a building block in a cascade, is the multiple-amplifier filter. Such a filter uses more than one active element (operational amplifier), typically three. Although these filters have a more complex structure than the single-amplifier ones, and usually have a greater number of passive RC elements as well, they have excellent sensitivity properties and are easy to design. They are the most important general-purpose filter structures.

One of the best-known multiple-amplifier active filters is the state-variable filter, also known as the KHN (from the names of its inventors, W. J. Kerwin, L. P. Huelsman, and R. W. Newcomb) [**Fig. 6***a*]. Three outputs, V_{HP}, V_{BP}, and V_{LP}, for high-pass, band-pass, and low-pass transfer functions respectively, are available. The versatility of the state-variable filter is improved if a fourth operational amplifier is provided as a summing amplifier. By summing the high-pass, low-pass, and band-pass outputs, a biquadratic filter function can be realized. This includes notch filters,

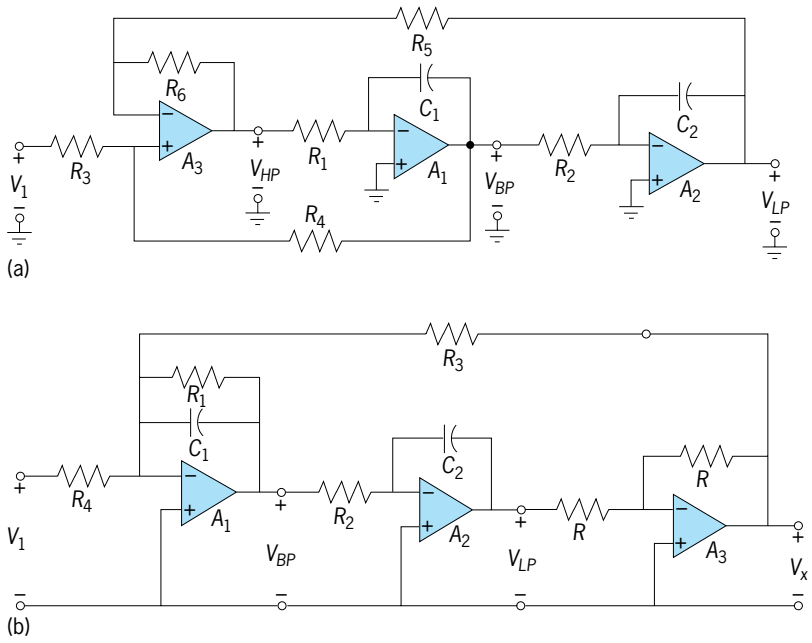


Fig. 6. Multiple-amplifier filters. (*a*) State-variable (KHN) filter. (*b*) Tow-Thomas filter.

elliptic (Cauer) filters, and all-pass filters. Realizations of higher than second order can be produced by cascading appropriately designed second-order realizations.

Another versatile multiple-amplifier filter, the Tow-Thomas filter (Fig. 6*b*), has the advantage of being even easier to design than the state-variable filter, but has the disadvantage of not providing a high-pass output in its basic configuration.

Direct realizations. A completely different class of active filters than the single- and multiple-amplifier ones described above is referred to as direct realizations. This class uses a passive *RLC* filter realization as a prototype, then directly derives an active-*RC* configuration from the prototype. Since passive filters can be of any order, the resulting class of active filters can also have any desired order. These high-order active filters are said to be directly realized in contrast to the cascade higher-order realizations described above. There are two main classes of direct realization filters.

The first type of direct realization is one in which the inductor element is replaced by a simulated inductor. The other elements in the circuit remain the same. This technique is especially applicable to the realization of high-pass filters in which the inductor elements all appear with one terminal connected to ground. This arrangement simplifies the connection of the power supplies for the active elements. An example of a simulated inductor realization is a fifth-order high-pass filter (**Fig. 7***a*). In a direct active-*RC* realization of the filter (Fig. 7*b*), each of the inductors is simulated by an active-*RC* circuit consisting of four resistors, a capacitor, and two operational amplifiers.

Another example of an element-replacement direct realization is one in which the elements are first subjected to an impedance transformation of $1/s$, in which $R \rightarrow C$, $L \rightarrow R$, and C becomes an element with an impedance $1/Cs^2$. This element is called a frequency-dependent negative resistor (FDNR), and is indicated in circuit diagrams by a four-bar symbol. Such a transformation leaves the voltage transfer function invariant but permits the realization of a low-pass filter with grounded active-*RC* elements. An example is a fifth-order *RLC* low-pass filter (**Fig. 8**), in which two active-*RC* circuits, each consisting of two capacitors, three resistors, and two operational amplifiers, realize the FDNRs.

The second type of direct realization is one in which the voltages and currents of a passive *RLC* filter are simulated by the voltages in a circuit consisting of damped and undamped inverting integrators. The resulting configuration has a series of feedback paths and is called a leapfrog filter.

Computer-aided design. The design of various active filters (as well as their passive counterparts) can be simplified by the use of various computer-aided design programs which operate on a personal computer. *See* COMPUTER-AIDED DESIGN AND MANUFACTURING.

Parasitics. In general, active filters are designed with the assumption that the active elements are ideal (no parasitics). For example, the operational amplifier is assumed to have infinite gain, infinite input

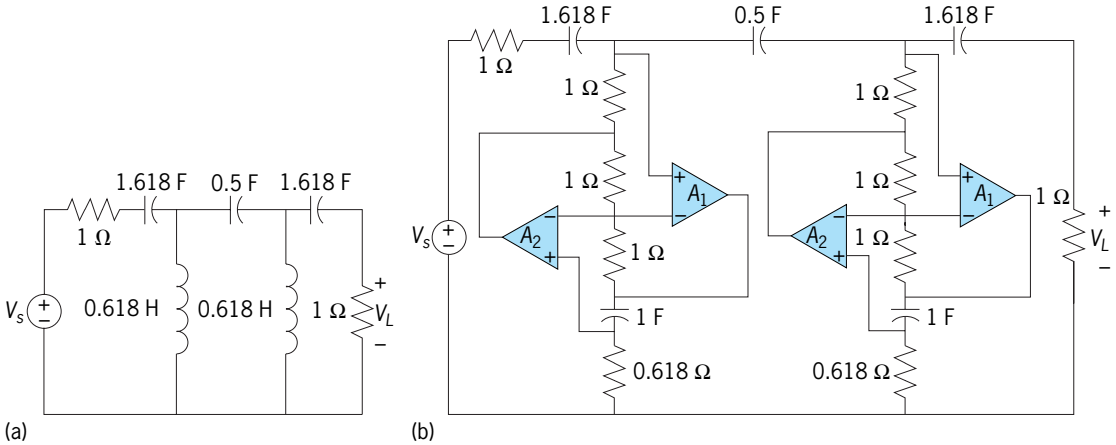


Fig. 7. Simulated-inductor realization of fifth-order high-pass filter. (*a*) *RLC* prototype, normalized for cutoff frequency of 1 rad/s (0.159 Hz). (*b*) Direct active-*RC* realization.

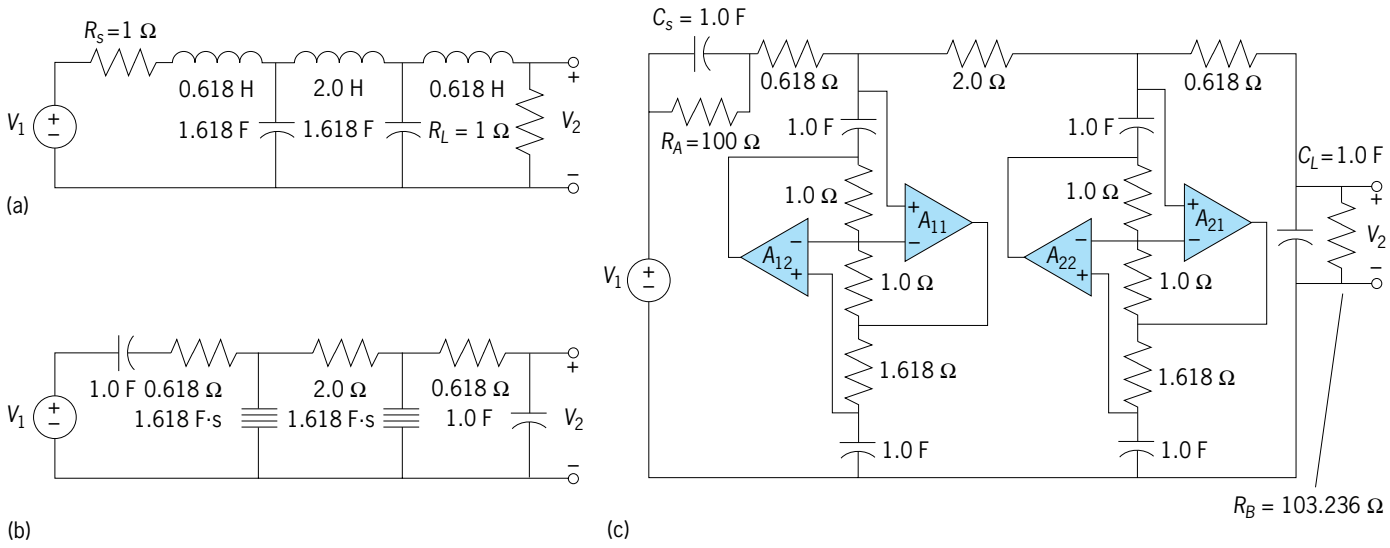


Fig. 8. FDNR realization of fifth-order low-pass filter. (*a*) *RLC* prototype, with normalized low-pass function with cutoff frequency of 1 rad/s (0.159 Hz). (*b*) Filter after impedance transformation. (*c*) Final realization, in which active-*RC* circuits realize FDNRs.

impedance, zero output impedance, and an infinite frequency range (gain-bandwidth). All practical filter realizations must be evaluated for the effect that the nonideal parasitics have on actual filter performance. One of the most troublesome of the parasitics is the operational-amplifier gain bandwidth. Typically, for a given application, tuning charts may be developed which can be used to provide compensation for such operational-amplifier limitations. *See* AMPLIFIER; CIRCUIT (ELECTRONICS); OPERATIONAL AMPLIFIER.

Lawrence P. Huelsman

Bibliography. H. Blinchikoff and A. Zverev, *Filtering in the Time and Frequency Domains*, 1976, reprint 1986; M. Hasler and J. Neirynck, *Electric Filters*, 1986; L. P. Huelsman, *Active and Passive Analog Filter Design*, 1993; S. Mitra and C. Kurth, *Miniature and Integrated Filters*, 1989; R. Schaumann and M. Van Valkenburg, *Design of Analog Filters: Passive, Active-RC, and Switched Capacitor*, 2d ed., 2000.

Electric furnace

A chamber heated to high temperature by electricity. The furnace consists of a refractory shell to resist the high temperatures attained within the chamber and minimize the heat losses to the surrounding area, plus a power source and electrical circuit to provide the heat. The operating temperatures range from a few hundred degrees Celsius to the melting points of refractory metals. The heating mechanism may be resistance, induction, microwave, or an electric arc in the largest installations. Electric furnaces are used primarily in the metallurgical industries for heat-treating, melting, and smelting operations. Resistance furnaces are used for heat treating. Induction furnaces are used to melt relatively low-melting alloys for casting, while arc furnaces are used for melting and smelting applications. The range of furnace powers, sizes, and geometries varies immensely. Powers range from a few kilowatts to 100 megavolt-amperes in a single unit. The furnace volume may be less than a cubic meter to more than 300 m^3. Both cylindrical and rectangular geometries are common. *See* ELECTRIC HEATING; ELECTROMETALLURGY; FURNACE CONSTRUCTION; HEAT TREATMENT (METALLURGY); INDUCTION HEATING; METALLURGY; REFRACTORY; RESISTANCE HEATING.

Electric-arc furnaces, in which the prime mechanism of heat transfer is from an electric arc, may be of the direct-arc or indirect-arc type. In the direct-arc furnace, an arc is struck between the charge being melted (usually scrap steel) and an electrode. In the indirect-arc type, an arc is struck between adjacent electrodes, and this arc then heats the charge. The arc may also be submerged below the surface of the charge. In this case, some of the current may heat the charge by resistance heating through the charge, while the rest of the current passes through the arc. Submerged-arc furnaces are among the largest units in operation and are used primarily for the smelting of ferroalloys and silicon metal, as well as phosphorus, calcium carbide, calcium silicon, copper, and copper-nickel mattes. *See* ARC HEATING.

Both alternating current and direct current are used to operate electric-arc furnaces. Historically, alternating-current devices with three or six carbon electrodes were used. A cylindrical geometry with the electrodes arranged in a triangle is common for three electrode systems, while the largest smelting furnaces are usually rectangular with the six electrodes installed in a line. Direct-current furnaces are cylindrical with a single axial electrode acting as cathode, while the furnace charge and bottom becomes the anode. Submerged-arc furnaces are continuously charged with feed, and there is a continuous discharge of gas from the furnace roof. Molten metal and/or slag are removed by tapping at discrete time intervals. Melting furnaces have a removable

roof for charging and often a means of tilting the furnace to remove the molten product. Although most arc furnaces use graphite electrodes, some smaller-capacity plasma furnaces use nonconsumable water-cooled metal electrodes, which allow for better control of the furnace atmosphere.

Presently one-third of the world's steel production and 50% of the United States' steel production comes from electric-arc furnaces. The largest units have graphite electrodes 60–80 cm (24–31 in.) in diameter and a tap weight of 80–100 metric tons (88–110 tons) of liquid steel. The active power may be up to 100 megawatts. A typical FeNi submerged-arc smelting furnace would have six electrodes and an operating load of 75 MVA. *See* STEEL MANUFACTURE.

Richard J. Munz

Bibliography. N. Barcza, The application of dc plasma-arc technology to extractive metallurgical processing, *High Temp. Mater. Proc.*, 1:1–15, 1997; F. Fereday, *Commentary: Submerged Arc Furnaces*, TechCommentary published by CMP, EPRI Center for Materials Production, Carnegie Mellon Research Institute, 15230-2950, 1996; J. Kunze and R. Degel, History, current status and new trends of submerged arc furnace technology for ferroalloy metals, *Steel GRIPS*, 1(3):164–169, 2003; D. Neuschütz, Metallurgical applications of high power are heating systems, *High Temp. Mater. Proc.*, 4(3):309–321, 2000.

Electric heating

Methods of converting electric energy to heat energy by resisting the free flow of electric current. Electric heating has several advantages: it can be precisely controlled to allow a uniformity of temperature within very narrow limits; it is cleaner than other methods of heating because it does not involve any combustion; it is considered safe because it is protected from overloading by automatic breakers; it is quick to use and to adjust; and it is relatively quiet. For these reasons, electric heat is widely chosen for industrial, commercial, and residential use.

Types of electric heaters. There are four major types of electric heaters: resistance, dielectric, induction, and electric-arc.

Resistance heaters. Resistance heaters produce heat by passing an electric current through a resistance—a coil, wire, or other obstacle which impedes current and causes it to give off heat. Heaters of this kind have an inherent efficiency of 100% in converting electric energy into heat. Devices such as electric ranges, ovens, hot-water heaters, sterilizers, stills, baths, furnaces, and space heaters are part of the long list of resistance heating equipment. *See* RESISTANCE HEATING.

Dielectric heaters. Dielectric heaters use currents of high frequency which generate heat by dielectric hysteresis (loss) within the body of a nominally nonconducting material. These heaters are used to warm to a moderate temperature certain materials that have low thermal conducting properties; for example, to soften plastics, to dry textiles, and to work with other materials like rubber and wood. *See* DIELECTRIC HEATING.

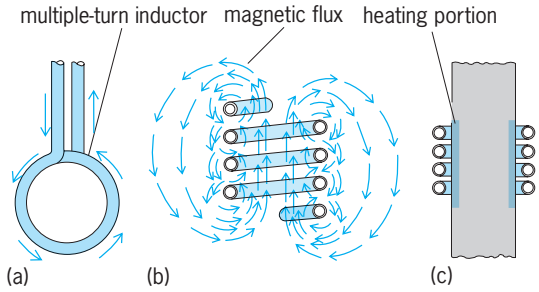


Fig. 1. Induction heater. Arrows indicate current flow. (*a*) End-on view. (*b*) Magnetic fields. (*c*) Heating pattern.

Induction heaters. Induction heaters produce heat by means of a periodically varying electromagnetic field within the body of a nominally conducting material. This method of heating is sometimes called eddy-current heating and is used to achieve temperatures below the melting point of metal. For instance, induction heating is used to temper steel, to heat metals for forging, to heat the metal elements inside glass bulbs, and to make glass-to-metal joints. **Figure 1** shows the magnetic fields and heating pattern produced by an induction heater. *See* INDUCTION HEATING.

Electric-arc heaters. Electric-arc heating is really a form of resistance heating in which a bridge of vapor and gas carries an electric current between electrodes. The arc has the property of resistance. Electric-arc heating is used mainly to melt hard metals, alloys, and some ceramic metals (**Fig. 2**). *See* ARC HEATING.

General design features. All electrical parts must be well protected from contact by operators, work materials, and moisture. Terminals must be enclosed within suitable boxes, away from the high heat zone, to protect the power supply cables. Repairs and replacements should be possible without tearing off heat insulations.

Resistance heaters are often enclosed in pipes or tubes suitable for immersion or for exposure to difficult external conditions. Indirect heating is done by circulating a heat transfer medium, such as special oil or Dowtherm (liquid or vapor), through jacketed vessels. This permits closer control of heating-surface temperature than is possible with direct heating.

Some conducting materials can be heated by

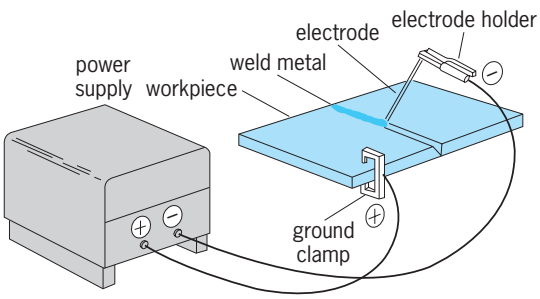


Fig. 2. Typical electric-arc heater.

passing electric current through them, as is done in the reduction of aluminum. Some conducting liquids can be heated by passing an electric current between immersed electrodes. Heat is produced by the electrical resistance of the liquid.

The supply of necessary electric power for large heating installations necessitates consultation with the utility company. The demand, the power factor of the load, and the load factor all affect the power rates. Large direct-current or single-phase alternating-current loads should be avoided. Polyphase power at 440–550 V permits lower current and reduced costs. *See* FURNACE CONSTRUCTION.

Electric heating for houses. Electricity is one choice for heating houses, but with only a 35% efficiency rate, electricity has been a less attractive option than the direct use of gas and oil for heating homes.

One type of electric heater for houses is the heat pump. The heat pump can be used alone or to supplement the output of direct resistance heating (heat strips). The heat pump works on a reversed refrigeration cycle; the pump pulls the heat from the cold outside air and forces it into the house. The efficiency of this process is directly affected by the ambient temperature and the desired interior temperature; the larger the difference in these temperatures, the lower the efficiency of the pump. Generally, however, the efficiency of the heat pump is higher than the efficiency of resistance heating, so that the efficiency of the heat pump is actually higher than 100%. *See* HEAT PUMP.

Since the heat pump performs both heating and cooling and is more efficient than other forms of electric heat, it can be an economical way to heat houses.

Common electric heating systems in houses are central heating employing an electric furnace with forced air circulation; central heating employing an electric furnace with forced water circulation; central heating using radiant cables; electrical duct heaters; space (strip) heaters which use radiation and natural convection for heat transfer; and portable space heaters.

Mo-Shing Chen

Bibliography. C. J. Erikson, *Handbook of Electric Heating for Industry*, 1994; D. G. Fink and H. W. Beaty (eds.), *Standard Handbook for Electrical Engineers*, 14th ed., 1999; L. Miles, *Heat Pumps*: *Theory and Service*, 1993; M. Orfeuil, *Electric Process Heating*, 1987; H. J. Sauer, *Heat Pump Systems*, 1990.

Electric insulator

A material that blocks the flow of electric current across it. Insulators are to be distinguished from electrolytes, which are electronic insulators but ionic conductors. *See* ELECTROLYTE; IONIC CRYSTALS.

Electric insulators are used to insulate electric conductors from one another and to confine electric currents to specified pathways, as in the insulation of wires, electric switchgear, and electronic components. They provide an electrical, mechanical, and heat-dissipation function. The electrical function of an insulator is characterized by its resistivity, its dielectric strength (breakdown voltage), and its dielectric constant (relative permittivity). Insulators can be solid, liquid, or gaseous (see **table**).

Resistivity. The resistivity of a material is a measure of the electric current density that flows across it in response to an applied electric field. The electric current density is the sum of contributions from all species of mobile charge carriers, including electrons and ions. Each contribution is proportional to the concentration of the mobile species, its charge, and its mean (drift) velocity in response to the applied field. Solid and liquid insulators have direct-current resistivities of 10^{10} ohm-meters at room temperature as compared to 10^{-8} Ω-m for a good metal or 10^{-3} Ω-m for a fast ionic conductor. Insulators minimize the concentrations of mobile charge carriers by having a filled valence band separated from an empty conduction band by a large energy gap and by having all equivalent lattice sites either occupied or vacant of ionic species. An energy gap greater than 5 eV can be obtained in either ionically or covalently bonded crystals and molecules. Oxides may have the advantage of being inert to further oxidation, but they are more subject to short-range, if not long-range, ionic motion. Protons introduced by the absorption or adsorption of water are the most prevalent source of ionic conduction. Impurities and crystalline defects are sources of mobile electrons accessible at lower temperatures and electric fields. *See* BAND THEORY OF SOLIDS; ELECTRICAL RESISTANCE; ELECTRICAL RESISTIVITY; SEMICONDUCTOR.

Dielectric strength. A material used in the electrical industry for insulation, capacitors, or encapsulation may be exposed to large voltage gradients that it must withstand for the operating life of the system. Power lines operate at transmission voltages in excess of 1 MV, for example. Failure occurs if an electric short-circuit develops across the material. Such a failure is called dielectric breakdown. The breakdown voltage gradient, expressed in kilovolts per millimeter, is a measure of the dielectric strength. Dielectric breakdown of a solid is destructive; liquids and gases are self-healing.

Breakdown is initiated by the injection of mobile charge carriers into the conduction or valence band of the insulator. There are three basic types: intrinsic, thermal, and avalanche. Intrinsic breakdown occurs where electrons are injected across a metal-insulator interface. It depends primarily on the applied field gradient, it occurs in less than 1 microsecond, and it is independent of sample shape. Thermal breakdown is associated with thermal excitation of electrons into a conduction band; resistive losses of the excited electrons lead to thermal runaway. This mechanism depends primarily on the temperature rather than the applied field gradient, and is dependent on both sample shape and the rate of application of the field. Avalanche breakdown occurs where collisions between mobile and stationary electrons multiply

Characteristics of electric insulators*

Insulator	Breakdown voltage, kV/mm	Resistivity, Ω-m,	Relative permittivity (*k*)
Solid			
Paper	20	10^{10}	3–4
Nylon	16	10^{13}	3–4
Glass	20–40	10^{10}–10^{14}	5
Rubber	16–50	10^{13}–10^{14}	3
Mica	25–200	10^{11}–10^{15}	5
Poly(tetrafluoroethylene)	60	10^{14}–10^{15}	2.1
Porcelain	10–20	10^{16}	5–7
Polystyrene	25	10^{16}	2.5
Liquid			
Saturated hydrocarbons†	140–200	10^{17}–10^{18}	
Aromatic hydrocarbons†	160–280	10^{17}–10^{18}	
Transformer and silicon oils‡	15		
Gaseous§			
Air	3		
Carbon dioxide	2.7		
Sulfur hexafluoride	7.5		

*At 20°C (68°F) and atmospheric pressure (101.3 kPa)
†Physically and chemically purified.
‡Commercial high purity.
§Between two electrodes 25 cm (10 in.) in diameter.

the concentration of mobile electrons. It generally occurs in conjunction with intrinsic or thermal breakdown.

In addition, there are three pseudo-types of breakdown of great practical importance: discharge, electrochemical, and mechanical. Dielectric discharge is associated with a gaseous breakdown at the surface or within included pores of the material. Electrochemical breakdown is a result of chemical reactions that, over time, reduces the dielectric strength. Mechanical breakdown is precipitated by cracks or other defects that distort the applied field. *See* BREAKDOWN POTENTIAL; ELECTRICAL BREAKDOWN; ELECTRICAL CONDUCTION IN GASES.

Dielectric constant. An insulator is also known as a dielectric. The dielectric constant k is defined as the ratio of the capacitance of a flat-plate condenser, or capacitor, with a dielectric between the plates to that with a vacuum between the plates; this ratio is also the relative permittivity of the dielectric. The dielectric constant is a measure of the ability of the insulator to increase the stored charge on the plates of the condenser as a result of the displacement of charged species within the insulator. Displacement of a charge q through a distance δ creates an electric dipole of moment $p = q\delta$. The displacement of charge takes time, so in an alternating-current circuit $k = k'\text{-}ik''$ is a complex number; a measure of the losses introduced by the charge displacements is the loss tangent, $\tan \delta \equiv k''/k'$. Low dielectric constants are desirable for the insulation of alternating-current circuits.

In ceramics and glasses, four primary short-range motions contribute to the total polarization of the dielectric: electronic, atomic, dipolar, and interfacial. Electronic polarization is due to a shift of the valence-electron cloud relative to the atomic nucleus. It gives a resonance absorption peak at optical frequencies. Atomic polarization is due to a displacement of positive and negative ions with respect to each other. It gives a resonance in the infrared range 10^{12}–10^{13} Hz that may be quite broad where several types of ions are present. Dipolar, or orientational, polarization is a perturbation of the thermal motion of ionic or molecular dipoles to give a net dipolar orientation in the direction of the applied field. The preferred orientation of permanent dipoles (Stevels deformation polarization), which may be particularly important in liquid and gaseous insulators, occurs in the 10^{11}–10^{12}-Hz frequency range and the switching of dipoles between two equivalent equilibrium orientations in a solid occurs at lower frequencies, 10^3–10^6 Hz. Interfacial, or space-charge, polarization occurs where the motions of mobile charges are blocked by internal barriers such as two-phase interfaces. This mechanism produces a large capacitance and dielectric constant at low frequencies. Orientational and interfacial polarizations introduce losses in insulators from 10^{-3} to 10^9 Hz; atomic and electronic polarizations in glasses and ceramics are important for optical communication. *See* CAPACITANCE; DIELECTRIC MATERIALS; FERROELECTRICS; OPTICAL COMMUNICATIONS; PERMITTIVITY; POLARIZATION OF DIELECTRICS.

Mechanical properties. Successful application of solid insulating materials also depends on their mechanical properties. Insulation assemblies commonly must withstand thermal-expansion mismatch, tension, compression, flexing, or abrasion as well as a hostile chemical-thermal environment. The introduction of cracks promotes the penetration of moisture and other contaminants that promote failure, and the presence of pores may cause damaging corona discharge on the surface of a high-voltage conductor. As a result, composite materials and engineered shapes must be tailored to meet the challenges of a particular operational environment. *See* CORONA DISCHARGE.

For example, glasses and varnishes are used as sealants, and oil is used to impregnate high-voltage, paper-insulated cables to eliminate air pockets. Porcelain is a commonly used material for the suspension of high-voltage overhead lines, but it is brittle. Therefore, a hybrid insulator was developed that consists of a cylindrical porcelain interior covered by a mastic sealant and a silicone elastomer sheath heat-shrunk onto the porcelain core. The circular fins of the outer sheath serve to shed water. However, twisted-pair cables insulated with poly(tetrafluoroethylene) are used for high-speed data transmission where a small dielectric constant of the insulator material is needed to reduce signal attenuation. *See* CONDUCTOR (ELECTRICITY); ELECTRIC POWER TRANSMISSION; FLUORINE; PORCELAIN; SILICONE RESINS.

Liquid and gaseous insulators. Liquid or gas insulation provides no mechanical strength, but it may provide a cheap, flexible insulation not subject to mechanical failure. Biphenyls are used as insulating liquids in capacitors; alkyl benzenes in oil-filled cables; and polybutenes for high-pressure cables operating at alternating-current voltages as high as 525 kV. Sulfur hexafluoride (SF_6) is a nonflammable, nontoxic electron-attracting gas with a breakdown voltage at atmospheric pressure more than twice that of air. Fluorocarbons such as C_2F_6 and C_4F_8 as well as the Freons are also used, and breakdown voltages have been increased significantly in gas mixtures through a synergistic effect. Used as an insulating medium in high-voltage equipment at pressures up to 10 atm (1 megapascal), sulfur hexafluoride can reduce the size of electrical substations by a factor of 10 over that of air-insulated equipment. Enclosure of metal cable in a metallic conduit filled with sulfur hexafluoride gas has been used to shield nearby workers from exposure to high electric fields. *See* CIRCUIT BREAKER; FLUOROCARBON; POLYCHLORINATED BIPHENYLS.

Thermal conductivity. Finally, the ability to transfer heat may be an overriding consideration for the choice of an electric insulator. Electrical machines generate heat that must be dissipated. In electronic devices, the thermal conductivity of the solid substrate is a primary consideration. Where mechanical considerations permit, circulating liquid or gaseous insulation is commonly used to carry away heat. Liquids are particularly good transporters of heat, but they are subject to oxidation. In transformers, for example, the insulators are generally mineral or synthetic oils that are circulated, in some places with gaseous nitrogen to inhibit oxidation, to carry away the heat generated by the windings and magnetic core. Silicone fluids, while flammable, have a higher fire point than mineral oil and a lower dissipation factor. Circulating gaseous sulfur hexafluoride is also used for transformer insulation. Hydrogen is a gaseous insulator used to cool rotating electrical machines because it has a heat-conducting capacity over seven times that of air. *See* ELECTRIC ROTATING MACHINERY; ELECTRICAL INSULATION; HYDROGEN; TRANSFORMER.

John B. Goodenough

Bibliography. R. Bartnikas, *Electrical Insulating Liquids in Dielectrics*, vol. 3, 1994; R. Bartnikas et al. (eds.), *Engineering Dielectrics*, vol. 1, 1979, vol. 2A, 1983, vol. 3, 1987; C. Johnson, *Handbook of Electrical and Electronic Technology*, 1994; F. H. Kreuger, *Industrial High Voltage*, vol. 1: *Electric Fields, Dielectrics, Constructions*, 1991; W. T. Shugg, *Handbook of Electrical and Electronic Insulating Materials*, 2d ed., 1995.

Electric organ (biology)

An effector organ found in six different groups of fishes; output is an electric pulse (**Table 1**; **Fig. 1**). Voltages large enough to aid in prey capture or predator deterrence are produced by various strongly electric fishes. These include the electric eel (*Electrophorus electricus*) from South America; the electric catfish (*Malapterurus electricus*) from Africa; the family of electric rays, Torpedinidae, which are widely distributed in the world's oceans; and possibly the stargazer genus, *Astroscopus*, of the western Atlantic. Weakly electric fishes emit a lower voltage, the energy source for an active electrosensory system that monitors electrical impedance in the environment. These weak signals also serve in intra- and interspecific communication. There are three groups of weakly electric fishes. First, the South American knifefishes, the Gymnotiformes, are a large and diverse group of several families that also include *Electrophorus*. Second, the electrically active African Mormyriformes are composed of the numerous species of the family Mormyridae and the single species *Gymnarchus niloticus* in the family Gymnarchidae. Finally, many species of skates and rays of the family Rajidae occur in marine waters worldwide.

The strongly electric fishes are remarkable for the high voltage and power of their discharges (**Table 2**). For example the electric eel generates pulses in excess of 500 V. In contrast, a large torpedo generates a smaller voltage, about 50 V in air, but the maximum current is larger, and the peak pulse power can exceed 1 kW (or about 1 hp). The electric organs may constitute a substantial fraction of the body mass. The weakly electric fishes have much smaller organs emitting pulses of a few tenths of a volt to several volts; such amplitudes are still large compared to those recorded outside ordinary, nonelectric fishes.

Electric organ discharge is explicable in terms of the properties of ordinary excitable cells, that is, those of nerve and muscle. The single generating cells of electric organs are called electrocytes. They are modified striated muscle fibers that have lost the ability to contract, although they retain muscle filaments to varying degrees, and they produce electric signals as muscle fibers do.

Electrocyte operation. The basic operation of an electrocyte is as follows. Consider a flattened cell which contains a high concentration of potassium ions (K^+), a low concentration of sodium ions (Na^+), and an equivalent number of unspecified anions. Let

TABLE 1. Groups of electric fishes*

Common name	Family	Genera and species	Strength of organ discharge	Distribution
Skates, ordinary rays	Rajidae	*Raja*, many species, a number of other genera not known to be electric	Weak	Marine, cosmopolitan
Electric rays, torpedos	Torpedinidae	About 20 genera	Strong, up to 60 V or 1 kW, some perhaps weak	Marine, cosmopolitan
Mormyrids, elephant-nosed fish (many lack enlarged chin or snout)	Mormyridae	About 11 genera, several with many species	Weak	Fresh-water, Africa
Gymnarchus	Gymnarchidae	1 species, *Gymnarchus niloticus*	Weak	Fresh-water, Africa
Gymnotid eels, electric eel, knifefishes	Electrophoridae	1 species, *Electrophorus electricus*	Strong, more than 500 V	Fresh-water, South America
	Gymnotidae	1 species, *Gymnotus carapo*	Weak	Fresh-water, South America
	Sternopygidae	About 8 genera, a number of species	Weak	Fresh-water, South America
	Sternarchidae (Apteronotidae)	About 5 genera, a number of species	Weak	Fresh-water, South America
Electric catfish	Malapteruridae	1 species, *Malapterurus electricus*	Strong, more than 300 V	
Stargazers	Uranoscopidae	1 electric genus, *Astroscoupus,* several species	Strong (?), about 6 V in air from small animals	Marine, western Atlantic

*From M. V. L. Bennett, Electric organs, in W. S. Hoar and D. J. Randa ll (eds.), *Fish Physiology,* vol. 5, pp. 347–491, Academic Press, 1971.

it be bathed in a solution which is high in Na^+, and low in K^+. Initially its membrane is specifically permeable to potassium ions (K^+) over its entire surface, and in this condition the membrane potential is uniform and no currents flow (**Fig. 2***a*). The cell interior will be at some negative voltage, at which the tendency for K^+ to diffuse out and down its concentration gradient is just balanced by the negativity of the interior, tending to prevent the positively charged ions from diffusing out. This "equilibrium" potential is about 90 mV inside negative for most excitable cells. When the cell is activated, one of its membrane faces becomes much more permeable to Na^+. The potential at which this membrane would be at equilibrium in terms of Na^+ flow is about 50 mV inside positive, but the cell interior is negative. Thus, Na^+ flows inward through the activated face, making the interior more positive. The other membrane face that still is specifically permeable to K^+ is displaced from its equilibrium potential, and as a result K^+ flows outward through it. In effect, the two faces act as batteries oriented in series (Fig. 2*b*). All operate just this simply; apposed membranes have different permeabilities during activity, allowing ions to flow down their concentration gradients and thereby generating electric currents. The proximate sources of energy for organ discharge are the differences in ionic concentrations between interior and exterior of the electrocytes. During activity these concentrations become more alike; between periods of activity, metabolic energy is expended to restore the concentration differences. *See* BIOPOTENTIALS AND IONIC CURRENTS.

During activity, currents crossing all of the electrolyte membranes in series flow around the outside of the electric organ through the surrounding body

TABLE 2. Discharge characteristics of electric fishes

Species	Approximate discharge voltage, V	Duration, ms	Frequency	Number of electrocytes in series
Electrophorus electricus, electric eel	700	2	0–250	6000
Malapterurus electricus, electric catfish	300	1.5	0–200	?
Torpedo nobiliana, electric ray	60	5	0–150	1000
Astroscopus γ-graecum, stargazer	6	5	0–150	200
Gnathonemus petersii, mormyriform	5	0.3	2–100	100
Gymnarchus niloticus, mormyriform	1	2	250	140
Gymnotus carapo, gymnotiform knifefish	1	1.5	50–150	90
Sternarchus albifrons, sternarchid (apteronotid) knifefish	1	1	800	90
Sternopygus macrurus, gymnotiform knifefish	1	10	50 male, 150 female	90
Raja erinacea, rajid skate	2	100	0–5	200

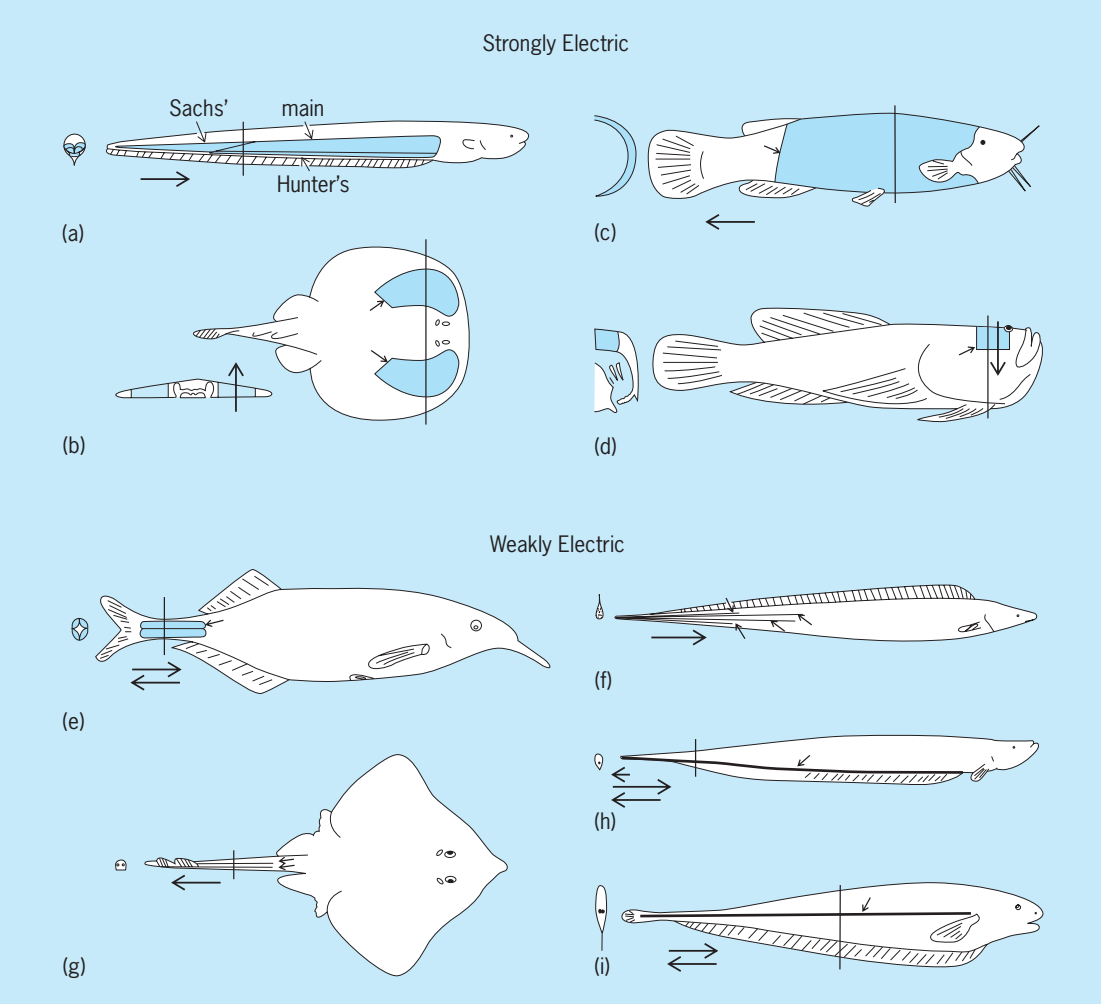


Fig. 1. Representative electric fishes: (*a*) *Electrophorus*; (*b*) *Torpedo*; (c) *Malapterurus*; (*d*) *Astroscopus*; (e) *Gnathonemus*; (*f*) *Gymnarchus*; (*g*) *Raja*; (*h*) *Gymnotus*; (*i*) *Sternarchus*. Electric organs are shaded or solid and indicated by small arrows. Cross-sectional views through the organs are also shown at the levels indicated by the solid lines. The large arrows indicate the direction of active current flow through the organs; more than one arrow indicates successive phases of activity, and the relative lengths are proportional to relative amplitudes of the phases. (*After W. S. Hoar and D. J. Randall, eds., Fish Physiology, vol. 5, Academic Press, 1971*)

and water, producing a field, similar in form to that around a bar magnet (Fig. 2*c*). A voltage drop occurs across any object in the field. If the objects differ in conductivity from water, they distort the pattern of current flow, and these distortions can be detected by the electrosensory system (see below).

Ions enter or leave a cell by passing through specific protein channels in the cell membrane. The extracellular fields generated by electrocytes are significantly larger than those generated by ordinary excitable cells, but the difference is not in the nature of the ionic permeabilities or the magnitude of the membrane voltages. Rather it is in the shape of the cells and the distribution and density of ionic channels. The maximum voltage generated by a single electrocyte is not much over 0.1 V. The extraordinarily large voltages that are generated by strongly electric organs result from the arrangement of many elements in series (Fig. 2*c*) and in parallel. For example, in the electric eel many thousands of cells occur in series (Table 2). In the torpedos there are hundreds of cells in series.

Electrocyte membranes. The arrangement of specific kinds of membrane differs in different electrocytes. In the strongly electric fishes, electrocytes are always flattened (hence the terms electroplaque and electroplax, **Fig. 3**). The simplest are those of the electric ray and stargazer. One face is densely innervated and generates only postsynaptic potentials (PSPs), as at the nerve-muscle junction. The PSPs result from a permeability increase to Na^+ and K^+, so that the membrane is at electrical equilibrium near 0 V instead of inside positive. The PSPs are produced by action of a neurotransmitter substance, acetylcholine. The transmitter is secreted by the innervating nerves. It combines briefly with receptor molecules in the electrocyte membrane, causing the ion channels to open; then it is inactivated by the enzyme acetylcholinesterase, which is found at high concentration at the synapses. The uninnervated face of these electrocytes is inexcitable; its permeability does not change when a potential is applied. At rest the uninnervated face has a much lower electric resistance than the innervated face. During

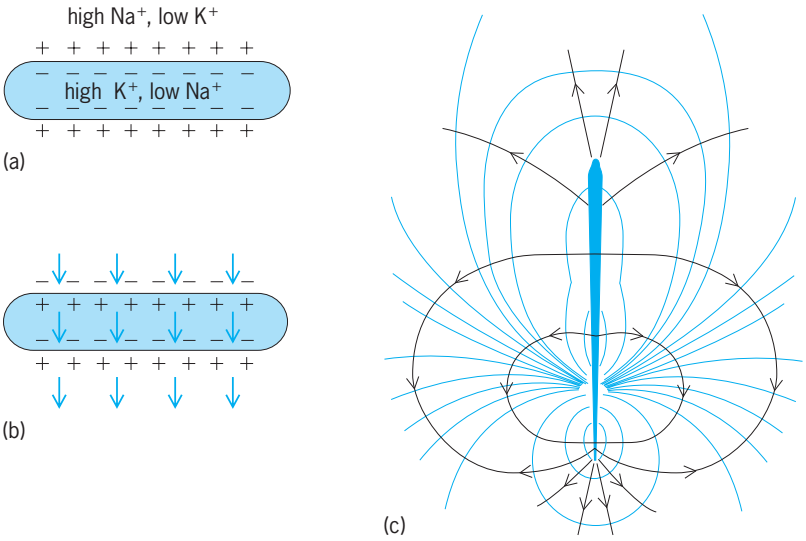


Fig. 2. Current flow through a single electrocyte and fields around an electric fish. (*a*) Cross section of a flattened cell at rest. Each face is specifically permeable to K^+, and the cell interior is at a negative potential. Because equal potentials are opposed, no current flows. (*b*) During activity that greatly increases the permeability to Na^+ in the upper face, this membrane moves toward inside positivity. As a result, two batteries act in series and current flows. (*c*) Flow of current (black lines) and isopotentials (colored lines) around a weakly electric gymnotid during a head positive phase of organ discharge. (*After W. Heiligenberg, Principles of Electrolocation and Jamming Avoidance in Electric Fish, Springer, 1977*)

PSPs, however, the resistance of the innervated face greatly decreases and the two faces have comparable resistances. These characteristics maximize the effectiveness of the cells as generating elements.

The electrocytes of the electric eel (and of several weakly electric relatives) are slightly more complex. In these cells, as in those of the electric ray and stargazer, one face is innervated and one is not, the uninnervated face being of low resistance and inexcitable. In addition to PSP channels, the innervated membrane has channels specifically permeable to Na^+; they open whenever the cell interior becomes about 10 mV or so more positive than its resting potential (depolarization). The PSPs generated by the innervated membrane cause some of these channels to open, greatly increasing the permeability to Na^+. The initial opening of Na^+ channels allows an influx of Na^+, which makes the interior yet more positive, causing a further opening of Na^+ channels, and so forth; the response is said to be regenerative. The positive feedback action continues until virtually all the Na^+ channels are open, and an impulse resembling the action potential of axons and muscle fibers is generated. In contrast to the electrically opened Na^+ channels, the PSP generation channels are little affected by membrane potential. In the electric eel the density of PSP channels is much lower than in the electric ray, and most of the change in membrane potential results from the electrically opened Na^+ channels. Two mechanisms can contribute to termination of the impulse. In the eel the Na^+ channels simply close after a brief period, and the membrane returns to the K^+ equilibrium potential. In others there are additional K^+ channels that are opened by depolarization somewhat more slowly than the Na^+ channels, and these speed the return of the membrane potential to its resting value.

The three classes of response, mediated by the PSP channels and by the Na^+ and K^+ channels opened by depolarization, can be mixed together in a number of different arrangements. In some weakly electric fishes, both innervated and uninnervated membranes generate impulses. First, current flows in one direction through the cell as the innervated membrane generates its impulse; this current then excites the uninnervated face, though with some delay, producing an impulse with current flow in the opposite direction. Thus, a biphasic external potential results. In the electric catfish, for example, one face generates a briefer impulse than the other. The resting resistances of the two faces are about equal, and the role of the briefer impulse appears to be to increase the K^+ permeability of that face, thus making the resistances of the two faces comparable during the later part of the longer lasting impulse, thereby increasing current flow. A similar phenomenon occurs in some skates, except that there are no electrically opened Na^+ channels. PSPs generated in the innervated face depolarize the uninnervated face, opening K^+ channels in it, and thus increasing current flow.

In several weakly electric fishes (for example, *Gymnarchus*) the uninnervated membrane has a very low permeability to K^+ or to any other ions. This membrane does have a large electrical capacity that permits significant current flow. Because the current flow is capacitative, the discharge has no direct current (dc) component. The absence of a dc component can be related to the functioning of the discharge in the electrosensory system.

Discharge. In all strongly electric fishes the activity of individual electrocytes is highly synchronized, thus maximizing the voltage and current produced. The discharges are monophasic pulses which are presumably more effective in stimulating prey or predator (A and A′ in **Fig. 4**). In weakly electric fishes the discharges are more diverse. There are a few with discharges like the strongly electric fishes, but others generate diphasic or triphasic pulses that need not be synchronous throughout the electric organ (B and B′ in Fig. 4). Triphasic discharges can result from out-of-phase addition of diphasic responses from two sets of electrocytes (*Gymnotus*). Alternatively a single electrocyte can generate a triphasic response (some mormyrids). The shape of electrocytes is also more variable in weakly electric fishes. The cells can be drum-shaped (*Gymnotus*, Fig. 3) or spindle-shaped (*Sternopygus*, Fig. 3). The innervation can occur on a thin process or stalk from the main body of the cell (*Malapterurus*, Fig. 3; some gymnotids). The stalks may also be multiple (mormyrids, Fig. 3; some gymnotids) and may even turn around and pass through holes in the flattened part to the opposite side before receiving their synapses (some mormyrids). Propagated activity in these stalks introduces a small component in overall organ discharge.

A number of gymnotids have a separate small "accessory" electric organ in the head region.

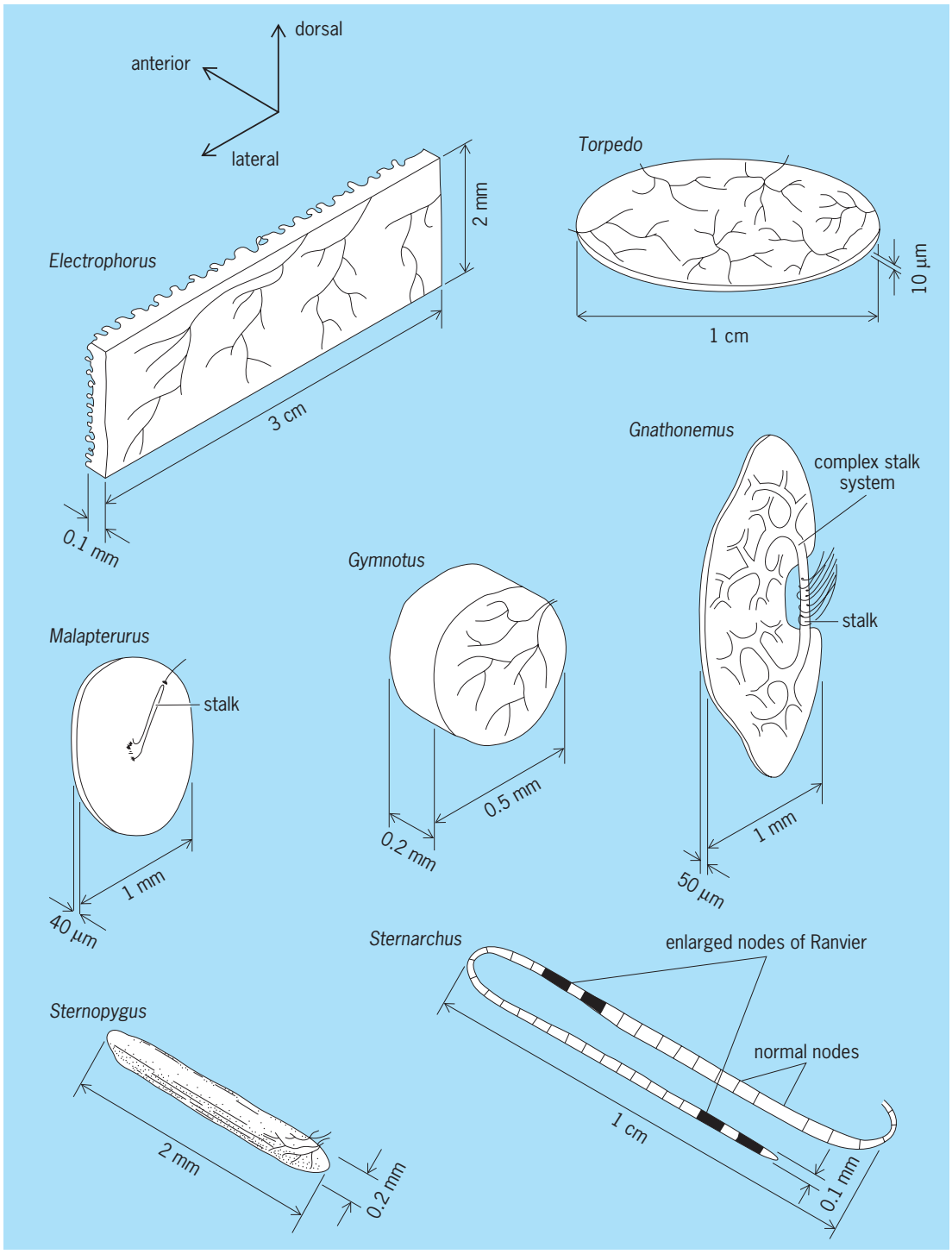


Fig. 3. Representative electrocytes. Body axes are indicated by the arrows except that the *Torpedo* electrocyte is shown ventral surface up. Innervation is indicated by branching, except for *Malapterurus* in which a single nerve fiber makes a single contact on the stalk tip. Also, the electrocyte of *Sternarchus* is a nerve fiber.

Electroreceptors are also concentrated on the head, and presumably the accessory organ functions in detecting localized objects in this region.

In strongly electric fishes the shape of the electric organ can be related to the conductivity of the environment (Fig. 1). In the marine forms, the torpedos and the stargazers, the organ is short and wide, particularly in the torpedos. In the fresh-water forms, the electric catfish and eel, the organs are long and relatively narrower. In the marine, as compared to fresh-water forms, fewer cells occur in series and more in parallel, and the voltages are lower and the currents higher in the higher-conductance medium.

In weakly electric fishes, including the marine rajids, the organs are generally long and slender. They have fewer than a dozen cells in parallel and of the order of 100 in series (Table 2).

In view of the high power that can be generated

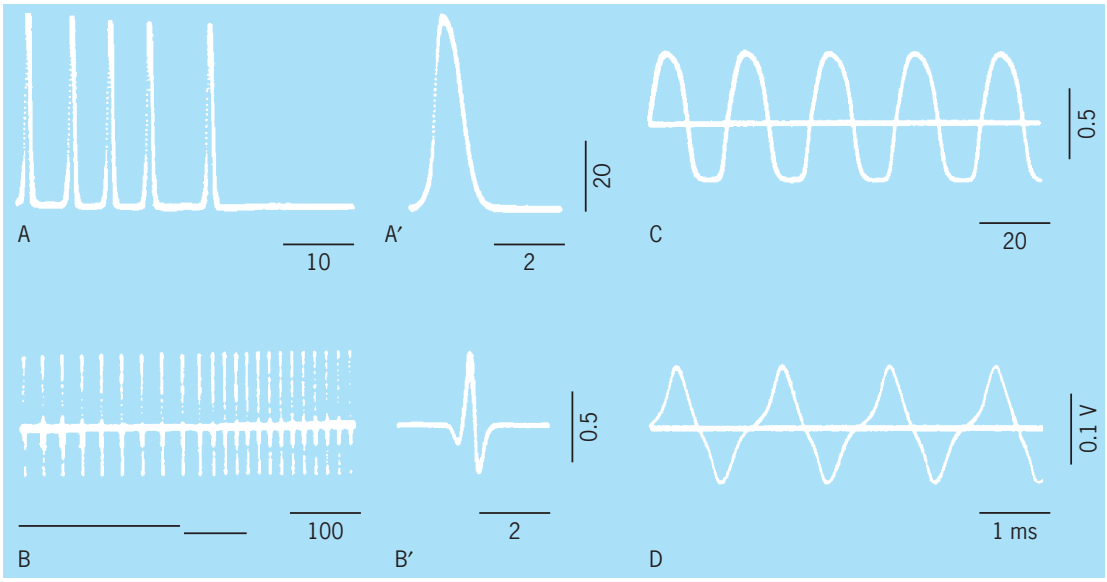


Fig. 4. Patterns of electric organ discharge. (A) An electric catfish; potentials recorded between head and tail in a small volume of water with head negativity upward. Pulses attain a maximum frequency of 190/s. (A′) Single pulse. (B–D) Weakly electric gymnotids immersed in water; discharges recorded between head and tail, head positivity upward. (B) *Gymnotus*; pulses are emitted at a basal frequency of approximately 35/s. Tapping the side of the fish at the time indicated by the downward step in the lower trace causes an acceleration up to about 65/s. (B′) Faster sweep showing the triphasic pulse shape. (C) *Sternopygus*; pulse frequency is about 55/s. The horizontal line indicates the zero potential level. (D) *Sternarchus*; pulse frequency is about 800/s. The horizontal line indicates the zero potential level. (*After W. S. Hoar and D. J. Randall, eds., *Fish Physiology, vol. 5, Academic Press, 1971*)

by strongly electric fishes, it is not surprising that they discharge the organs only infrequently, and generally in response to an external stimulus (A in Fig. 4). The fresh-water weakly electric fishes, however, emit pulses continually (B–D in Fig. 4), and the range of frequency is wide, from a few pulses per second to more than a thousand. The frequencies and discharge patterns are species-specific and differ considerably. Two broad types can be defined, pulse species and continuous-wave species. The pulse species emit brief pulses separated by a much longer interval; nonspecific sensory stimulation usually causes an increase in the frequency of discharge (B and B′ and Fig. 4). Continuous-wave species emit monophasic or biphasic pulses separated by intervals of about the same duration as the pulses (C and D in Fig. 4). These species generally modify their discharge frequency very little, although all show small changes when presented with a stimulus at a frequency close to their own. These changes, termed jamming avoidance responses, are presumed to reduce interference in electrolocation.

The discharges of rajids under normal conditions have not been described. They are evoked with difficulty by handling and consist of long-lasting, repetitive, asynchronous discharges of the generating cells.

Neural control. The neural control of electric organ discharge has been studied in representative species of most groups. Generally the frequency is set in a command nucleus, a compact group of neurons in the brain stem. The command neurons integrate various neural inputs, both excitatory and inhibitory, and the group of neurons generates a highly synchronous impulse. Each synchronous discharge is the command signal that is responsible for each electric organ discharge. Synchronization is mediated by electrical synapses that interconnect the cells. In the continuously discharging fishes the command neurons are probably spontaneously active, but they still receive sensory and other inputs that modulate discharge frequency. The command signal may be relaxed by one or two intervening groups of cells, the ultimate group innervating the electrocytes. (In the sternarchids the electric organ actually consists of the axons of the final neurons; Fig. 3.) Since different parts of the electric organ are at significantly different distances from the brain, compensatory mechanisms are present to adjust for the variations in nerve conduction and thus to ensure synchrony of organ discharge. For example, nerves to nearer parts may take more devious pathways, and they may conduct impulses at lower velocities.

Electrosensory systems. A knowledge of the electrosensory systems is integral to understanding the biological significance of electric organs. A number of groups of fishes, both electric and nonelectric, have receptors specialized for the detection of electric fields. These groups include all the elasmobranchs; the primitive bony fishes, lungfishes, bichirs, sturgeons, and paddlefish; the catfishes; and the two teleostean groups of fresh-water electric fishes. Electrosensory systems can be divided into two categories: those activated by fields from extrinsic sources, or passive systems, and those activated by the fishes' electric organs, or active systems. The passive systems generally are sensitive to voltages of low frequency such as are generated by respiratory movements in fishes, muscle or skin lesions, and geomagnetic sources. They can be extraordinarily sensitive. For example, elasmobranchs can respond to

voltage gradients as low as 0.01 microvolt/cm or 1 V uniformly distributed over 1000 km. The active systems utilize the electric organ discharge to test the conductivity of the surrounding medium, thereby revealing changes or inhomogeneities. The active systems are also involved in intra- and interspecific communication.

The electroreceptors are cutaneous organs distributed over the body surface. They are innervated by cranial nerves of the acousticolateralis system. The gymnotids and mormyrids have separate sets of receptors for active and passive systems; the electric organ discharges are often of high frequency with little dc component, and the two receptor systems detect electric fields over different frequency bands. The Mormyriformes have a third set of receptors, very sensitive to high frequencies, that appears to be specialized for detecting the electric discharges of other fishes. The diversity of organ discharges in terms of both frequency and pulse shape may be partly accounted for in terms of interspecific communication and interaction.

In the skate the electric organ discharge is sufficiently long-lasting to be capable of activating the passive electrosensory system. It is unknown whether their discharges are utilized for electrolocation or for communication.

Evolution. The electrosensory systems allow an explanation of the evolution of strongly electric organs. Charles Darwin supposed that a weakly electric organ would have been an intermediate stage in the gradual development of a strongly electric organ, but could see no use for a weak organ. Later, Hans Lissman, who initially described the active electrosensory system, proposed that they provide a function for Darwin's hypothesized intermediate stage. This notion is supported by the electric eel and at least one torpedinid; both have strongly and weakly electric organs, and the eel is known to utilize the weak pulses for electrolocation in the same way as the related weakly electric gymnotids. The electric catfish has a passive electrosensory system, as do most nonelectric catfish. No weakly electric catfish are known, and the hypothetical intermediate leading to *Malapterurus* is either extinct or has not yet been discovered. The electrosensory capabilities of the stargazer have not been adequately investigated.

Electrocution. A question often asked about the strongly electric fishes is how they keep from electrocuting themselves. When immersed in water, they do not appear to be at all affected by their own discharges, although their electroreceptors are undoubtedly activated. In air, when the organ is not loaded down by the conductivity of the water, they often do twitch when they discharge their organs, but as would be expected, they are much more resistant than other fishes. One factor is that their nerves and central nervous system are well insulated by many layers of connective and fatty tissue. Whether their brains are more resistant to stunning by a given current density has not been determined; certainly the excitability of individual cells is no different from that in animals in general.

Source of data. Electric fishes have proved useful in defining extremes of excitability properties. Historically they served to validate Luigi Galvani's proposal of animal electricity, and in modern times have provided, and are continuing to provide, a wealth of biophysical data of general and comparative interest. They also supply important starting materials for the biochemical characterization of excitable membranes, for example, the acetylcholine receptor at the synapse. Because evolution has led to large volumes of tissue with high concentrations of the relevant macromolecules, the electric organs provide a rich source for isolation and analysis. *See* ACETYLCHOLINE; SYNAPTIC TRANSMISSION.

Michael V. L. Bennett

Electric power generation

The production of bulk electric power for industrial, residential, and rural use. Although limited amounts of electricity can be generated by many means, including chemical reaction (as in batteries) and engine-driven generators (as in automobiles and airplanes), electric power generation generally implies large-scale production of electric power in stationary plants designed for that purpose. The generating units in these plants convert energy from falling water, coal, natural gas, oil, and nuclear fuels to electric energy. Most electric generators are driven either by hydraulic turbines, for conversion of falling water energy; or by steam or gas turbines, for conversion of fuel energy. Limited use is being made of geothermal energy, and developmental work is progressing in the use of solar energy in its various forms. *See* GENERATOR; PRIME MOVER.

Fossil-fuel and hydroelectric plants, and various types of advanced power sources, are discussed in detail below. For a discussion of nuclear power plants *see* NUCLEAR POWER; NUCLEAR REACTOR.

General Considerations

An electric load (or demand) is the power requirement of any device or equipment that converts electric energy into light, heat, or mechanical energy, or otherwise consumes electric energy as in aluminum reduction, or the power requirement of electronic and control devices. The total load on any power system is seldom constant; rather, it varies widely with hourly, weekly, monthly, or annual changes in the requirements of the area served. The minimum system load for a given period is termed the base load or the unity load-factor component. Maximum loads, resulting usually from temporary conditions, are called peak loads, and the operation of the generating plants must be closely coordinated with fluctuations in the load. The peaks, usually being of only a few hours' duration (**Figs. 1** and **2**), are frequently served by gas or oil combustion-turbine or pumped-storage hydro generating units. The pumped-storage type utilizes the most economical off-peak (typically 10 P.M. to 7 A.M.) surplus generating capacity to

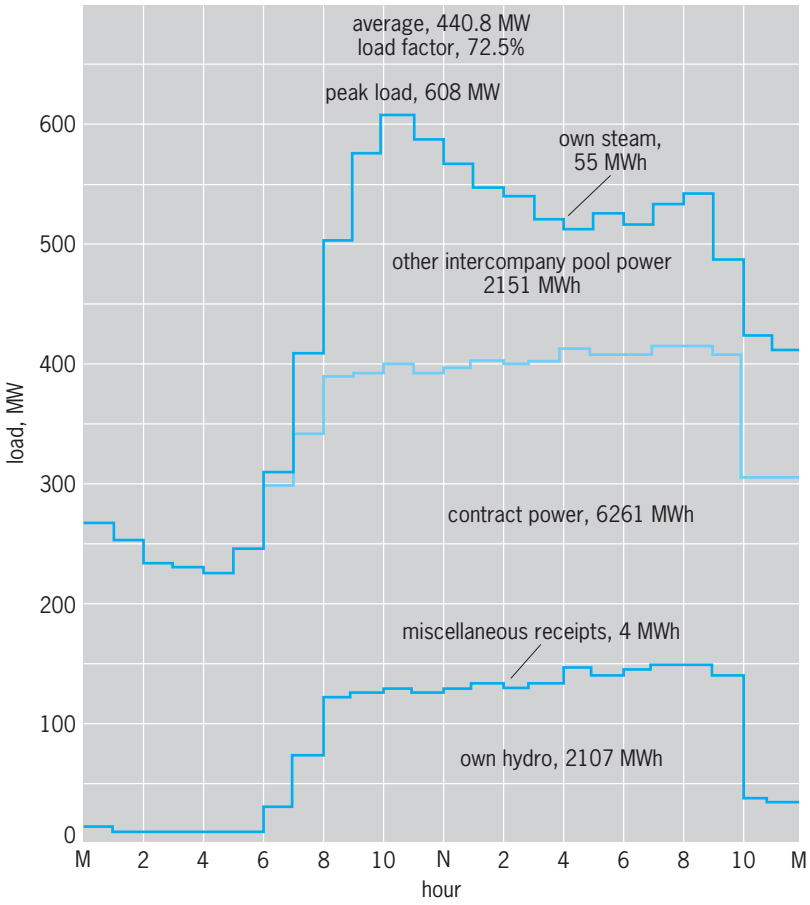


Fig. 1. Load graph indicates net system load of a metropolitan utility for typical 24-h period (midnight to midnight), totaling 10.578 MWh. Such graphs are made to forecast probable variations in power required.

released through hydraulic turbine generators during peak periods. This type of operation improves the capacity factors or relative energy outputs of base-load generating units and hence their economy of operation.

Actual variations in the load with time are recorded, and from these data load graphs are made to forecast the probable variations of load in the future. A study of hourly load graphs (Figs. 1 and 2) indicates the generation that may be required at a given hour of the day, week, or month, or under unusual weather conditions. A study of annual load graphs and forecasts indicates the rate at which new generating stations must be built; they are an inseparable part of utility operation and are the basis for decisions that profoundly affect the financial requirements and overall development of a utility.

Generating unit sizes. The size or capacity of electric utility generating units varies widely, depending upon type of unit; duty required, that is, base-, intermediate-, or peak-load service; and system size and degree of interconnection with neighboring systems. Base-load nuclear or coal-fired units may be as large as 1200 MW each, or more. Intermediate-duty generators, usually coal-, oil-, or gas-fueled steam units, are of 200 to 600 MW capacity each. Peaking units, combustion turbines or hydro, range from several tens of megawatts for the former to hundreds of megawatts for the latter. Hydro units, in both base-load and intermediate service, range in size up to 825 MW.

The total installed generating capacity of a system is typically 20 to 30% greater than the annual

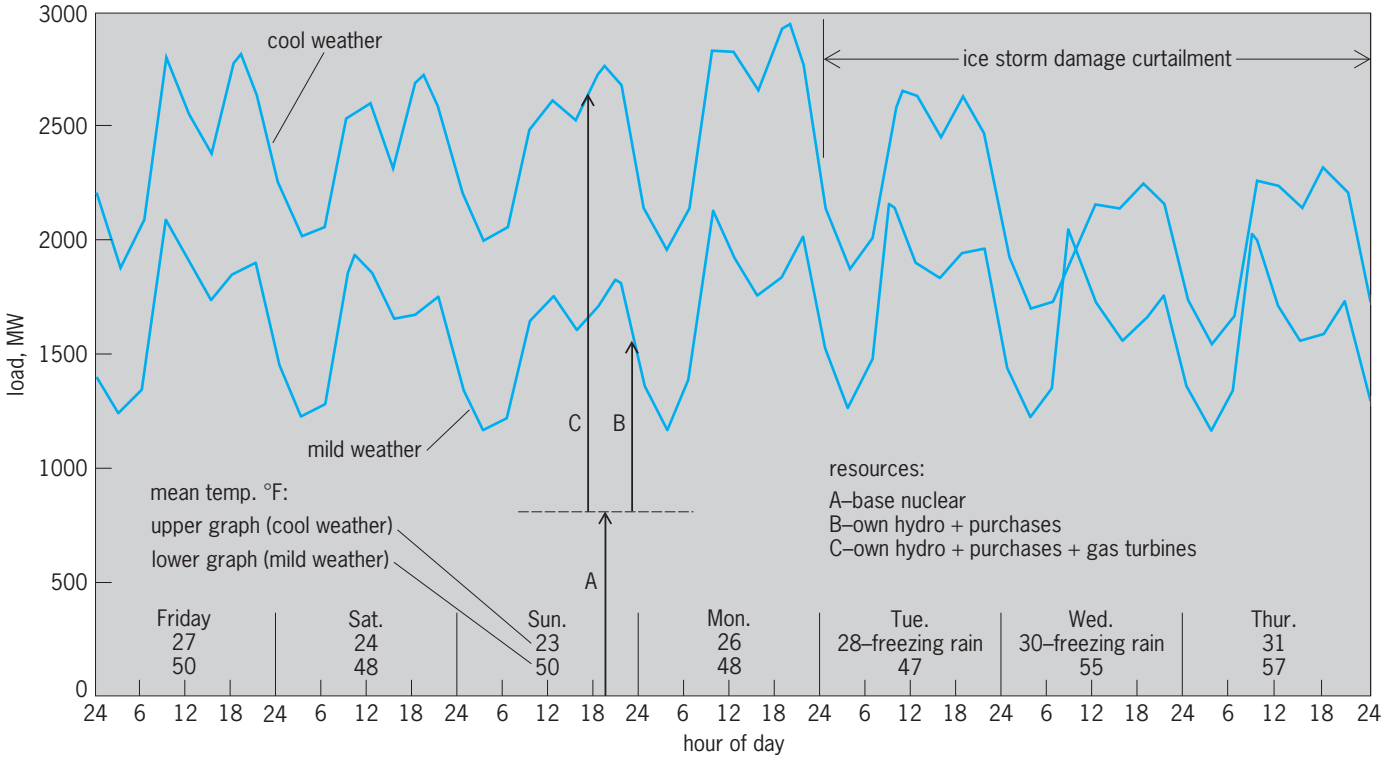


Fig. 2. Examples of northwestern United States electric utility weekly load curves showing same-year weather influence.

predicted peak load in order to provide reserves for maintenance and contingencies.

Power-plant circuits. Both main and accessory circuits in power plants can be classified as follows:

1. Main power circuits to carry the power from the generators to the step-up transformers and on to the station high-voltage terminals.
2. Auxiliary power circuits to provide power to the motors used to drive the necessary auxiliaries.
3. Control circuits for the circuit breakers and other equipment operated from the plant's control room.
4. Lighting circuits for the illumination of the plant and to provide power for portable equipment required in the upkeep and maintenance of the plant. Sometimes special circuits are installed to supply the portable power equipment.
5. Excitation circuits, which are so installed that they will receive good physical and electrical protection because reliable excitation is necessary for the operation of the plant.
6. Instrument and relay circuits to provide values of voltage, current, kilowatts, reactive kilovoltamperes, temperatures, pressures, flow rates, and so forth, and to serve the protective relays.
7. Communication circuits for both plant and system communications. Telephone, radio, transmission-line carrier, and microwave radio may be involved.

It is important that reliable power service be provided for the plant itself, and for this reason station service is usually supplied from two or more sources. To ensure adequate reliability, auxiliary power supplies are frequently provided for start-up, shutdown, and communication services.

Generator protection. Necessary devices are installed to prevent or minimize other damage in cases of equipment failure. Differential-current and ground relays detect failure of insulation, which may be due to deterioration or accidental overvoltage. Overcurrent relays detect overload currents that may lead to excessive heating; overvoltage relays prevent insulation damage. Loss-of-excitation relays may be used to warn operators of low excitation or to prevent pulling out of synchronism. Bearing and winding overheating may be detected by relays actuated by resistance devices or thermocouples. Overspeed and lubrication failure may also be detected.

Not all of these devices are used on small units or in every plant. The generator is immediately deenergized for electrical failure and shut down for any overlimit condition, all usually automatically. *See* ELECTRIC PROTECTIVE DEVICES.

Voltage regulation. This term is defined as the change in voltage for specific change in load (usually from full load to no load) expressed as percentage of normal rated voltage. The voltage of an electric generator varies with the load and power factor; consequently, some form of regulating equipment is required to maintain a reasonably constant and predetermined potential at the distribution stations or load centers. Since the inherent regulation of most alternating-current (ac) generators is rather poor (that is, high percentagewise), it is necessary to provide automatic voltage control. The rotating or magnetic amplifiers and voltage-sensitive circuits of the automatic regulators, together with the exciters, are all specially designed to respond quickly to changes in the alternator voltage and to make the necessary changes in the main exciter or excitation system output, thus providing the required adjustments in voltage. A properly designed automatic regulator acts rapidly, so that it is possible to maintain desired voltage with a rapidly fluctuating load without causing more than a momentary change in voltage even when heavy loads are thrown on or off.

In general, most modern synchronous generators have excitation systems that involve rectification of an ac output of the main or auxiliary stator windings, or other appropriate supply, using silicon controlled rectifiers or thyristors. These systems enable very precise control and high rates of response. *See* SEMICONDUCTOR RECTIFIER; VOLTAGE REGULATOR.

Generation control. Computer-assisted (or on-line controlled) load and frequency control and economic dispatch systems of generation supervision are being widely adopted, particularly for the larger new plants. Strong system interconnections greatly improve bulk power supply reliability but require special automatic controls to ensure adequate generation and transmission stability. Among the refinements found necessary in large, long-distance interconnections are special feedback controls applied to generator high-speed excitation and voltage regulator systems.

Synchronization of generators. Synchronization of a generator to a power system is the act of matching, over an appreciable period of time, the instantaneous voltage of an alternating-current generator (incoming source) to the instantaneous voltage of a power system of one or more other generators (running source), then connecting them together. In order to accomplish this ideally the following conditions must be met:

1. The effective voltage of the incoming generator must be substantially the same as that of the system.
2. In relation to each other the generator voltage and the system voltage should be essentially 180° out of phase; however, in relation to the bus to which they are connected, their voltages should be in phase.
3. The frequency of the incoming machine must be near that of the running system.
4. The voltage wave shapes should be similar.
5. The phase sequence of the incoming polyphase machine must be the same as that of the system.

Synchronizing of ac generators can be done manually or automatically. In manual synchronizing an operator controls the incoming generator while observing synchronizing lamps or meters and a synchroscope, or both. The operator closes the connecting switch or circuit breaker as the synchroscope needle slowly approaches the in-phase position.

Automatic synchronizing provides for automatically closing the breaker to connect the incoming machine to the system, after the operator has properly

adjusted voltage (field current), frequency (speed), and phasing (by lamps or synchroscope). A fully automatic synchronizer will initiate speed changes as required and may also balance voltages as required, then close the breaker at the proper time, all without attention of the operator. Automatic synchronizers can be used in unattended stations or in automatic control systems where units may be started, synchronized, and loaded on a single operator command. *See* ALTERNATING-CURRENT GENERATOR; PHASE-ANGLE MEASUREMENT; SYNCHROSCOPE.

Eugene C. Starr

Fossil-Fuel Plants

Fossil fuels are of plant or animal origin and consist of hydrogen and carbon (hydrocarbon) compounds. The most common fossil fuels are coal, oil, and natural gas. The less common ones include peat, oil shale, and biomass (wood and so forth), as well as various waste or by-products such as steel mill blast furnace gas, coke-oven gas, and refuse-derived fuels. Fossil-fuel electric power generation uses the combustion heat energy from these fuels to produce electricity. *See* COAL; COKE; FOSSIL FUEL; NATURAL GAS; OIL SHALE; PEAT; PETROLEUM.

Steam power plants. A fossil-fuel steam power plant operation essentially consists of four steps (**Fig. 3**): (1) Water is pumped at high pressure to a boiler, where (2) it is heated by fossil-fuel combustion to produce steam at high temperature and pressure. (3) This steam flows through a turbine, rotating an electric generator (connected to the turbine shaft) which converts the mechanical energy to electricity. (4) The turbine exhaust steam is condensed by using cooling water from an external source to remove the heat rejected in the condensing process. The condensed water is pumped back to the boiler to repeat the cycle. Figure 3 also shows features to increase cycle efficiency, including preheating of the boiler feedwater by using steam extracted from the turbine, and reheating the high-pressure turbine exhaust steam before it enters the intermediate-pressure turbine.

Steam cycle. Modern fossil-fuel power plants are based on a steam cycle first proposed by W. J. M. Rankine in 1908. The basic Rankine cycle underwent conceptual evolution through the 1950s that improved cycle efficiency to current levels but which also increased the complexity of building and operating a fossil-fuel power plant. Commercial development lagged behind conceptual cycle development because of the unavailability of suitable materials and fabrication techniques to accommodate the higher steam pressures and temperatures required to increase cycle efficiency. **Table 1** summarizes this steam cycle evolution with respect to turbine inlet steam pressure, turbine heat rate, and turbine cycle and overall thermal efficiencies, along with related available main steam piping materials from the 1930s through the 1950s. *See* RANKINE CYCLE; THERMODYNAMIC CYCLE; VAPOR CYCLE.

During the 1960s and 1970s, there was no significant cycle efficiency improvement, primarily due to the advent of nuclear power and the abundant supply of low-cost fossil fuel. In the 1980s, the interest in cycle efficiency improvement was renewed due to the uncertain future of nuclear energy and escalating fossil-fuel costs. However, due to the return of relative cost and price stability of fossil fuels since the late 1980s, interest in cycle efficiency

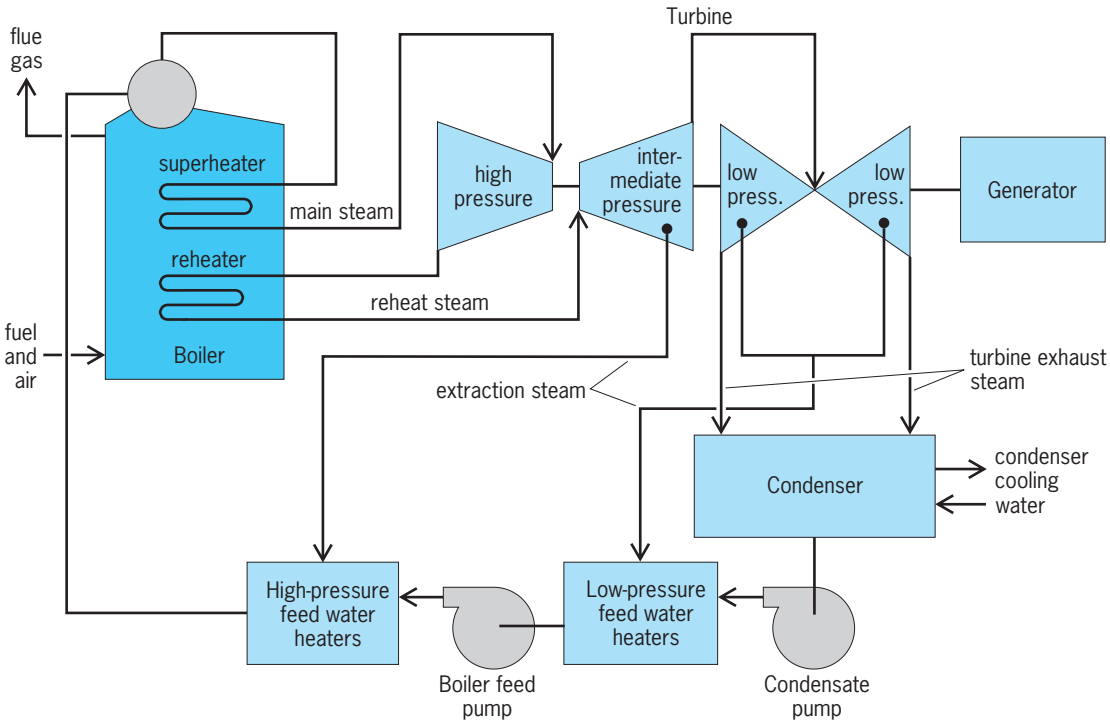


Fig. 3. Steam power plant flow diagram.

TABLE 1. Steam cycle development in fossil-fuel power plant

Years	Steam cycle, psig/°F/°F* (MPa/°C/°C)	Main steam piping material	Turbine cycle† heat rate, Btu/kWh (MJ/kWh)	Thermal efficiency (%)	
				Turbine cycle†	Overall‡
Late 1930s	1250/950 (8.7/510)	Chromium-molybdenum alloy(0.2% C, 0.5% Cr, 0.5% Mo)	9350 (9.86)	36.50	29.2
Mid 1940s	1450/1000/1000 (10.1/538/538)	Chromium-molybdenum alloy (0.2% C, 1.0% Cr, 0.5% Mo)	8150 (8.60)	41.9	33.5
Late 1940s	2000/1050/1050 (13.9/566/566)	Chromium-molybdenum alloy (0.2% C, 1.0% Cr, 0.5% Mo)	7700 (8.12)	44.3	35.5
Early 1950s	2400/1100/1050	Chromium-molybdenum	7500 (7.91)	45.5	36.4

*Main steam pressure/main steam temperature/reheat steam temperature.
†Cycle heat rate and thermal efficiency are at rated load for a condenser backpressure of 1.43 in. Hg (36.3 mmHg or 4.84 kPa) absolute and a condensate temperature of 90°F (32.2°C).
‡Based on a boiler efficiency of 85% and a power plant auxiliary power requirement of 5% of plant output.

improvement has diminished significantly. In fact, few if any large steam power plants with main steam pressure, main steam temperature, and reheat steam temperature higher than 2400 psig, 1100 °F, and 1050 °F (16.6 MPa, 593 °C, and 566 °C) respectively (the level achieved in the early 1950s) have been built since 1990. With the rapid development of Asian and Pacific rim countries including China, there is renewed interest in smaller-size 200–300-MW units, with emphasis on reducing capital cost.

Plant size and facilities. A given power plant's size is determined by the utility company's need for power as dictated by the electrical demand growth forecast. The additional generating capacity needed is matched to commercially available turbine-generator capacity sizes. The boiler is selected to suit the turbine steam flow requirements, and the remaining accessory equipment is sized to serve the needs of the resulting boiler-turbine steam cycle.

A typical large fossil-fuel power plant consists of several major facilities and equipment (**Fig. 4**), including fuel handling and processing, boiler (including furnace), turbine and electric generator, condenser and condenser heat removal system, feedwater heating and pumping system, flue gas–cleaning system, and plant controls and control system.

1. *Fuel handling*. Fuel type determines the fuel-handling system requirements. Natural gas is normally delivered by pipeline directly and fired in the furnace, needing no special handling or storage. For coal and oil, which are usually delivered by rail, ship, or pipelines, the fuel-handling system includes fuel unloading, storage, reclaiming from storage, and fuel preparation for combustion (such as coal pulverizing and pneumatic conveying to the burners, and oil heating and atomizing). Coal is usually stored in open stockpiles and oil in closed tanks. Normally, 30 to 180 days of fuel requirement is stored at the plant site to maintain operation during fluctuations in fuel delivery.

2. *Boiler*. The fuel type also determines the size, design, and operation of the furnace and boiler, which become larger and more complex (in ascending order) for gas, oil, and coal. For coal, the size and complexity also increase with the decline in coal rank and quality (in ascending order)

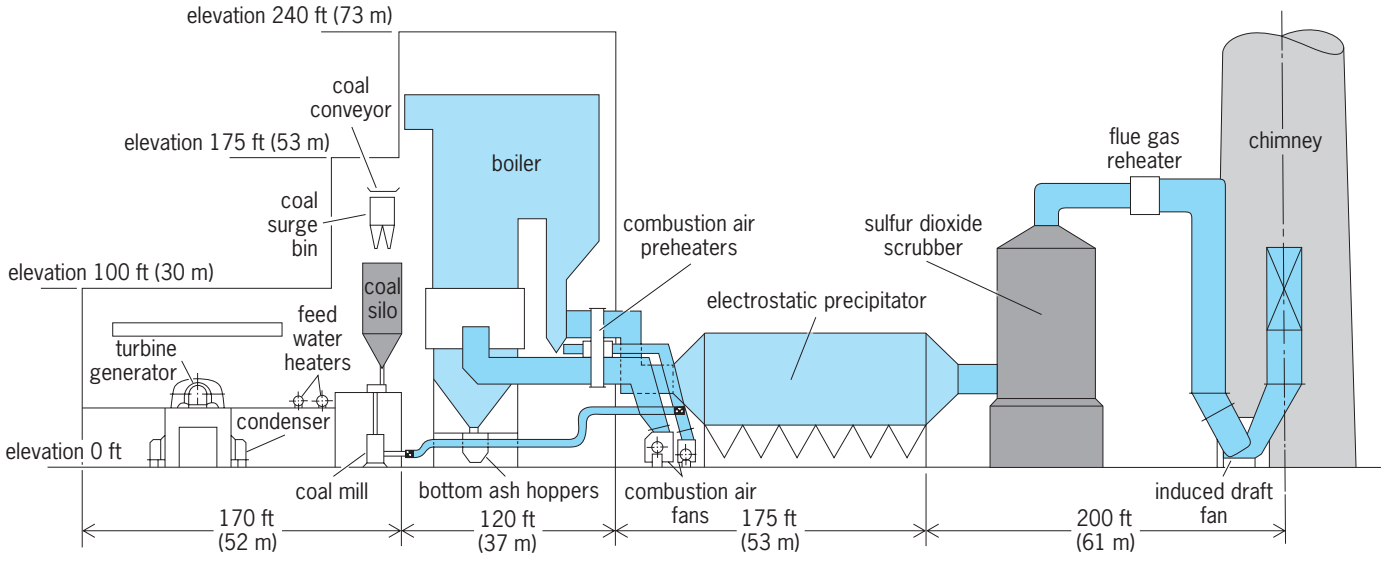


Fig. 4. Schematic cross section of a 600-MW coal-fired steam power plant.

for bituminous, subbituminous, lignite, and brown coal. Specialized boiler designs are required to handle peat, shale, and the various waste fuels. One evolving technology, fluidized-bed boilers, can effectively burn a wide range of low-quality solid fuels in the same boiler. Units of up to 200 MW are commercially proven and available. Most operating plants are in the size range of 120–180 MW. *See* BOILER; FIRE-TUBE BOILER; FLUIDIZED-BED COMBUSTION; STEAM-GENERATING FURNACE; STEAM-GENERATING UNIT; WATER-TUBE BOILER.

3. *Turbines*. Modern power plants, depending on the size and steam conditions, have single or multiple-casing turbines with high-, intermediate−, and low-pressure sections. High-pressure steam from the boiler after expanding through one or more turbines is exhausted to a condenser under vacuum, where the steam is condensed by cooling water. *See* STEAM TURBINE.

4. *Condenser*. About 50% of the fuel heat input to the boiler is ultimately rejected as waste heat in the condenser. Condenser heat-removal-system selection depends on the geography, climatology, and water availability at each plant site. There are basically two types of heat rejection systems: open (or once-through) cooling and closed cooling. The open system is the most simple and normally used where water is abundantly available. It takes water from an ocean, lake, or river and discharges the heated water to a downstream location. The closed system is used when it is determined not to return heated water to the water source due to environmental or other constraints. The closed system rejects heat to the atmosphere by using a secondary heat rejection system such as a cooling tower or pond. However, closed-system operation involves evaporating a portion of the water circulated, requiring a water supply capable of restoring the quantity of water evaporated. Dry cooling (with air condensers) is less efficient, but is used in arid areas where the water supply is limited. The temperature of the cooling medium (water or air) at the condenser inlet determines the cycle thermal efficiency by establishing the steam condensation temperature. *See* COOLING TOWER; STEAM CONDENSER; VAPOR CONDENSER.

5. *Feedwater heating*. The condensed steam is pumped back to the boiler inlet to complete the cycle, with heat from turbine extraction steam transferred along the way to increase cycle efficiency. Condensate pumps deliver feedwater from the condenser through low-pressure feedwater heaters to the boiler feed pumps, which increase the feedwater pressure and pump it through high-pressure feedwater heaters to the boiler inlet. A modern power plant typically has six feedwater heaters, which can be increased in number to improve cycle efficiency, but with offsetting increases in plant complexity and cost.

6. *Flue gas and liquid waste*. Fossil-fuel power plants produce gaseous and liquid wastes which are treated prior to discharge to minimize effects on the environment. In many countries, environmental regulations define the extent of flue gas and liquid waste treatment required.

Major pollutants in the flue gas include particulates (except for natural gas fuel), sulfur dioxide, and nitrogen oxides. Incombustible matter in coal and oil yields particulates upon combustion which are collected in devices such as electrostatic precipitators and fabric filters (baghouses). These devices are capable of removing more than 99% of the particulate matter. *See* AIR FILTER; ELECTROSTATIC PRECIPITATOR.

A variety of chemical processes is available for the removal of sulfur dioxide from the flue gas. Flue gas desulfurization systems use alkaline chemicals such as lime, limestone, sodium, or magnesium compounds in solution, slurry, or dry form to react with and remove sulfur dioxide. The choice depends on the fuel sulfur content, performance requirement, capital and operating costs, and market for the desulfurization by-products. The throw-away wet limestone slurry process is the most widely used in the United States due to the limited market for by-products. A similar process producing gypsum as the by-product is common in Europe and Japan. *See* GAS ABSORPTION OPERATIONS.

Nitrogen oxides emission can be reduced by combustion modifications which limit their formation during combustion. These techniques involve reduction of flame temperature by prolonging the combustion time (staged combustion) and reducing the combustion air quantity (low–nitrogen oxides burners).

Processes to remove nitrogen oxides after their formation involve injection of ammonia into the hot flue gas in the presence or absence of catalysts. They are referred to as the selective catalytic reduction and noncatalytic reduction processes, respectively. The selective catalytic reduction process is widely used in Japan and Europe, particularly in Germany, in both large-scale electric utilities and in small-scale industrial boilers. The noncatalytic reduction process is used in the United States in some small-scale industrial boilers. Typical nitrogen oxide reductions with the selective catalytic reduction process range from 70 to 90% and with the noncatalytic reduction process from 30 to 70%.

Waste liquid pollutants include suspended solids, toxic metal compounds, dissolved chlorine, and high-acidity or high-alkalinity, high-temperature, and oily compounds. Treatment prior to discharge or reuse usually involves one or more of the following: neutralization, sedimentation, clarification, filtration, dechlorination, and incineration. *See* CLARIFICATION; FILTRATION; SEDIMENTATION (INDUSTRY).

7. *Plant controls*. The power plant size and steam cycle pressure and temperature determine plant controls and control systems requirements. In ascending order of complexity and sophistication, these include pneumatic analog controls, electronic analog controls, hard-wired digital controls, and distributed digital control with microprocessors and minicomputers. Modern power plants typically use pneumatic local controls and electronic analog remote

TABLE 2. Typical cost distribution of major systems and equipment for a coal-fired steam power plant*

Systems and equipment	% of capital cost
Boiler (including furnace)	30
Turbine and electric generator	18
Flue gas desulfurization system	16
Particulate collection system	7
Coal-handling equipment	5
Condenser and cooling power	5
Ash-handling equipment	3
Pumps	3
Chimney	2
Water treatment	1
Other mechanical equipment	10

*Excluding land, site preparation, and building costs.

controls with electromechanical or solid-state interlocking devices and variable-speed motor drives. *See* CONTROL SYSTEMS; DIGITAL CONTROL; DISTRIBUTED SYSTEMS (CONTROL SYSTEMS); MICROCOMPUTER; MICROPROCESSOR; PROCESS CONTROL.

A typical cost distribution by major systems and equipment for a coal-fired steam power plant is given in **Table 2**.

Gas turbine plants. Power plants with gas turbine–driven electric generators are often used to meet short-term peaks in electrical demand. They are generally small (up to 150–200 MW), have a low thermal efficiency, require clean fuels such as natural gas or light oils, and consequently are more expensive to operate. However, these plants have a relatively low capital cost and can be used in areas with low power demand and limited water availability. Diesel engine–driven electric generators are also used under similar conditions. *See* DIESEL ENGINE.

Gas turbine power plants use atmospheric air as the working medium, operating on an open cycle where air is taken from and discharged to the atmosphere and is not recycled. In a simple gas turbine plant (**Fig. 5**), air is compressed and fuel is injected into the compressed air and burned in a combustion chamber. The combustion products expand through a gas turbine and exhaust to the atmosphere. Variations of this basic operation to increase cycle efficiency include regeneration (where exhaust from the turbine is used to preheat the compressed air before it enters the combustion chamber) and reheating (where the combustion gases are expanded in more than one stage and are reheated between stages).

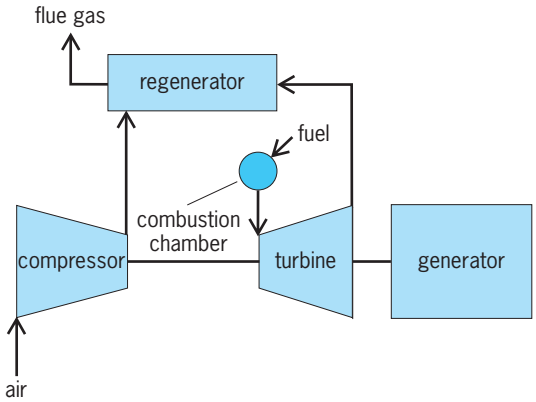


Fig. 5. Gas turbine plant flow diagram.

Combined-cycle plants use both gas and steam turbines and offer overall thermal efficiency improvement and operating economy. The gas turbine exhaust heat is used to generate steam for operating steam turbine generators. Combined-cycle plants with overall thermal efficiencies greater than 50%, where operating economies justify it, are commercially proven and available. Some very large plants in the 500–1500-MW size range are in commercial operation. *See* GAS TURBINE; STEAM TURBINE.

Cogeneration is another method of improving overall thermal efficiency, where steam exhausted from the steam turbine generator is used in industrial processes. Substantial energy is saved as heat rejected in the condenser is put to use. *See* COGENERATION.

Advanced concepts. Advanced fossil-fuel power generation concepts such as integrated coal gasification combined cycle and magnetohydrodynamics offer the promise of significant overall thermal efficiency improvement. In the integrated coal gasification combined cycle, gas produced from coal (after proper purification) is burned and expanded through a gas turbine, with the exhaust gas heat used to generate steam for a steam turbine generator. In the magnetohydrodynamic process, an electrically conducting medium such as hot flue gas from a coal combustor (seeded with alkali metal salts to increase the flue gas electrical conductivity) is passed through a magnetic field (magnetohydrodynamic channel) to produce electricity. The magnetohydrodynamic channel exhaust gas is then used to generate steam in a conventional steam power plant. The alkali salts are recovered and reused. Power generation based on these concepts can reach a thermal efficiency of 48–52%, whereas a steam turbine plant has an efficiency of 30–40% and a conventional gas turbine plant 22–28%. *See* COAL GASIFICATION; MAGNETOHYDRODYNAMIC POWER GENERATOR.

Maris T. Fravel; Natarajan Sekhar

Hydroelectric Plants

Hydroelectric generation is an attractive source of electric power because it is a renewable resource and a nonconsumptive use. In the broadest sense, hydroelectric power is a form of solar power; the resource is renewed by the solar cycle in which water is evaporated from the oceans, transported by clouds, and falls as precipitation on the landmasses, and returns through rivers to the ocean, generating power on the way. Hydroelectric power can be defined as the generation of electricity by flowing water; potential energy from the weight of water falling through a vertical distance is converted to electrical energy. The amount of electric power P that can be

generated is given by the equation below, where Q

$$P = \frac{Qb}{k} \qquad \text{kW}$$

is the volume flow of water, b is the height through which the water falls, and k is a constant equal to 11.8 when Q is given in cubic feet and b in feet (English units), and equal to 0.102 when Q is in cubic meters and b is in meters (SI units).

Development. In the United States, the first hydroelectric generating station was built in 1882 in Appleton, Wisconsin. The total power output was sufficient to operate some 250 electric lights. Rapid expansion followed, and by 1900 about 300 hydroelectric plants were in operation worldwide. This rapid expansion was a consequence of the development in the nineteenth century of the hydraulic reaction turbine. Also, by the same date, the concept of pumped storage had been introduced in Switzerland, but extensive development in the United States did not occur until the advent of the reversible pump–turbine design in the early 1950s.

Major components. A typical hydroelectric development consists of a dam to divert or store water; waterways such as a forebay, canals, tunnels, and penstocks to deliver the water to the hydraulic turbine and a draft tube, tunnel, or tailrace to return the water to the stream; hydraulic turbines and governors; generators and exciters; electrical controls to provide protection and to regulate frequency, voltage, and power flow; a powerhouse to enclose the machinery and equipment; transformers and switching equipment; and a transmission line to deliver the power to the load center for ultimate distribution (**Fig. 6**).

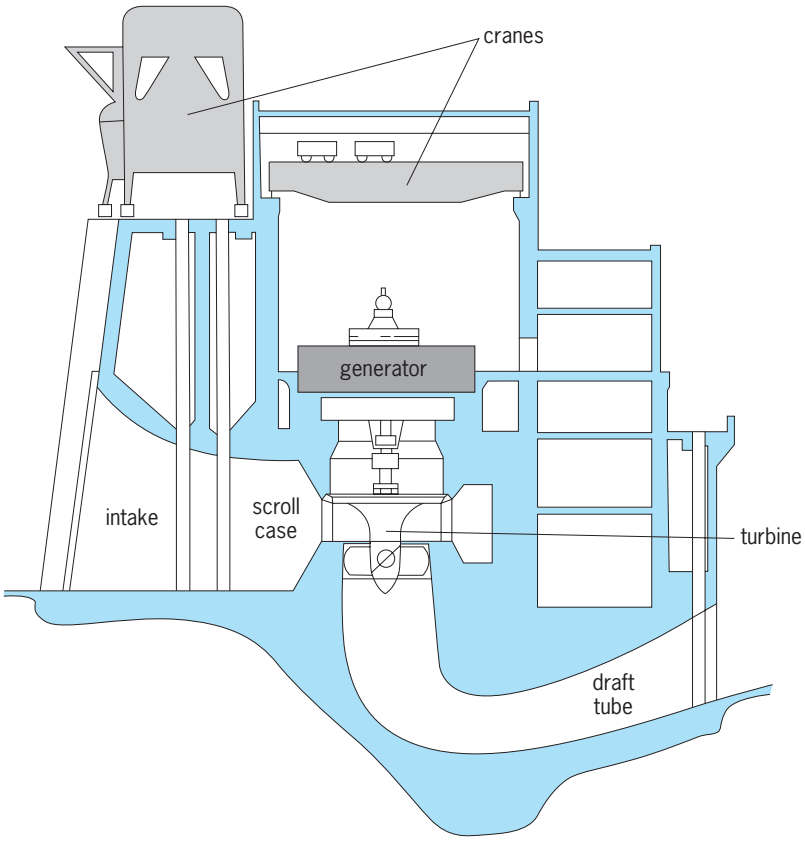


Fig. 6. Cross section of a hydroelectric powerhouse. Arrangement is typical for an adjustable propeller-type reaction turbine. The intake and powerhouse are integral; waterways consist of intake, scroll case, and draft tube. Turbine and generator are set with the axis vertical.

Dam. Dams are functionally important as they regulate the water supply to the plant and also are frequently the largest single element of cost in a hydroelectric development. Dams may be classified according to materials or type of construction, height, or use. *See* DAM.

Waterways. Waterways may be very short or as much as 10 mi (16 km) or more in length. The type of waterway is selected on the basis of length, pressure, plant arrangement, and economy. Very short waterways are required in installations where the powerhouse and the dam are integral structures; then the waterway is short and the length dictated by the plant arrangement. In other low-head plants, canals and forebays are frequently used. Conversely, very long waterways are also common in the developments where the head is great. Tunnels and penstocks are commonly used in high-head developments to deliver the water to a powerhouse some distance from the impoundment.

Some of the important aspects of the hydraulic design of waterways are provision for dewatering to facilitate inspection and maintenance, provision for water hammer resulting from surges caused by sudden start-up or shutdown, and optimization of hydraulic losses for an economic design. *See* WATER HAMMER.

An intake structure may be provided at the entrance to the waterway to reduce the hydraulic losses and to provide a closure gate for dewatering the waterway. Surge chambers are required at abrupt vertical changes in the direction of the waterways to provide relief of surges on shutdown and to prevent separation of the water column on start-up. When tunnels are used for the tailrace, surge chambers may be required for surge suppression.

Penstocks are defined as large-diameter pipelines for delivering water to the turbine or the steel-lined part of a tunnel adjacent to the powerhouse. Materials other than steel have been used in penstock construction; in the past, wood staves were common; and fiber glass is now used selectively.

Turbines. Hydraulic turbines are used to drive an electric generator, thereby converting the potential energy of the water to electrical energy. Reaction- or impulse-type turbines are used depending upon the head available, with a reaction turbine being used for heads under 1200 ft (365 m) and an impulse turbine for greater heads. Reaction turbines operate in a closed system from headwater to tailwater. For impulse turbines, only the portion to the turbine is closed and under pressure; from that point, flow is by gravity. As water passes through the reaction turbine, conversion from potential to mechanical energy takes place and, through a shaft,

drives the generator. For an impulse turbine, the potential energy of the water is first converted to kinetic energy in high-velocity jets which impinge on buckets at the circumference of the impulse wheel. There the kinetic energy is converted to mechanical energy for driving the generator. Reaction turbines are classified as Francis or propeller types, and propeller types are further categorized as fixed-blade and adjustable-blade, or Kaplan-type. Both classes of reaction turbines are normally installed with the axis vertical. *See* HYDRAULIC TURBINE; IMPULSE TURBINE; REACTION TURBINE.

Generator. In vertical installations, the generator is located above the turbine. When the turbine rotates, so does the rotor of the synchronous generator which produces electrical energy in a conventional way. An operating speed is chosen, as well as a generator with the correct number of poles to develop the desired frequency. From the generator, electricity is carried by buses to switching apparatus and then to transformers, where the voltage is stepped up for transmission. *See* ALTERNATING-CURRENT GENERATOR; HYDROELECTRIC GENERATOR.

Bulb units. Bulb units, which take their name from their shape, are an adaptation of the propeller unit and differ from the conventional unit in two significant respects: the alignment is horizontal, and the generator is totally enclosed within the bulb and mounted in the water passage. Water flows to the unit in a horizontal conduit and passes in the annular space between the bulb and the walls of the conduit. As the water passes the bulb, it drives the propeller and exits from the unit through a divergent horizontal water passage. Flow is controlled in a manner similar to the vertical units by wicket gates.

Classification of developments. Hydroelectric developments may be classified according to purpose, use, location of powerhouse, and head. The descriptive classifications commonly used are the run-of-river or storage, high- or low-head, base- or peak-load, multipurpose or single-purpose, and conventional or pumped-storage. Powerhouses may be indoor or outdoor, aboveground (surface) or underground. A number of these classifications may apply to any specific project such as conventional, run-of-river with surface, outdoor powerhouse, or pumped-storage with an underground powerhouse.

Multipurpose projects. A multipurpose project has a number of uses, some of which may be governed by competing criteria. Hydroelectric generation is generally compatible with other uses because it is nonconsumptive. Such projects normally require large reservoirs to accommodate the multiple uses. The most frequent purposes are irrigation, power, flood control, navigation, water supply, and recreation. Some adjustment must be made in either the amount of or the timing of the generation for compatibility with other uses. On major river systems in the United States, multipurpose projects have been developed in conjunction with navigation, flood control, and power. In the semiarid states, irrigation has played a dominant role in multipurpose developments.

Pumped storage. Pumped storage is a process for converting large quantities of electrical energy to potential energy by pumping water to a higher elevation where it can be stored indefinitely, then released to pass through hydraulic turbines and generate electrical energy on demand. Storage is desirable, as the consumption of electricity is highly variable according to the time of day or week, as well as seasonally. Consequently, there is excess generating capacity at night and on weekends. This excess capacity can be used to generate energy for pumping, hence storage. Normally, pumping energy can be obtained cheaply at night or on weekends, and its value will be upgraded when used for daytime peak loads.

In a typical operation, water is pumped at night or on weekends from a lower to an upper reservoir, where it is stored. The water can be retained indefinitely without deterioration or significant loss. During the daylight hours when the loads are greatest, stored water is released to flow from the upper to the lower reservoir through hydraulic turbines which generate electricity. No water is consumed in either the pumping or the generating cycles. A hydroelectric pumped-storage development is similar to a typical hydro installation, except that it has two reservoirs of essentially equal size situated to maximize their difference in elevation. Also, since it is a closed system, it need not be located directly on a large stream but must have a source for initial filling and to make up for the losses of evaporation and seepage. The second principal difference is the pump-turbine which is capable of rotation in either direction and acts as a pump in one direction and a turbine in the other. *See* ENERGY STORAGE; WATERPOWER.

Dwight L. Glasscock

Advanced Power Sources

Interest has developed in alternative or advanced power sources. Most of this attention has focused on solar or solar-related technologies, although several additional options including geothermal and fusion have been examined. Genuine interest in alternatives developed following the oil embargo in 1973.

The supplies of fossil and nuclear fuels are classified as nonrenewable and therefore are gradually being depleted. The advanced power sources of interest are those which are renewable and linked to the Sun, naturally occurring steam or hot water in the Earth's crust, or fusion. Since all solar sources suffer from intermittent availability, they require special consideration. As these various options are considered, three significant problems must be addressed. First, availability and the primary form of occurrence of the resource must be assessed. Second, a conversion system must be developed to transform the source into a usable form of energy, typically electricity. Third, the electricity production source must be effectively and efficiently appended to the existing electrical supply system.

Solar power. The solar options consisting of photovoltaics, wind, solar thermal process, and biomass will be discussed below. Other solar alternatives,

including solar ponds, passive solar, and ocean-thermal energy conversion, and other ocean power systems such as tide and wave power will not be treated. *See* TIDAL POWER.

Photovoltaics. The collection of solar energy takes many forms, but one of its most desirable configurations is direct conversion to electricity. The device configurations are modular in form, making them convenient from a mass-production viewpoint.

The heart of the photovoltaic system is a thin flat layer of semiconductor material. When the material is struck by sunlight, electrons are freed, producing an electric current. The direct-current (dc) power is passed through a dc load, into a storage battery, or converted to alternating current (ac) for general use in electric utility grids. Typically, individual solar cells are ganged together to form photovoltaic modules about 5 ft^2 (0.5 m^2) in size with a generating capacity of about 50 W. Typically, about half the cost of a solar system lies with the solar cell modules, and the remainder is directed toward power conditioning, electrical wiring, installation, and site preparation.

Typically, silicon or gallium arsenide are used to fabricate solar cells, although other semiconductor materials are being developed. Silicon technology is the most advanced because it is the least expensive, and takes many different forms including single-crystal, polycrystal, and amorphous configurations. The efficiency of these configurations ranges from about 6 to about 14% in module form. Amorphous silicon offers significant potential because of low manufacturing cost.

One way to advance the power yield from photovoltaics is to use concentrator systems. Special lenses and mirrors are used to focus sunlight on the cells, thereby raising the output of a module. Concentrator systems result in larger electric current which increases cell losses but improves overall economic efficiency of the system. Concentrator technology can raise module efficiencies to about 20%. Photovoltaic systems have potential application at the household level; however, the best near-term application is in array configurations for utility application. *See* SOLAR CELL.

Wind. The use of wind energy dates back to sailing ships and windmills. Today the interest in wind is for electricity generation by wind turbines. Wind energy conversion systems consist of several major subsystems. Among the most important is the mechanical system and electromechanical rotary converter. A wind energy conversion system is designed to rotate at either constant or variable speed as the wind varies. The variable-speed system usually offers high wind-collection efficiency; however, constant-speed units permit simpler electrical systems.

There are a multitude of aeroturbine designs available, with each striving to enhance the wind to mechanical power conversion effectiveness. A measure of this effectiveness is the power coefficient (C_p). The power coefficient is expressed as a function of the tip speed ratio, that is, the ratio of the blade tip speed to the wind speed. In variable-speed systems, C_p is tuned to its optimum value, whereas constant-speed systems cannot operate at optimum C_p over the entire wind speed range. This does not suggest that variable speed is a panacea, since it exposes the mechanical drive train and tower to potential resonance problems and requires more power conditioning of the electrical output.

The characteristics of the aeroturbine and the electric generator are not totally compatible; therefore, an element is required to effectively tie these two components. The mechanical interface has two major assignments. First, aeroturbines usually rotate at low speeds, which are not desirable for electrical production. Typically, a transmission is used to convert low-speed, high-torque mechanical power into high-speed, low-torque mechanical power for electrical conversion. Second, the mechanical interface can be used to regulate the drive train shaft stiffness, and thus shape its dynamic performance.

Ultimately the mechanical energy that has been produced by the aeroturbine and conditioned by the mechanical interface must be converted to an electrical form. In general, the type of wind system, constant- or variable-speed, determines the type of electric generator—synchronous, induction, or dc. Variable-speed systems tend to produce initially poor-quality ac or dc electric power, and then use power conditioners in the electrical interface to enhance it. Usually, constant-speed systems produce high-quality electric power at the generator terminals.

Further development of wind turbine technology is required to provide economical systems; however, wind electric systems show promise. Systems under development will produce as much 4–5 MW per machine. Typically, units larger than 200 kW are used by utilities. *See* WIND POWER.

Solar thermal. The process of collecting and concentrating solar energy at a focal point is known as solar thermal. There are two broad classes of solar thermal collection systems: power tower and point- and line-focus. The power tower is a large central power-generating system which shows significant promise, while the point- and line-focus systems are smaller and would be deployed as so-called dispersed sources. In both cases, solar energy is focused at a point or a line and heat is transferred to a working fluid.

The power tower concept has received significant attention because of its potentially large power production capability. A 10-MW demonstration plant, called Solar One, developed by the U.S. Department of Energy near Barstow, California, consists of 1818 heliostats, each containing 12 separate mirror facets designed to aim their energy at a central receiver on top of a structural steel tower. The heliostats are computer-controlled and adjust during the day to track the Sun so that optimal collection efficiency is maintained. The working fluid in this system is water, which is converted to steam and then passed through

a conventional steam turbine for eventual conversion to electricity. One significant application for the solar thermal system is as a repowering plant. Repowering is a term applied to the concept of adding a solar thermal plant to a conventional power plant to supplement steam production.

The point- and line-focus systems are smaller facilities which produce power at lower temperatures (200–600°F or 100–300°C). A typical point system is a parabolic dish which concentrates sunlight on a receiver a few feet away. Line-focus or trough systems consist of a series of cylindrical or parabolic trough lines with mirrors to collect and concentrate the Sun's radiation. Both the point- and line-focus systems could be used in smaller cogeneration facilities for commercial and industry electric power systems. *See* SOLAR ENERGY.

Biomass. Biofuel applications, involving consumption of wood products and garbage, have existed for many years. Perhaps the single largest hurdle for this application of wood and other products is the creation of effective distribution and supply systems. In general, these materials are low-density, thus causing significant transportation problems in delivering them to the chosen site for consumption.

An alternative to direct burning of biofuels is conversion to combustible liquids, gases, and solids. One process is classified as thermochemical—pyrolysis and gasification. Pyrolysis is a low-temperature process for forming gases, liquids, and solids by heating them in the absence of air, while gasification is a high-temperature process for producing gaseous fuels in the presence of an oxidant (air). A second class of processes for converting biomass to an alternative form is biochemical, which is better known as fermentation, that is, microbial transformation of organic feed materials that takes place without oxygen and produces alcohols and organic chemicals, such as methane, ethanol, acetic acid, and acetone. *See* FERMENTATION; PYROLYSIS.

The typical process of converting solar radiation to electricity by using vegetation as the collector is very inefficient (**Table 3**). It is clear that direct conversion by means other than vegetation is desirable. *See* BIOMASS.

Geothermal power. The natural emissions of steam (geysers), hot springs, and volcanoes represent potential sources of electricity production and are indicative of the steam and hot water potential embedded in the Earth's crust. Geothermal energy manifests itself in three basic types: hydrothermal, geopressured, and petrothermal. Hydrothermal is the natural production of steam when water is vaporized by coming into contact with hot rock. In geopressured systems, hot water is generated in deep reservoirs embedded in sand and shale formations within the Earth. The hot water is trapped and pressurized and is saturated with natural gas. Petrothermal systems consist of hot rock at the Earth's surface which could have a fluid injected into it and pumped out again, to extract theral energy. The amount of geothermal potential for electricity production is estimated at 220×10^{15} Btu (2.3×10^{20} joules), which is about three times the United States' energy consumption in 1980.

Presently, most geothermal power-producing facilities are limited to application of geysers or so-called dry steam sources. Since dry steam represents less than 1% of the total geothermal potential, other systems are under development to harness the Earth's natural heat. The largest near-term potential appears to be the hydrothermal resource. The present approach of converting high-temperature (above 410°F or 210°C) water is direct-flash technology. Hot water is extracted from the earth and its pressure is dropped, causing the water to vaporize (flash-boil). The steam is applied to conventional steam turbine technology. The overall efficiency of this process requires improvement since significant energy is lost when flashing the hydrothermal fluid. Since the hot water arrives from the geothermal well as a two-phase mixture of steam and water, a centrifugal process is under development to separate steam and water, capturing natural steam and optimizing conversion of the hot water to steam.

Since much of the geothermal potential occurs at lower temperatures, below 410°F (210°C), a process is being developed to capture this energy. The process is called binary since hot geothermal energy is used to vaporize a secondary fluid such as a hydrocarbon like isobutane or isopentain which boils at a lower temperature. The heat from the geothermal source is transferred to the secondary fluid through a heat exchanger, which in turn produces steam for a conventional steam cycle.

Although the various geothermal sources follow different paths to the production of steam, several common problems exist. Most hydrothermal fluids contain dissolved minerals. As the temperature of the fluid drops, the minerals precipitate out causing scaling in most flow channels and thus reducing the efficiency and effectiveness of the system. The corrosive nature of the brine hydrothermal fluids is a general problem. Perhaps the other general problem is the noncondensable gases found dissolved in the hydrothermal fluids. These gases reduce thermal efficiency, accelerate equipment wear, and cause environmental concerns. *See* GEOTHERMAL POWER.

TABLE 3. Typical energy losses in using vegetation for generation of electricity

Energy losses	% of energy in incident solar radiation
Due to leaf reflection and ineffective absorption	9
Due to invisible light, unusable for photosynthesis	55
Due to photosynthetic conversion	31
Due to plant respiration and metabolism	2
Due to conversion to electricity	2
Available as electricity	1

Fusion power. The use of nuclear energy as a power source has been confined to nuclear fission or the splitting of atomic nuclei for energy production. In the process of fusion, atomic nuclei merge by overcoming their normal electrostatic repulsion. When the nuclei fuse, energy is released as a result of the freeing of fast-moving nuclear particles. The Sun is a natural fusion reactor and is indicative of the potential from this source of energy.

Atomic mixtures used in fusion, called plasmas, consist of positively and negatively charged particles resulting from heating a gas to a high temperature. Fusion requires that the working plasma have not only extreme temperature to free the nuclei but also a sufficiently dense or compact environment to merge them. As a gas heats up, it will expand unless confined, resulting in a very low-density mixture that will not allow nuclear fusion. A successful fusion effort must therefore combine both high temperature and confinement.

The Sun accomplishes the process of fusion with its own high temperature (15×10^6°C) and its enormous mass which holds the plasma by gravity at proper densities. On Earth, such gravitational systems are not possible and thus higher temperatures will be necessary. The development of fusion reactors has concentrated on the two problems of temperature and confinement. Heating appears a less formidable problem than confinement. Electric currents, radio-frequency waves, and neutral-beam heaters have been effective approaches to the heating problem. Containment systems have focused on two approaches: magnetic and inertial.

The leading contender in magnetic confinement is a doughnut-shaped device called a tokamak. A magnetic field is used to control the plasma by deflecting the charged particles into a cylindrical or magnetic bottle and preventing them from striking the walls of the containment vessel. Such an occurrence would drop the temperature of the plasma, preventing the fusion process. In inertial confinement, a pulsed-energy source, called a drive, is used to compact and heat the fusion fuel in a single step. This results in rapid burning of the fuel to yield energy release. *See* ELECTRIC POWER SYSTEMS; NUCLEAR FUSION.

Thomas W. Reddoch

Bibliography. P. N. Cheremisinoff, *Air Pollution Control and Design for Industry*, 1993; D. G. Fink and H. W. Beaty (eds.), *Standard Handbook for Electrical Engineers*, 14th ed., 2000; L. L. Freris, *Principles of Wind Energy Conversion Systems*, 1990; General Electric Co., *Updating Excitation Systems*, Publ. GET-6675B, 1986; K. W. Li and A. P. Priddy, *Power Plant System Design*, 1985; P. Longrigg and E. H. Buell, *Electric Power from Renewable Energy Technologies*, Solar Energy Research Institute, May 1983; A. J. Pansini and K. D. Smalling, *Guide to Electric Power Generation*, 1994; J. G. Singer, *Combustion-Fossil Power Systems*, 4th ed., 1993; S. C. Stultz and J. B. Kitto, *Steam: Its Generation and Use*, 40th ed., 1990; R. Van Overstraeten and R. P. Mertens, *Physics, Technology and Use of Photovoltaics*, 1986; C. L. Wadhwa, *Electrical Power Systems*, 2d ed., 1992; J. Weisman and L. E. Eckart, *Modern Power Plant Engineering*, 1985; J. H. Willenbrock and H. R. Thomas (eds.), *Planning, Engineering and Construction of Electric Power Generation Facilities*, 1980; A. J. Wood and B. J. Wollenberg, *Power Generation, Operation, and Control*, 2d ed., 1996; K. E. Zweibel, *Harnessing Solar Power: The Photovoltaics Challenge*, 1990.

Electric power measurement

The measurement of the time rate at which electrical energy is being transmitted or dissipated in an electrical system. The potential difference in volts between two points is equal to the energy per unit charge (in joules/coulomb) which is required to move electric charge between the points. Since the electric current measures the charge per unit time (in coulombs/second), the electric power p is given by the product of the current i and the voltage v (in joules/second = watts), as in Eq. (1).

$$p = vi \quad \text{(watts)} \tag{1}$$

See ELECTRIC CURRENT; ELECTRICAL UNITS AND STANDARDS; POWER.

Alternate forms of the basic definition can be obtained by using Ohm's law, which states that the voltage across a pure resistance is proportional to the current through the element. This results in Eq. (2),

$$p = i^2R = \frac{v^2}{R} \quad \text{(watts)} \tag{2}$$

where R is the resistance of the element and i and v are the current through and voltage across the resistive element. Other commonly used units for electric power are milliwatts (1 mW = 10^{-3} W), kilowatts (1 kW = 10^3 W), megawatts (1 MW = 10^6 W), and, in electromechanical systems, horsepower (1 hp = 746 W). *See* ELECTRICAL RESISTANCE; OHM'S LAW.

These fundamental expressions yield the instantaneous power as a function of time. In the dc case where v and i are each constant, the instantaneous power is also constant. In all other cases where v or i or both are time-varying, the instantaneous power is also time-varying. When the voltage and current are periodic with the same fundamental frequency, the instantaneous power is also periodic with twice the fundamental frequency. In this case a much more significant quantity is the average power defined by Eq. (3), where T is the period of the periodic wave.

$$P = \frac{1}{T}\int_t^{t+T} vi\,dt \quad \text{(watts)} \tag{3}$$

The average power is usually the quantity of interest since in most cases the electric power is converted to some other form such as heat or mechanical power and the rapid fluctuations of the power are smoothed by the thermal or mechanical inertia of the output

system. The remainder of this article is concerned primarily with average power in systems with periodic voltages and currents.

Sinusoidal ac waves. A number of basic concepts in electric power measurement are directly related to the behavior of systems where the voltages and currents are simple sinusoids.

Average power. In a circuit containing only resistance, the voltage and current waves have the same points of zero crossing and are said to be in phase. The power curve, obtained by directly multiplying the instantaneous values of voltage and current, is a double-frequency sinusoidal wave which is entirely positive. It is clear from the symmetry of the power wave that the average power over one cycle is one-half of the peak power.

When a circuit contains both resistance and inductance, the current lags behind the voltage wave. The power curve has the same double-frequency sinusoidal shape but is shifted downward so that over a portion of the cycle the instantaneous power is negative. During this part of the cycle, power is actually being delivered back to the source from the magnetic field of the inductance. The average of the double-frequency power wave is less than one-half of the peak because of the portion which has negative values.

Quantitative relationships describing such results can be obtained by writing the voltage and current waves as Eqs. (4) and (5), where V_m and I_m are max-

$$v = V_m \sin 2\pi ft = \sqrt{2}\, V \sin 2\pi ft \quad \text{(volts)} \quad (4)$$

$$i = I_m \sin (2\pi ft - \phi) = \sqrt{2}\, I \sin (2\pi ft - \phi) \quad \text{(amperes)} \quad (5)$$

imum values, V and I are root-mean-square (rms) or effective values, f is the frequency, and ϕ is the angle by which i lags v. The instantaneous power is then given by Eq. (6), which can be rewritten as Eq. (7), from which the average power is given by Eq. (8).

$$p = vi = 2\, VI \sin 2\pi ft \, \sin (2\pi ft - \phi) \quad \text{(watts)} \quad (6)$$

$$P = VI[\cos \phi - \cos(4\pi ft - \phi)] \quad \text{(watts)} \quad (7)$$

$$P = VI \, \cos \phi \quad \text{(watts)} \quad (8)$$

See ALTERNATING CURRENT; ALTERNATING-CURRENT CIRCUIT THEORY.

Power factor, apparent power, and reactive power. The average power P in Eq. (8) is also called the real or active power and is distinguished from the simple product VI by the factor cos ϕ. The product VI is called the apparent power (which is measured in volt-amperes), and the factor cos ϕ is called the power factor so that, by definition Eq. (9*a*) holds or, in words, Eq. (9*b*).

$$\cos \phi = \frac{P}{VI} \quad (9a)$$

$$\text{Power factor} = \frac{\text{average power}}{\text{apparent power}} \quad (9b)$$

One additional quantity is introduced by rewriting Eq. (7) in the alternative form of Eq. (10), which indicates that the single double-frequency sinusoidal

$$p = VI[\cos \phi \, (1 - \cos 4\pi ft) + \sin \phi \sin 4\pi ft] \quad \text{(watts)} \quad (10)$$

instantaneous power wave of a circuit containing both resistance and inductance can be broken into two double-frequency sinusoidal waves. One of the waves is entirely above the zero line, and the average of this wave is the average power given by Eq. (8). The other wave is symmetric above and below the zero line and represents the instantaneous power flow into and out of the magnetic field of the inductor. This portion of the instantaneous power is termed the reactive power. The peak amplitude is given by Eq. (11), and is measured in volt-amperes-

$$Q = VI \, \sin \phi \quad \text{(vars)} \quad (11)$$

reactive or vars. The quantity Q is nominally taken as positive for an inductive load. It is evident from Eqs. (8) and (11) that the apparent power, the average power, and the reactive power are related by Eq. (12). *See* VOLT-AMPERE.

$$V^2I^2 = P^2 + Q^2 \quad \text{(volt-amperes)} \quad (12)$$

Nonsinusoidal periodic waves. For more complex periodic waves, the voltage and current can be expressed as Fourier series summations consisting of a dc component V_0, a fundamental component V_1, and a set of harmonics V_n as in Eq. (13), with the same

$$v = V_0 + \sum_{n=1}^{\infty} \sqrt{2}\, V_n \sin (2\pi fnt + \phi) \quad \text{(volts)} \quad (13)$$

general form for the current. It is easily shown by the methods used previously that only terms with the same frequency contribute to the power. All other products contribute only reactive power. The average power can be found by repeated application of Eq. (8) for each frequency present in both v and i. *See* FOURIER SERIES AND TRANSFORMS; HARMONIC (PERIODIC PHENOMENA); NONSINUSOIDAL WAVEFORM.

Low-frequency measurements. The measurement of power in a dc circuit can be carried out by simultaneous measurements of voltage and current by using standard types of dc voltmeters and ammeters. The product of the readings typically gives a sufficiently accurate measure of dc power. If great accuracy is required, corrections for the power used by the instruments should be made. In ac circuits the phase difference between the voltage and current precludes use of the voltmeter-ammeter method unless the load is known to be purely resistive. When this method is applicable, the instrument readings lead directly to average power since ac voltmeters and ammeters are always calibrated in rms values. *See* AMMETER.

Wattmeter measurements. In power-frequency circuits the most common instrument for power measurement is the moving-coil, dynamometer wattmeter. This instrument can measure dc or ac power by carrying out the required multiplication and averaging on a continuous analog basis. The instrument has four terminals, two for current and two for voltage, and reads the average power directly. It can be built with an accuracy better than 0.25% of full-scale deflection and with frequency response up to about 1 kHz. For harmonics within its frequency range, the reading is independent of the voltage and current waveforms. *See* WATTMETER.

Digital wattmeters. Electronic wattmeters are available which give a digital indication of average power. Their primary advantage, in addition to minimizing errors in reading the instrument, is that the frequency range can be greatly extended, up to 100 kHz or more, with good accuracy. In addition, such instruments also often provide other measurements such as rms voltage and current, apparent power, reactive power, and power factor. These instruments are of great value in power measurements involving very distorted waveforms as are common at the input of many electronic power supplies. Accurate measurement of power in these situations can be accomplished only with the wide bandwidth of these electronic instruments.

Polyphase power measurements. Measurement of the total power in a polyphase system is accomplished by combinations of single-phase wattmeters or by special polyphase wattmeters which are integrated combinations of single-phase wattmeter elements. A general theorem called Blondel's theorem asserts that the total power supplied to a load over N wires can be measured by using $N - 1$ wattmeters. The theorem states that the total power in an N-wire system can be measured by taking the sum of the readings of N wattmeters so arranged that each wire contains the current coil of one wattmeter. One voltage terminal of each wattmeter is connected to the same wire as its current coil, and the second voltage terminal is connected to a common point in the circuit. If this common point is one of the N wires, one wattmeter will read zero and can be omitted.

The most common polyphase system is the three-phase system, and the two most common connections are the three-phase, three-wire system, in which the sources and loads may be Y-connected or delta (Δ)-connected, and the three-phase, four-wire system, which utilizes Y-connected sources and loads. The fourth wire is the neutral wire, which connects the neutral point of the source to the neutral point of the load.

In one arrangement for measuring the power in a four-wire system, each three-phase wattmeter directly measures the power in one of the load elements. The sum of the three readings is the total power. However, this connection will read the correct total power even if the load is more complex, including, for example, one or more Δ-connected loads in parallel with the load illustrated in the figure. With multiple loads, only the sum of the meter readings is significant. For three-wire systems, the most common connection is the two-wattmeter arrangement. With this arrangement, again only the sum of the two readings is of significance. Care must be exercised in making the connection (the positive and negative terminals must be connected as shown) since one of the wattmeters will have a negative reading when the power factor is low or the load unbalanced. When this occurs, the negative-reading wattmeter must have its current or voltage coil reversed to obtain the value of the negative meter reading. Thus, careful attention to the positive and negative terminal connections is a necessity in this method.

High-frequency measurements. At frequencies significantly above power frequencies, dynamometer wattmeters become inaccurate and cannot be used. The newer digital wattmeters have usable ranges well above audio frequencies and make accurate audio-frequency power measurements quite feasible. Generally, however, power measurements at higher frequencies are based on indirect methods which do not attempt to carry out the multiplication and integration required for direct electric power measurement.

Voltmeter-based methods. For frequencies up to a few hundred megahertz, the voltage across a standard resistance load can be measured and the power calculated from V^2/R. Instruments which combine the resistive load and voltmeter are called absorption power meters. For the lower frequency end of the range, such instruments often contain an input transformer with taps to provide a range of input resistance values. Adjusting the transformer to achieve a maximum meter reading allows matching the effective load resistance to the source output impedance to maximize the power output from the source.

Diode detector–based methods. A diode may be used as a detector for radio-frequency (rf) power measurement. This type of instrument is simple and easy to use but is less accurate than the thermally based systems described below. The diode is operated on the square-law portion of its characteristic curve, and as a result the voltage on the capacitor C is proportional to the average rf power input. There are a number of limitations to the performance of these instruments. The dc signal at the output is very small, requiring amplification to enable accurate measurement, and care must be taken to avoid errors caused by electronic and thermoelectric effects. Temperature changes affect the diode resistance, and hence the sensitivity and reflection coefficient can vary; the matching resistor, usually 50 ohms, is used to terminate the input source to reduce these temperature effects. *See* DIODE.

Thermocouple-based methods. A thermocouple consists of two dissimilar metals joined at one end. When the joined end is heated and the other end is at a lower temperature, an electric current is produced (the

thermoelectric or Seebeck effect). The current is proportional to the temperature difference between the two ends. For electric power measurements, the hot end is heated by a resistor supplied from the rf power source to be measured. Modern devices use thin-film techniques to make the thermocouple and resistor to ensure good thermal coupling. The output is a low-level dc signal as in diode detector systems. Since the thermocouple method is affected by drift in the thermocouple and associated electronic circuits, most instruments contain a source arranged to be connected in place of the rf input to calibrate the instrument. Thermocouple instruments generally have a wider range of measurement and are more rugged than most other high-frequency power measurement instruments. *See* THERMOCOUPLE.

Bolometric methods. A bolometer is basically an electric bridge circuit in which one of the bridge arms contains a temperature-sensitive resistor. In principle, the temperature is detected by the bridge circuit. Bolometric bridges use either barreters or thermistors as the temperature-sensitive resistor. A barreter is simply a short, thin wire, usually of platinum, whose resistance changes significantly for small amounts of power dissipation. Since the element must be operated at relatively high temperatures, near its burn-out level, it is easily destroyed by accidental overloads. The thermistor, which has largely replaced the barreter since it is more rugged, is a semiconductor device with a negative temperature coefficient. For rf power measurements, the thermistor is fabricated as a small bead with very short lead wires so that essentially all the resistance is in the bead. *See* BARRETTER; BOLOMETER; THERMISTOR.

In power measurements, absolute measurement of the resistance variation is not generally acceptable since the change in resistance can affect the reflection coefficient of the probe. Thus, to expand the useful power range of the instrument, comparison methods are employed. For example, the bridge can be arranged so that, without any input power, a dc or low-frequency bias is applied to the sensing element and the bridge is balanced. The input power is then applied causing an imbalance. The bias is then changed to return the bridge to balance, and the change in the bias power is used to indicate the input power. A major advantage of thermistor-based bolometric techniques is that they operate at high signal levels and shielding requirements are minimized.

Calorimeter methods. The most accurate high-frequency power measurement methods involve calorimetric techniques based on direct determination of the heat produced by the input power. The power to be measured is applied to the calorimeter, and the final equilibrium temperature rise is recorded. Then the input signal is removed, and dc power applied until the same equilibrium temperature is attained. The dc power is then the same as the signal power. Most power measurement calorimeters use a flow-through coolant rather than a static fluid coolant, and can be built to measure large amounts of power. Calorimeter methods are usually used in standards laboratories rather than in industrial applications. *See* CALORIMETRY.

Sensors and connections. Ideally the power sensor should accept all of the high-frequency power to be measured. There are, however a number of factors that reduce the calibration factor of the sensor below the ideal of 100%. For example, mismatch between the impedance of the rf or microwave source and that of the sensor will result in some reflection of the incident power. Some power may also be lost in the conducting walls of the sensor, and some may be radiated into space. Typically, the calibration factor of a commercial power sensor at various frequencies within its range is supplied by the manufacturer.

The accuracy of high-frequency power measurements is very dependent on correct setup of the instrumentation to minimize losses. Most often, directional couplers and attenuators are used to supply the power meter with the desired signal at the appropriate power level for the instrument. Care must be exercised to assure that all portions of the test circuit are properly matched to avoid reflections and that losses are minimized. *See* DIRECTIONAL COUPLER; ELECTRICAL MEASUREMENTS; IMPEDANCE MATCHING; TRANSMISSION LINES.

Donald W. Novotny

Bibliography. M. Braccio, *Basic Electrical and Electronic Tests and Measurements*, 1978; *Code for Electrical Metering*, ANSI C12, 1982; D. Fink and H. W. Beaty (eds.), *Standard Handbook for Electrical Engineers*, 14th ed., 1999; Instrument Society of America, *ISA Standards and Practices for Instrumentation*, 9th ed., 1988; P. Kantrowitz, G. Kousouri, and L. Zucker, *Electronic Measurements*, 1979; F. F. Mazda, *Electronic Instruments and Measurement Techniques*, 1987.

Electric power substation

An assembly of equipment in an electric power system through which electrical energy is passed for transmission, distribution, interconnection, transformation, conversion, or switching. *See* ELECTRIC POWER SYSTEMS.

Specifically, substations are used for some or all of the following purposes: connection of generators, transmission or distribution lines, and loads to each other; transformation of power from one voltage level to another; interconnection of alternate sources of power; switching for alternate connections and isolation of failed or overloaded lines and equipment; controlling system voltage and power flow; reactive power compensation; suppression of overvoltage; and detection of faults, monitoring, recording of information, power measurements, and remote communications. Minor distribution or transmission equipment installation is not referred to as a substation.

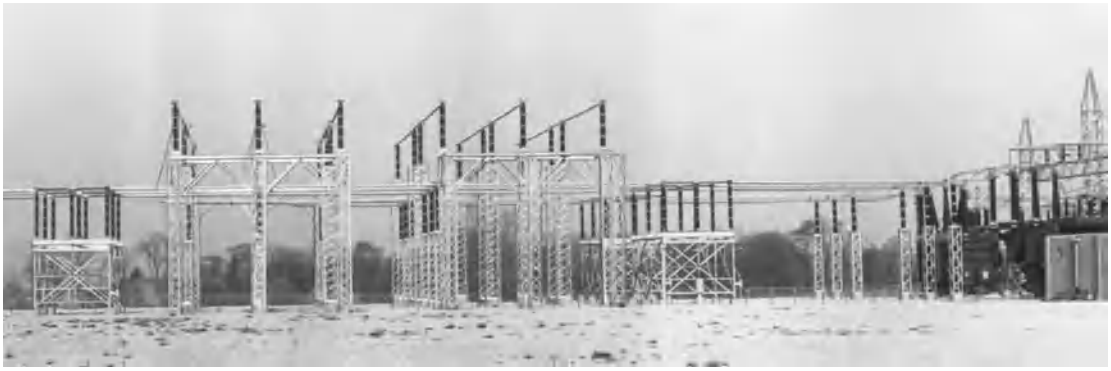


Fig. 1. A typical transmission substation (345 kV).

Classification. Substations are referred to by the main duty they perform. Broadly speaking, they are classified as transmission substations (**Fig. 1**), which are associated with high voltage levels; and distribution substations (**Fig. 2**), associated with low voltage levels. *See* ELECTRIC DISTRIBUTION SYSTEMS.

Substations are also referred to in a variety of other ways, which may be described as follows.

Transformer substations. These are substations whose equipment includes transformers.

Switching substations. These are substations whose equipment is mainly for various connections and interconnections, and does not include transformers.

Customer substations. These are usually distribution substations located on the premises of a larger customer, such as a shopping center, large office or commercial building, or industrial plant.

Converter stations. The main function of converter stations is the conversion of power from ac to dc and vice versa. Converter stations are complex substations required for high-voltage direct-current (HVDC) transmission or interconnection of two ac systems which, for a variety of reasons, cannot be connected by an ac connection. The main equipment includes converter valves usually located inside a large hall, transformers, filters, reactors, and capacitors. **Figure 3** shows a ± 250-kV 500-MW HVDC converter station together with a 245-kV ac substation. The building in the center of the converter station contains all the thyristor valves. On each side of the building against the wall are converter transformers connected to the valves inside. Further away from the building are filters and capacitors. Going away from the building is the dc line, with a typical dc line tower in the uppermost part of the figure. *See* DIRECT-CURRENT TRANSMISSION.

Air-insulated substations. Most substations are installed as air-insulated substations, implying that the busbars and equipment terminations are generally open to the air, and utilize insulation properties of ambient air for insulation to ground. Modern substations in urban areas are esthetically designed with low profiles and often within walls, or even indoors.

Metal-clad substations. These are also air-insulated, but for low voltage levels; they are housed in metal cabinets and may be indoors or outdoors (Fig. 2).

Fig. 2. View of a residential outdoor distribution substation with double-ended underground supply to transformers which are positioned at either end of the metal-clad switchgear.

Fig. 3. View of ±250-kV 500-MW HVDC converter station (upper part) and 245-kV ac substation (lower part). The large building of the converter station contains all the thyristor valves. (*General Electric Co.*)

Gas-insulated substations. Acquiring a substation site in an urban area is very difficult. Land in many cases is either unavailable or very expensive. Therefore, there has been a definite trend toward increasing use of gas-insulated substations, which occupy only 5–20% of the space occupied by the air-insulated substations. In gas-insulated substations, all live equipment and bus-bars are housed in grounded metal enclosures, which are sealed and filled with sulfur hexafluoride (SF_6) gas, which has excellent insulation properties. These substations have the appearance of a large-scaleplumbing (**Fig. 4**).

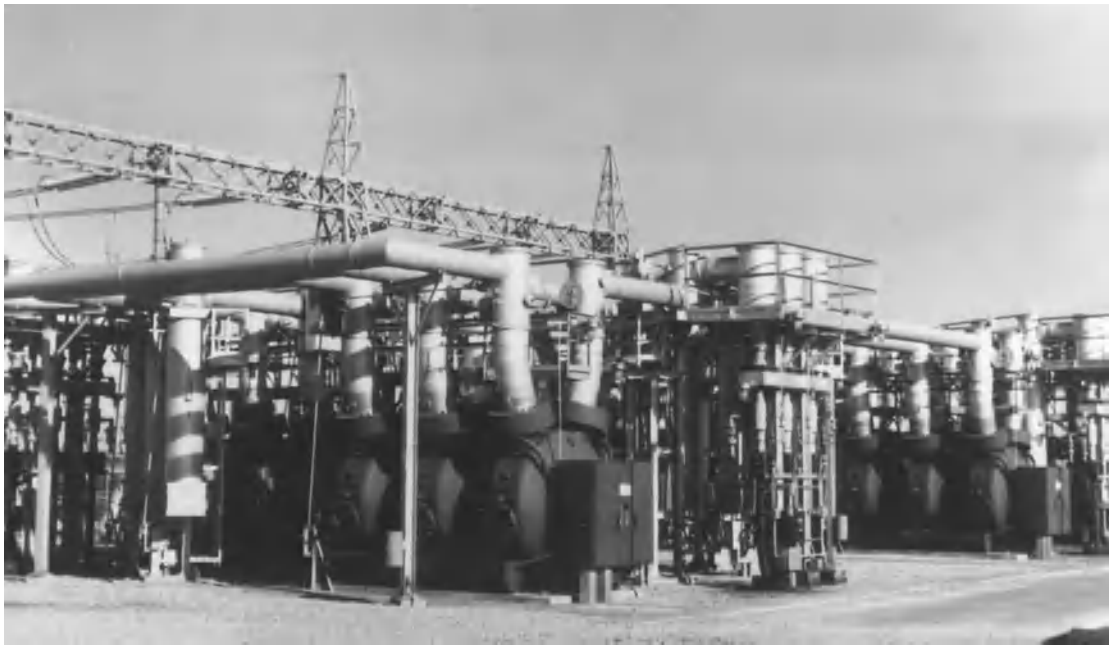


Fig. 4. A 345-kV ac gas-insulated substation. (*Gould-BBC*).

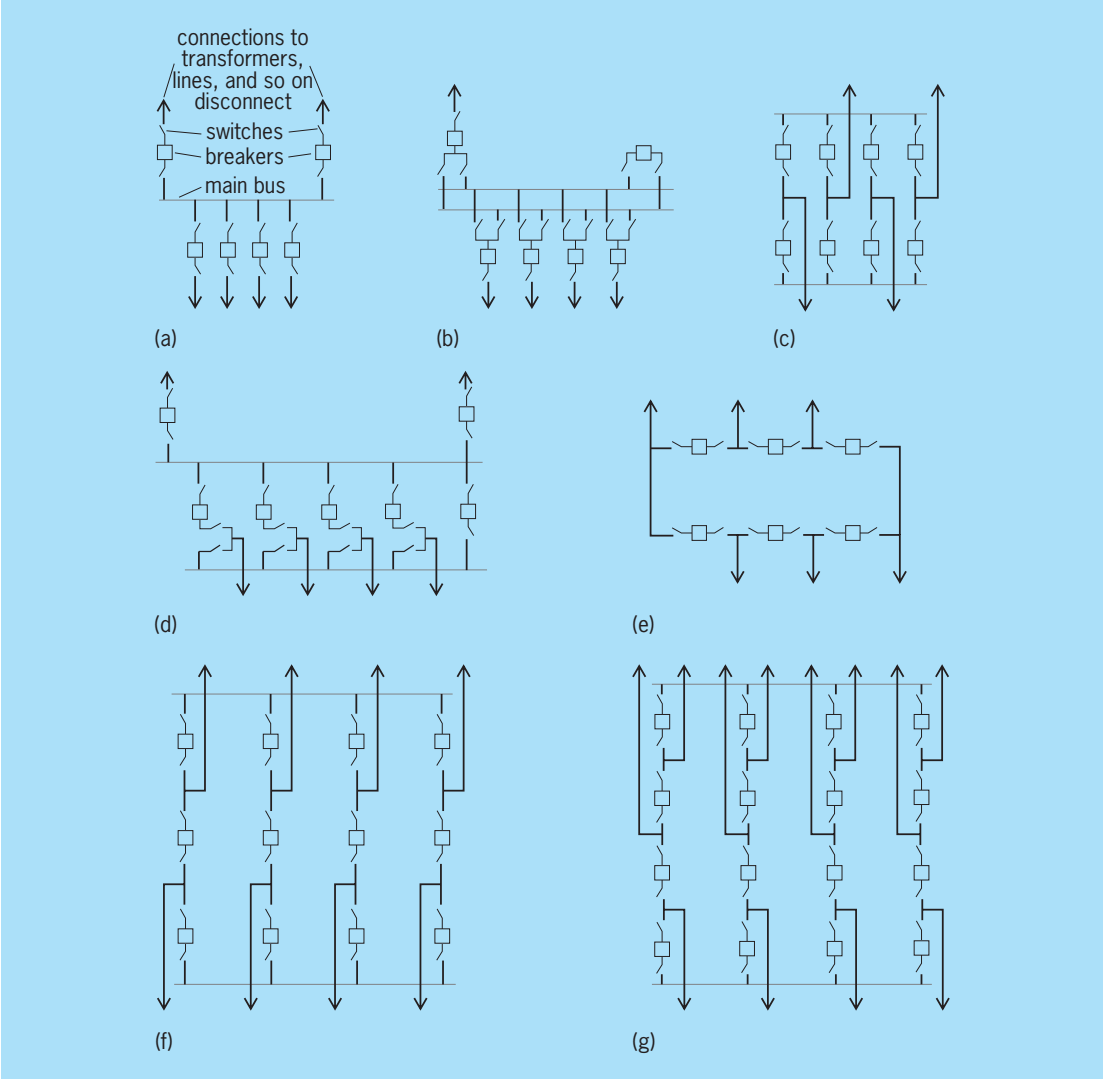


Fig. 5. One-line diagrams of substation switching arrangements. (*a*) Single bus. (*b*) Double bus, single breaker. (*c*) Double bus, double breaker. (*d*) Main and transfer bus. (*e*) Ring bus. (*f*) Breaker-and-a-half. (*g*) Breaker-and-a-third.

The disk-shaped enclosures in Fig. 4 are circuit breakers.

Mobile substations. For emergency replacement or maintenance of substation transformers, mobile substations are used by some utilities. They may range in size from 100 to 25,000 kVA, and from 2400 to 220,000 V. Most units are designed for travel on public roads, but larger substations are built for travel on railroad tracks. These substations are generally complete with a transformer, circuit breaker, disconnect switches, lightning arresters, and protective relaying.

Substation switching arrangement. An appropriate switching arrangement for "connections" of generators, transformers, lines, and other major equipment is basic to any substation design. There are seven switching arrangements commonly used (**Fig. 5**). Each breaker is usually accompanied by two disconnect switches, one on each side, for maintenance purposes. Selecting the switching arrangement involves considerations of cost, reliability, maintenance, and flexibility for expansion.

Single bus. This involves one common bus for all connections and one breaker per connection. This is the least costly arrangement, but also the least desirable for considerations of reliability and maintenance.

Double bus, single breaker. This involves two common buses, and each connection of the line equipment and so forth can be made to either bus through one or the other breaker. However, there is only one breaker per connection, and when a breaker is out for maintenance, the connected equipment or line is also removed from operation.

Double bus, double breaker. This arrangement has two common buses and two breakers per connection. It offers the most reliability, ease of maintenance, and flexibility, but is also the most expensive.

Main and transfer bus. This is like the single-bus arrangement; however, an additional transfer bus is provided so that a breaker can be taken out of service by transferring its connection to the transfer bus which is connected to the main bus through a breaker between the two buses.

Ring bus. This may consist of four, six, or more breakers connected in a closed loop, with the same number of connection points. A breaker can be taken out of service without also taking out a connection.

Breaker-and-a-half. This arrangement involves two buses, between which three breaker bays are installed. Each three-breaker bay provides two circuit connection points—thus the name breaker-and-a-half.

Breaker-and-a-third. In this arrangement, there are four breakers and three connections per bay.

Substation equipment. A substation includes a variety of equipment. The principal items are listed and briefly described below.

Transformers. These involve magnetic core and windings to transfer power from one side to the other at different voltages. Substation transformers range from small sizes of 1 MVA to large sizes of 2000 MVA. Most of the transformers and all those above a few MVA size are insulated and cooled by oil, and adequate precautions have to be taken for fire hazard. These precautions include adequate distances from other equipment, fire walls, fire extinguishing means, and pits and drains for containing leaked oil. *See* TRANSFORMER.

Circuit breakers. These are required for circuit interruption with the capability of interrupting the highest fault currents, usually 20–50 times the normal current, and withstanding high voltage surges that appear after interruption. Switches with only normal load-interrupting capability are referred to as load break switches. *See* CIRCUIT BREAKER.

Disconnect switches. These have isolation and connection capability without current interruption capability. *See* ELECTRIC SWITCH.

Bus-bars. These are connecting bars or conductors between equipment. Flexible conductor buses are stretched from insulator to insulator, whereas more common solid buses (hollow aluminum alloy tubes) are installed on insulators in air or in gas-enclosed cylindrical pipes. *See* BUS-BAR.

Shunt reactors. These are often required for compensation of the line capacitance where long lines are involved.

Shunt capacitors. These are required for compensation of inductive components of the load current.

Current and potential transformers. These are for measuring currents and voltages and providing proportionately low-level currents and voltages at ground potential for control and protection.

Control and protection. This includes (a) a variety of protective relays which can rapidly detect faults anywhere in the substation equipment and lines, determine what part of the system is faulty, and give appropriate commands for opening of circuit breakers; (b) control equipment for voltage and current control and proper selection of the system configuration; (c) fault-recording equipment; (d) metering equipment; (e) communication equipment; and (f) auxiliary power supplies. *See* ELECTRIC PROTECTIVE DEVICES; RELAY; VOLTAGE REGULATOR.

Many of the control and protection devices are solid-state electronic types, and utilize digital techniques with microprocessors. Most of the substations are fully automated locally with a provision for manual override. The minimum manual interface required, along with essential information on status, is transferred via communications channels to the dispatcher in the central office.

Other items which may be installed in a substation include: phase shifters, current-limiting reactors, dynamic brakes, wave traps, series capacitors, controlled reactive compensation, fuses, ac to dc or dc to ac converters, filters, and cooling facilities.

Substation grounding and shielding. Good substation grounding is very important for effective relaying and insulation of equipment; however, the safety of the personnel is the governing criterion in the design of substation grounding. It usually consists of a bare wire grid, laid in the ground; and all equipment grounding points, tanks, support structures, fences, shielding wires and poles, and so forth, are securely connected to it. The grounding resistance is reduced to be low enough that a fault from high voltage to ground does not create such high potential gradients on the ground, and from the structures to ground, to present a safety hazard. Good overhead shielding is also essential for outdoor substations, so as to virtually eliminate the possibility of lightning directly striking the equipment. Shielding is provided by overhead ground wires stretched across the substation or tall grounded poles. *See* GROUNDING; LIGHTNING AND SURGE PROTECTION.

Narain G. Hingorani

Bibliography. J. Arrilaga, *High Voltage Direct Current Transmission*, 2d ed., 1998; W. M. Flanagan, *Handbook of Transformer Design and Applications*, 2d ed., 1993; C. H. Flurscheim (ed.), *Power Circuit Breaker Theory and Design*, rev. ed., 1982; A. C. Franklin and D. P. Franklin, *The J&P Transformer Book,* 11th ed., 1983.

Electric power systems

Complex assemblages of equipment and circuits for generating, transmitting, and distributing electrical energy. The elements of a power system (**Fig. 1**) form complex networks which require energy control centers to monitor and regulate their operation. Various sources of primary energy, such as coal, nuclear fission, hydro power, geothermal, and wind, can be used to drive the electric generator in Fig. 1.

Since electrical energy plays a central role in industrialized societies, the reliability of electric power systems is a critical factor in their planning, design, and operation, and considerable effort is directed toward quantifying performance. In addition to the provision of electrical energy, attention is given to the myriad of items that consume this energy, that is, the power system load.

Principal Elements

Electrical systems require generating stations to produce electric power, transmission systems to carry

it to areas of consumption, substations to transform it for use in industrial and residential areas, and distribution systems to carry it to customers (Fig. 1). Coordinated interconnections between power systems play a critical role in ensuring an acceptable degree of reliability.

Generation. Electricity in the large quantities required to supply electric power systems is produced in generating stations, commonly called power plants. Such generating stations, however, should be considered as conversion facilities in which the chemical or nuclear energy of fuel (coal, oil, gas, or uranium); the kinetic energy of falling water, wind, or sea waves; or light from the Sun is converted to electricity. *See* ELECTRIC POWER GENERATION; POWER PLANT.

Steam stations. Most of the electric power used in the United States is obtained from generators driven by steam turbines. Units of 650, 800, and 950 MW are commonplace for new fossil-fuel-fired stations, with 1300 MW the largest in service. The most commonly installed units in nuclear stations are 800–1100 MW. *See* STEAM TURBINE.

Coal is an important fuel for more than half of steam turbine generation. Other fuels include natural gas and heavy fuel oil; the remainder is generated from the nuclear energy of slightly enriched uranium. As nuclear units have come into commercial operation, the contribution of uranium to the electrical energy supply has risen to about 15% of the total fuel generated. However, uranium's share should decrease, because no additional nuclear plants are slated for service and instead some existing nuclear power stations are slated for closure.

Combustion of coal produces sulfur dioxide and nitrogen oxides in the stack gases. To reduce these emissions below those permitted by the Environmental Protection Agency, modern coal-fired stations use either low-sulfur coal (that is, less than 0.3% sulfur by weight) or scrubbers that react the sulfur dioxide with a reagent to permit its removal. Nitrogen oxides are controlled by controlling combustion temperatures. Research is being conducted on clean coal technologies to minimize the production of carbon dioxide. *See* AIR POLLUTION.

Nuclear steam stations in the United States are mostly of the water-cooled and moderated types in which the heat of a controlled nuclear reaction is used to convert ordinary or "light" water into steam to drive a conventional turbine generator. *See* NUCLEAR POWER; NUCLEAR REACTOR.

Hydroelectric plants. Waterpower supplies about 10% of the electric power consumed in the United States, but this share will decline because very few sites remain undeveloped where sufficient water drops far enough in a reasonable distance to drive large hydraulic turbines. Much of the very small share of the planned additional hydroelectric capability will be used at existing plants to increase their effectiveness in supplying peak power demands, and as a quickly available source of emergency power. Special incentives from the federal government have made development of small hydroelectric plants of 15 MW or less financially attractive to private developers. *See* HYDRAULIC TURBINE; HYDROELECTRIC GENERATOR.

Some hydroelectric plants actually draw power from other generating facilities during light system-load periods, to drive their turbines in a reverse mode to pump water from a river or lake into an artificial reservoir at a higher elevation. From there, the water can be released through the hydraulic turbines when the power system needs additional generation. These pumped-storage installations consume about 35% more energy than they return to the power system and, accordingly, cannot be considered as primary energy sources. Their use is justified, however, by their ability to convert surplus power that is available during low-demand periods into prime power to serve system needs during peak-demand intervals, a need that otherwise would require building more generating stations for operation during the relatively few hours of high system demand. *See* ENERGY STORAGE; WATERPOWER.

Combustion turbine plants. Gas-turbine–driven generators, now commonly called combustion turbines because of the use of light oil as fuel, have gained wide acceptance as an economical source of additional power for heavy-load periods. In addition, they offer the fastest erection time and the lowest investment cost per kilowatt of installed capability. Offsetting these advantages, however, is their relatively less efficient consumption of more costly fuel. Combustion turbine units, even in the largest rating (100 MW), offer extremely flexible operation and can be started and run up to full load in as little as 10 minutes. Thus they are useful as emergency power sources, as well as for operating during the few hours of daily load peaks.

Combustion turbines have an additional role besides applying peaking or emergency power. Some installations use their exhaust gases to heat boilers that generate steam to drive steam turbine generators. Such combined-cycle units offer fuel economy comparable to that of modern steam plants and at considerably less cost per kilowatt. In addition, because only part of the plant uses steam, the requirement for cooling water is considerably reduced. However, wide acceptance is inhibited by the government's restrictions on light fuel oil for them. This barrier will be resolved by the development of systems for fueling combustion-turbine installations with gas derived from coal, natural gas, or by direct firing with pulverized coal. *See* COAL GASIFICATION; COGENERATION; GAS TURBINE.

Internal combustion plants. Internal combustion engines of the diesel type drive generators in many small power plants. In addition, they offer the ability to start quickly for operation during peak loads or emergencies. However, their small size, commonly about 2 MW per unit (although a few approach 10 MW), has limited their use. Such installations account for less than 1% of the total power-system generating capability in the United States, and make an

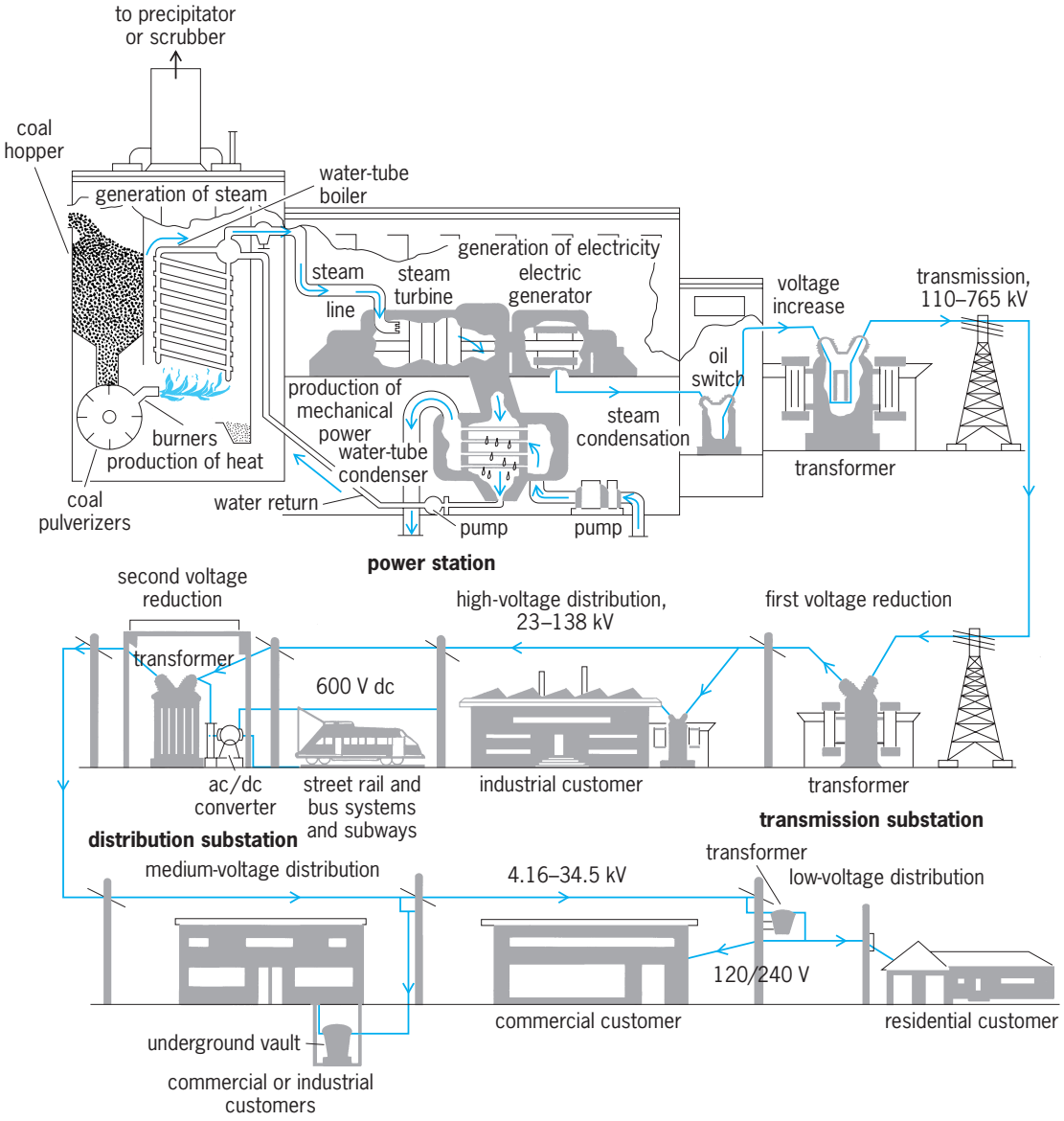


Fig. 1. Major steps in the thermal generation, transmission, and distribution of electricity.

even smaller contribution to total electric energy consumed. At present they are used primarily as standby units to provide electricity in case of emergencies and loss of electric grid. *See* DIESEL ENGINE; INTERNAL COMBUSTION ENGINE.

Renewable resources. Utilities actively seek to develop generating resources that do not consume fuel, including geothermal steam, wind, Sun, ocean waves, tides, and biomass-powered capacity. Of these, wind power is now the fastest growing segment of the power generation industry. Wind energy technology worldwide has grown 20% annually since 2000, and it is projected that by 2020 wind energy will meet 6% of the electric needs of the United States, matching the share of hydro power. *See* BIOMASS; ENERGY SOURCES; GEOTHERMAL POWER; SOLAR ENERGY; WIND POWER.

Three-phase output. Because of their simplicity and efficient use of conductors, three-phase 60-Hz alternating-current systems are used almost exclusively in the United States. Consequently, power-system generators are wound for three-phase output at a voltage usually limited by design features to a range from about 11 kV for small units to 30 kV for large ones. The output of modern generating stations is usually stepped up by transformers to the voltage level of transmission circuits used to deliver power to distant load areas. *See* TRANSFORMER.

Transmission. The transmission system carries electric power efficiently and in large amounts from generating stations to consumption areas. Such transmission is also used to interconnect adjacent power systems for mutual assistance in case of emergency and to gain for the interconnected power systems the economies possible in regional operation. Interconnections have so expanded that most of the generation east of the Rocky Mountains regularly operates in parallel and in synchronism. More than 90% of all

Power capability of typical three-phase open-wire transmission lines

Line-to-line voltage, kV	Capability, MVA
115 ac	60
138 ac	90
230 ac	250
345 ac	600
500 ac	1200
765 ac	2500
800 dc*	1500

*Bipolar line with grounded neutral.

generation in the United States and Canada, exclusive of Alaska and Hawaii, is or can be linked.

Transmission circuits are designed to operate up to 765 kV, depending on the amount of power to be carried and the distance to be traveled. The permissible power loading of a circuit depends on many factors, such as the thermal limit of the conductors and their clearances to ground, the voltage drop between the sending and receiving end and the degree to which system service reliability depends on it, and how much the circuit is needed to hold various generating stations in synchronism. A widely accepted approximation to the voltage appropriate for a transmission circuit is that the permissible load-carrying ability varies as the square of the voltage (see **table**).

Alternating-current transmission voltages have increased many times since transmission began as a distinct function around 1890 (**Fig. 2**). Lines at 500 and 765 kV were introduced in the 1960s. The very high capacity of lines at 765 kV limits their use, but 2454 mi (3950 km) are in service in the United States.

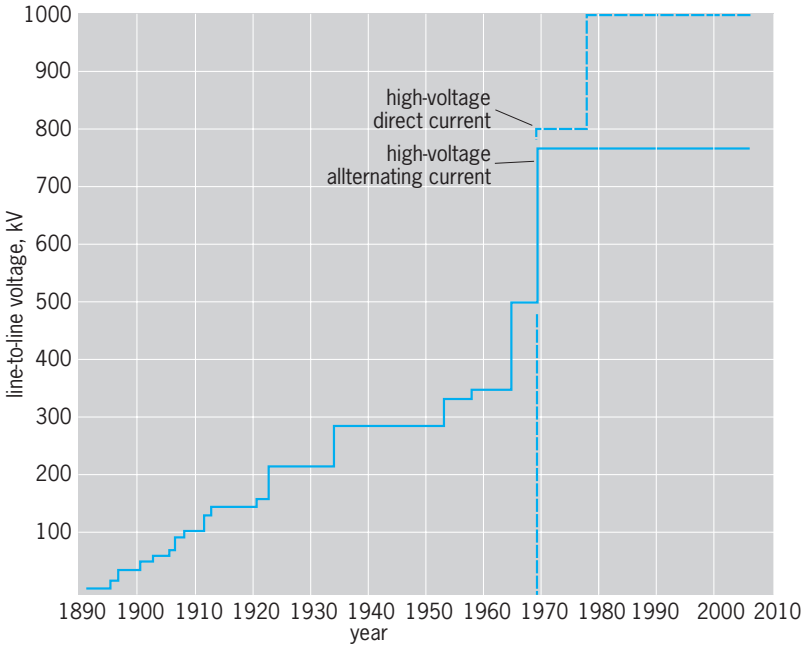


Fig. 2. Growth of alternating-current transmission voltages from 1890 onwards.

Transmission engineers have anticipated even higher voltages, of 1100 to 1500 kV, but they are fully aware that this objective may prove too costly in space requirements and funds to gain wide acceptance. Experience gained at 500 and 765 kV verifies that the prime requirement no longer is insulating the lines to withstand lightning discharges, but insulating them to tolerate voltage surges caused by the operation of circuit breakers. Audible noise levels, especially in rain or humid conditions, are high, requiring wide buffer zones. Environmental challenges have been brought on the basis of possible negative biological effects of the electrostatic field produced under EHV lines, although research has not shown any such effects. *See* ELECTRIC PROTECTIVE DEVICES; LIGHTNING AND SURGE PROTECTION.

Experience has indicated that, within about 10 years after the introduction of a new voltage level for overhead lines, it becomes necessary to begin connecting underground cable. The first 500-kV cable in the United States was placed in service in 1976.

Another approach to high-voltage long-distance transmission is high-voltage direct current (HVDC), which offers the advantages of less costly lines, lower transmission losses, and insensitivity to many system problems that restrict alternating-current systems. Its greatest disadvantage is the need for costly equipment for converting the sending-end power to direct current, and for converting the receiving-end direct-current power to alternating current for distribution to consumers. However, the development of solid-state converter valves made overhead HVDC economical for distances more than about 400 mi (650 km), but only 25–30 mi (40–50 km) in underground construction. Where systems that are out of synchronism must be interconnected, much shorter distances may be economic. An extreme example of this is at Eel River, New Brunswick, Canada, where back-to-back converters connect Canadian and United States systems. Starting in the late 1950s with a 65-mi (105-km) 100-kV system in Sweden, HVDC has been applied successfully in a series of special cases around the world, each one for a higher voltage and greater power capability. The first such installation in the United States was put into service in 1970. It operates at 800 kV line to line, and is designed to carry a power interchange of 1440 MW over an 841-mi (1354-km) overhead tie line between the winter-peaking Northwest Pacific coastal region and the summer-peaking southern California area. These HVDC lines perform functions other than power transfer, however. The Pacific Intertie is used to stabilize the parallel alternating-current transmission lines, permitting an increase in their capability; and back-to-back converters with no tie line between them are used to tie together two systems in Nebraska that otherwise could not be synchronized. The first urban installation of this technology was energized in 1979 in New York. *See* DIRECT-CURRENT TRANSMISSION.

In addition to these high-capability circuits, every large utility has many miles of lower-voltage transmission, usually operating at 110 to 345 kV, to carry bulk power to numerous cities, towns, and large industrial plants. These circuits often include extensive lengths of underground cable where they pass through densely populated areas. *See* ELECTRIC POWER TRANSMISSION; TRANSMISSION LINES.

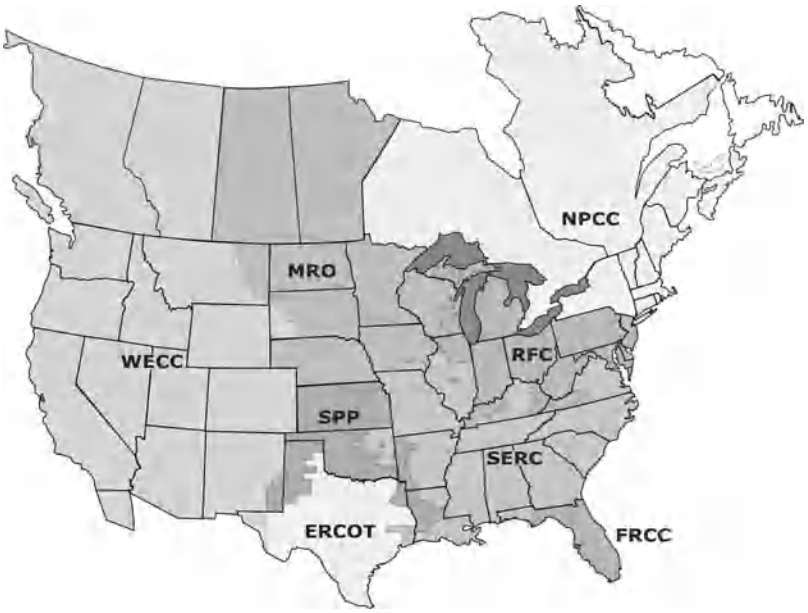
Interconnections. As systems grow and the number and size of generating units increase, and as transmission networks expand, higher levels of bulk-power-system reliability are attained through properly coordinated interconnections among separate systems. This practice began around 1930.

Most of the electrical utilities in the contiguous United States and a large part of Canada operate as members of power pools, and these pools in turn are interconnected into one gigantic power grid, known as the North American Power Systems Interconnection. The operation of this interconnection, in turn, is coordinated by the operating committee of the North American Electric Reliability Council (NERC). Each individual utility in such pools operates independently, but has contractual arrangements with other members in respect to generation additions and scheduling of operation. Their participation in a power pool affords a higher level of service reliability and important economic advantages.

The Northeast blackout of November 9, 1965, stemmed from the unexpected trip-out of a key transmission circuit carrying emergency power into Canada and cascaded throughout the northeastern states to cut off electrical service to some 30,000,000 people. It spurred the utilities into a chain reaction affecting the planning, construction, operation, and control procedures for their interconnected systems. They soon organized regional coordination councils to cover the entire contiguous United States and four Canadian provinces. Their objective was to further improve reliability of the planning and operation of their generation and transmission facilities. Currently, there are eight regional reliability councils (**Fig. 3**).

Then, in 1968, the North American Electric Reliability Council (NERC) was established to serve as an effective body for the collection, unification, and dissemination of various reliability criteria for use by individual utilities in meeting their planning and operating responsibilities (Fig. 3).

Increased interconnection capability among power systems reduces the required generation reserve of each of the individual systems. In most utilities the loss-of-load probability (LOLP) is used to measure the reliability of electrical service, and it is based on the application of probability theory to unit-outage statistics and load forecasts. A common LOLP criterion is 1 day in 10 years when load may exceed generating capability. The LOLP decreases (that is, reliability increases) with increased interconnection between two areas until a level is reached which depends upon the amount of reserve, unit sizes, and annual load shape in each area.



Regional Reliability Councils

ERCOT, Electric Reliability Council of Texas
FRCC, Florida Reliability Coordinating Council
MRO, Midwest Reliability Organization
NPCC, Northeast Power Coordinating Council
RFC, Reliability First Corporation
SERC, SERC Reliability Corporation
SPP, Southwest Power Pool Inc.
WECC, Western Electricity Coordinating Council

Fig. 3. Areas served by the eight regional reliability councils, coordinated by the North American Electric Reliability Council, that guide the planning, coordination, and operation of generation and transmission facilities. (*North American Electric Reliability Council, http://www.nerc.com*)

Traditionally, systems were planned to withstand all reasonably probable contingencies, and operators seldom had to worry about the possible effect of unscheduled outages. Reserve margin, which is the excess of capacity over load at maximum annual peak, while difficult for every system, is generally considered acceptable at 25%. Operators' normal security functions were to maintain adequate generation online and to ensure that such system variables as line flows and station voltages remained within the limits specified by planners. However, stronger interconnections, larger generating units, and rapid system growth spread the transient effects of sudden disturbances and increased the responsibilities of operators for system security.

System security is concerned with service continuity at standard frequency and voltage levels. The system is said to be insecure if a contingency would result in overloading some system components, in abnormal voltage levels at some stations, in change of system frequency, or in system instability, even if there is adequate capability as indicated by some reliability index. The concepts of power system

reliability and security are discussed in more detail below.

Substations. Power delivered by transmission circuits must be stepped down in facilities called substations to voltages more suitable for use in industrial and residential areas. On transmission systems, these facilities are often called bulk-power substations; at or near factories or mines, they are termed industrial substations; and where they supply residential and commercial areas, distribution substations.

Basic equipment in a substation includes circuit breakers, switches, transformers, lightning arresters and other protective devices, instrumentation, control devices, and other apparatus related to specific functions in the power system. *See* ELECTRIC POWER SUBSTATION.

Distribution. That part of the electric power system that takes power from a bulk-power substation to customers' switches, commonly about 35% of the total plant investment, is called distribution. This category includes distribution substations, subtransmission circuits that feed them, primary circuits that extend from distribution substations to every street and alley, distribution transformers, secondary lines, house service drops or loops, metering equipment, street and highway lighting, and a wide variety of associated devices.

Primary distribution circuits usually operate at 4160 to 34,500 V line to line, and supply large commercial institutional and some industrial customers directly. Since about 1950, by far the majority of lines constructed have been in the 15-kV class. Primary lines may be overhead open wire on poles, spacer or aerial cable, or underground cable. Legislation in more than a dozen states requires that all new services to developments of five or more residences be put underground. The bulk of existing lines are overhead, however, and will remain so for the forseeable future.

At conveniently located distribution transformers in residential and commercial areas, the line voltage is stepped down to supply low-voltage secondary lines, from which service drops extend to supply the customers' premises. Most such service is at 120/240 V, but other common voltages are 120/208 and 125/216 V, and for larger commercial and industrial buildings, 240/480, 265/460, or 277/480 V. These are classified as utilization voltages. *See* ELECTRIC DISTRIBUTION SYSTEMS.

Electrical utility industry. In the United States, which has the third highest per capita use of electricity in the world and more electric power capability than any other nation, electrical systems as measured by some criteria are the country's largest industry.

The utility industry in the United States is pluralistic in the nature of its ownership. Investor-owned utilities comprise about 78% of total installed capacity, and about the same percentage of customers served. Publicly owned utilities (that is, municipal, state, and power-district utilities) own and operate about 10% of the installed capacity and serve almost 14% of the customers; cooperatives own 3% of the capacity and serve about 10% of the customers; and federal agencies, while serving no measurable percentage of customers, own almost 10% of installed capacity. Electrical utilities are among the most capital-intensive of industries, primarily because of the huge investment in generating units.

Introduction of deregulation over the past few years has resulted in a significant reorganization of the electric utility industry. Instead of vertically integrated utilities being engaged in all three aspects—generation, transmission, and distribution—the industry is being organized horizontally with independent power producers (IPPs), independent system operators (ISOs), and independent distribution companies (DISCOs). The ISOs manage transmisson of power from the IPPs to the DISCOs to meet the contractual obligations of the IPPs and DISCOs. This structure is supplemented by electricity markets with bidding mechanisms for the sale and purchase of electricity, and long-term and short-term contracts between the IPPs and the DISCOs. In this reorganization, the ISO link is still regulated and has the responsibility of maintaining the reliability of electricity supply.

William C. Hayes; Om P. Malik

Power System Operation

The operation and control of the generation-transmission-distribution grid is quite complex because this large system has to operate in synchronism and because many different organizations are responsible for different portions of the grid. In North America and Europe, many public and private electric power companies are interconnected, often across national boundaries. Thus, many organizations have to coordinate to operate the grid, and this coordination can take many forms, from a loose agreement of operational principles to a strong pooling arrangement of operating together.

Power-system operations can be divided into three stages: operations planning, real-time control, and after-the-fact accounting. The main goal is to minimize operations cost while maintaining the reliability (security) of power delivery to customers. When all the utilities were vertically integrated—that is, the operator was also the owner of the generation and transmission—centrally optimizing the generation production was a main duty of the operator. Many regions of the world have deregulated generation ownership, and in those areas the market transactions determine generation production while the network operator is mainly concerned with the reliability of the grid. Operations planning requires the forecasting of load demand in the next few hours, days, weeks, or months that allows the market to determine the schedules and the operator to approve these schedules if they meet the reliability requirements. The market determines the buying and selling of power for the long and short terms, and after the markets clear (which involves matching the buy and sell bid prices) the

generating companies can schedule hydro resources, fossil fuels, and equipment maintenance over many weeks, and the commitment (startup and shutdown) of generating units over many hours. The operator has to run contingency (what-if) studies using these schedules to ensure that all reliability criteria are met or else the buy–sell transactions have to be adjusted to ensure that the reliability criteria are met.

Real-time control of the system is required to respond to the actual demand of electricity, which always deviates somewhat from the forecast, and any unforeseen contingencies (equipment outages). Maintaining reliability of the system so that a possible contingency cannot disrupt power supply is an integral part of the real-time control function.

After-the-fact accounting tracks actual delivery of energy between organizations so that billing can be generated.

For loosely coordinated operation of the grid, each balancing authority or utility takes responsibility for the operation of its own portion while exchanging all relevant information with its partners. For pool-type operations, a group of utilities sets up a pool where the operational decisions may be made centrally and then implemented by each utility. All of this requires significant data communication as well as engineering computation within a utility as well as between utilities. The use of modern computers and communications technology makes this possible, and the heart of system operations in a utility is the energy control center.

The monitoring and control of a power system from a centralized control center became desirable quite early in the development of electric power systems, when generating stations were connected together to supply the same loads. As electrical utilities interconnected and evolved into complex networks of generators, transmission lines, distribution feeders, and loads, the control center became the operations headquarters for each utility. Since the generation and delivery of electrical energy are controlled from this center, it is referred to as the energy control center or energy management system.

Although these centers have greatly advanced in technology, their basic control objective of maintaining reliability and security remains the same. In a vertically integrated utility the economic goal of minimizing the cost of supply is achieved by optimizing the dispatch and scheduling of its generators, whereas in deregulated regions the electricity markets optimize the supply costs. Increasingly, the generation is being privatized, and the generation companies have to compete in the electricity markets to sell their supply. The computerized tools needed to run the market and those used by the generation companies to determine their bid strategies are becoming more sophisticated. The rest of this section, however, describes the tools that are used to operate the grid itself in a reliable manner that avoids cascading failures and blackouts.

Evolution. In early control centers measurements of important power flows, voltages, and currents were displayed on banks of meters. Circuit-breaker and switch positions were sometimes monitored and displayed by the switching of lights, placed appropriately on a wall map of the power system. Hard-wired control switches were used to open or close breakers remotely from the control centers. These supervisory control and data acquisition (SCADA) functions became more versatile and flexible with the computerization of the control center.

Automatic generation control (AGC), which adjusts the generator outputs to follow the load changes, became absolutely necessary when utilities started to interconnect, and customized controllers were initially built for the purpose. The use of first analog and then digital computers greatly improved the capabilities of automatic generation control. For example, the problem of meeting the load with the cheapest combination of generating sources requires a multistep calculated solution, for which digital computers are ideal.

Digital computer–based SCADA-AGC control centers became common in the 1960s. They could handle more information, provide the information to the operator more efficiently, and provide monitoring and control of the power system more reliably. Also, the availability of unprecedented computational power made it possible to develop new functions for better control. Security analysis and scheduling functions based on on-line computer programs are now available to the control-center operator, providing more economic and secure control.

Configuration. The control-center computers are connected through communication channels, usually microwave links, although fiber optics are increasingly used, to remote terminal units. A remote terminal unit is placed at each substation and generating station, and it can gather all the measurements at that station and execute the control commands sent from the control center.

At the control center itself, many computers must be linked together to accomplish the various functions. For example, one workstation may handle the SCADA functions, another may be dedicated to security analysis, while several more may handle the graphics interface needed by the operator. Such distributed processing configurations (**Fig. 4**) are common, and open-system standards are used for computer operating systems and communications, making it much easier to expand and modify existing hardware configurations. *See* DISTRIBUTED SYSTEMS (COMPUTERS); DISTRIBUTED SYSTEMS (CONTROL SYSTEMS).

Redundancy has to be carefully designed into the control-center configuration so that the probability of losing the critical functions is very small. In addition to redundant computers, backup is usually provided for communication channels, remote terminal unit circuits, operator consoles, and other equipment. *See* MULTIPROCESSING.

Human-machine interface. The banks of control panels have been replaced by operator consoles which typically consist of color display screens, a keyboard with extra function buttons, logging

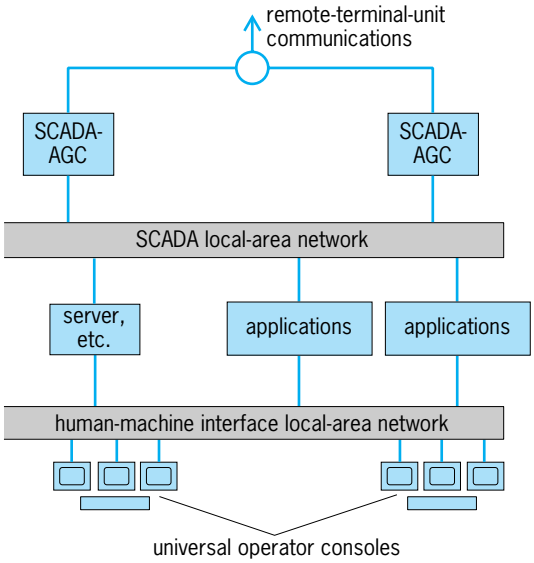


Fig. 4. Typical distributed-processing computer configuration for an energy control center.

printers, and telephones. All the measured and calculated data can be viewed at the display screens, and all control can be initiated and monitored from these consoles. A control center will usually have several consoles (**Fig. 5**), each dedicated to a particular aspect of system operation, like interchange monitoring or transmission control. Because the consoles are computer-controlled, their responsibilities can be changed at any time to accommodate changing procedures or availability of operators. The use of wall-mounted maps of the power system is quite common. Increasingly, these maps are very large liquid-crystal displays (LCDs) or plasma displays and are driven by the computers to denote breaker positions, line overloading, or out-of-limit voltages. *See* CATHODE-RAY TUBE; ELECTRONIC DISPLAY.

The control-center computers must handle thousands of displays, many of which are continually accessed and updated. This imposes a high demand on the computers, especially during emergencies, when the operator has to absorb and react to rapidly changing power-system conditions.

Control centers use display generator programs that allow the on-line creation of displays by the users, without computer-programming changes. Both the hardware and software techniques for graphics have improved dramatically over the years, and the quality of the human-machine interface is very high.

Supervisory control and data acquisition. All measurement data are scanned periodically by the computers, usually every 2 to 4 s. These data and others calculated from them are available to the operators through the displays. The data are checked for limit violations, and alarms are generated if necessary. Computerization and good database management programs have made it feasible for a control center to handle data from tens of thousands of information points. *See* DATABASE MANAGEMENT SYSTEM.

The supervisory control for switching of circuit breakers is usually done by using the displays and the function keys on the keyboard, with automatic checking and verification by the computers.

Automatic generation control. To match the load at all times by controlling generation, an automatic feedback control called load frequency control is used. The errors in frequency and scheduled power interchanges with neighboring utilities are continually checked, and control signals are sent to the generating units to adjust their generation levels. The units that can be controlled are those that the generation owners have agreed to put under control and those that have the ability to ramp their outputs up and down. The generators are compensated for this service at a contracted or market rate. (**Fig. 6**).

Fig. 5. Operator consoles at the energy control center of the New England Power Exchange (NEPEX) in Holyoke, Massachusetts. (*ESCA Corporation,* © *Richard G. Shaw*)

The generation control functions also include reserve monitoring to ensure adequate generation reserve, and interchange scheduling to automatically schedule contracted sales and purchases of energy between companies. If generation reserve is not adequate, the operator has to buy reserve capacity, and if interchange scheduling cannot be maintained at the contracted levels because of reliability considerations, the operator has to follow procedures to change the scheduled interchanges.

Security analysis functions. This set of functions (**Fig. 7**) is the direct result of the availability of powerful digital computers at control centers. A mathematical model of the power system can be stored in the computer and automatically updated in real time by using the measurements that are being continually monitored by the data-acquisition system. A program called the state estimator can track the real-time conditions of the power system with the stored model.

This real-time model can be used to provide the operator with answers to security questions. The contingency analysis program can analyze the effect of all probable contingencies on this model. If equipment failures at particular locations are determined to cause unacceptable operating conditions, the operator is alerted and can change the operating conditions so that these contingencies are no longer threatening to the system. This contingency analysis can be a static analysis (power flow), but for those systems that have stability problems a dynamic analysis is also used.

By using a study power flow program, the operator can also study this model for other purposes. If changes to the operating condition are contemplated by the operator, they can be first studied on this model to determine if they produce the desired results. An enhanced version, called the optimal power flow, can even determine the best operating strategies that should be tried to get the desired results.

Scheduling functions. For generation owners, optimizing costs is important. Steam-generating units have limits on how fast they can be started up or shut down. Hydroelectric units have limited available energy depending on the storage capacity of their reservoirs. These and other constraints affect the commitment and decommitment of generating units to meet the daily and weekly cycles of load. Unit commitment and hydrothermal coordination computer programs are available to determine the most economic scheduling pattern. With the availability of large computers these programs can be integrated with operations data so that the scheduler can update the results conveniently for changing conditions. These operations-planning functions, especially those that plan over many weeks, are often performed by engineering personnel distinct from system operators, on a separate computer system that shares data with the control-center computers.

In regions where generation owners are deregulated, the scheduling of generation depends on long-term (weeks) bilateral contracts and short-term

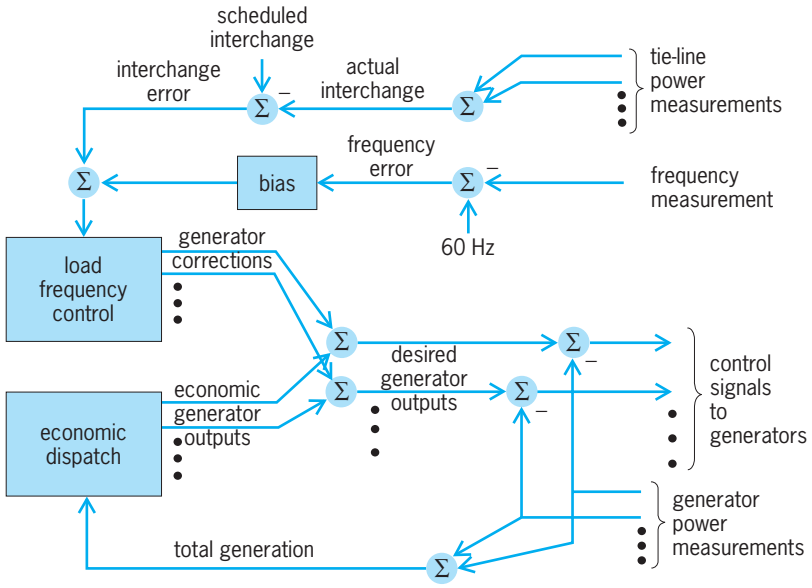


Fig. 6. Flow chart of the automatic generation control functions of an energy control center.

(24 hours or less) market transactions. These markets are computerized and use some sort of auction method to pick the bids and clear the market. Although the basic energy market is straightforward, the markets for reserve capacities and other ancillary services can be rather complicated. Moreover, even after the markets have cleared, the buy–sell transactions have to be checked by the system operator for their feasibility. This is because the markets clear on bid prices without regard to whether the transactions cause congestion (transmission overloads). Thus the operator has to check for congestion and manage the congestion by altering the transactions. This can be done iteratively between the market computers and the reliability studies or in one step by using the optimal power flow.

Logging and accounting. Computers have largely automated the logging and accounting functions, resulting in more comprehensive record keeping of the actual power flows and reduced labor costs. Daily, weekly, and monthly reports are produced automatically.

Computer programs are used to determine the cost of electric energy using the contract or market prices and the actual flows which are always

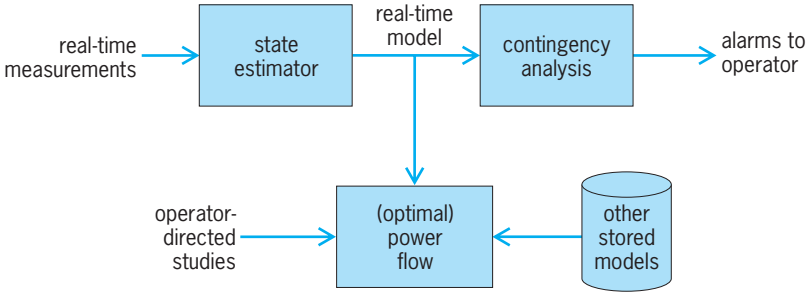


Fig. 7. Flow chart of the security analysis functions of an energy control center.

different from the contracted flows. The purpose is to automatically calculate the billing for energy that was bought and sold between generating companies and the distribution companies or loads. The same data can be used to compensate the transmission companies for their transmission service.

Operator training simulator. As the functions conducted at the control center have become more sophisticated and operating the power system more complex, the training of the operator has also become more important. The use of simulation training, like that used to train airline pilots and air-traffic controllers, has become common because there are no good alternatives for learning to handle emergency situations. The capability of computers to model the power system in real time has made this simulation training possible.

Impact of deregulation. The worldwide trend toward deregulating the electric power industry has had a significant impact on how the electric grid is operated. The presence of more organizations that has resulted from the separation of generating, transmission, and distribution companies has increased the need for new operational procedures to ensure the reliability and security of electricity supply. Here also, superior communication and computation technology play a vital role; in fact, such deregulation would not be feasible without them. Some of the trends in separating the economic functions from the reliability functions are described above, including some of the new tools like market systems, but changes are happening at different rates in different parts of the world and best practices for operating these deregulated systems are still emerging.

Anjan Bose

Power-System Planning

The two traditional objectives of power-system planning are (1) to provide electricity as cheaply as possible while (2) satisfying standards of reliability, safety, and quality. However, since about 1980, social, economic, and technical changes have altered the orientation of power-system planning everywhere. Power system planning, once done by vertically integrated utilities in isolation, has become a cooperative and even a competitive process.

Load forecasting. Planning begins with load forecasts. For decades, demand in the United States grew at about 7% per year and was correlated to the gross domestic product, but since about 1970 the rate of load growth has averaged about 2% per year. Overall demand is forecast in three ways: extrapolation of historic growth, econometric modeling, and end-use forecasting. A fourth technique, based on historic experience, is used to forecast local-area load growth. For day-to-day or hour-to-hour operations forecasting, weather and social patterns dominate.

Extrapolation. If load grows by the same percentage each year, the logarithm of load grows at a linear rate. A straight line passing through and beyond a plot of the logarithm of historical data gives an extrapolation forecast of the logarithm of load; load itself is forecast by taking the antilog. This method fails if the rate of load growth changes. *See* ALGEBRA; EXTRAPOLATION; LOGARITHM.

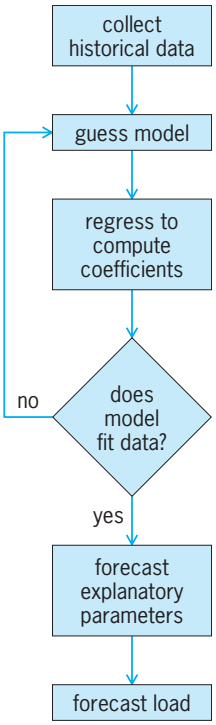


Fig. 8. Econometric load forecasting.

Econometric modeling. Mathematical models (formulas) can be developed that relate changes in load to changes in explanatory variables like population, gross domestic product, prices, and so forth (**Fig. 8**). Historical data are collected on loads and on possible explanatory factors. Often, the first formula or model tried is linear, like the equation below. Regression is

$$\text{LOAD} = k_1 + k_2 \times \text{POPULATION} + k_3 \times \text{GROSSDOMESTICPRODUCT} + \cdots$$

a mathematical procedure for computing the values of the coefficients k_1, k_2, and so forth, to make the model consistent with the historical data. Statistical tests measure just how closely the model matches the data. *See* STATISTICS.

Once the model adequately relates the historical explanatory data to past loads, the explanatory factors are forecast. Finally, the formula or model is used to compute the load forecast from the forecasts of explanatory factors.

Econometric modeling often works better than extrapolation. But both methods infer future load growth from historical data—a reasonable approach only if major changes in the economy do not occur.

End-use forecasting. Residential load forecasts can be built up from energy-use forecasts. The number of homes is forecast first. Major appliances are identified and their market penetrations are projected. The annual energy used by each type of appliance is also projected. This is multiplied by the appliance penetration and by the number of homes to give an energy forecast per appliance. The residential forecast is the sum of the appliance forecasts. This

method takes more effort than the other two and does not work well for commercial or industrial loads, but it depends less on the assumption that the past foretells the future.

Local-area forecasting. Forecasting loads in local areas is important for subtransmission and distribution system planning. Load often develops in an S-shaped pattern, with low historic loads succeeded by rapid growth as new residential or commercial development occurs, followed by low growth again when development saturates. Forecasting is based on experience with prior development patterns.

Generation planning. Centralized generation planning is done by utilities or other organizations having a statutory or implied "obligation to serve" and guaranteed recovery of costs. Generation options include building new utility-owned power plants, buying power from independent power producers or other entities, retiring old generating units or extending their lives, and repowering older units to increase their capacity or efficiency.

Which options are most attractive varies from country to country, and among regions in larger countries, depending on local conditions. For example, there are still vast unexploited hydroelectric power resources in Latin America, Asia, and Africa, while in much of Europe and the United States most of the economically attractive hydroelectric power has been developed.

Plans traditionally are evaluated in terms of two main criteria: reliability and cost (variable and fixed). They must also satisfy environmental and other requirements. Planners gage reliability in terms of reserves (installed generation capacity minus peak demand) or any of several probabilistic measures of the likelihood that enough generators will be available to meet the load.

Variable costs are roughly proportional to how heavily a plant is operated; fuel costs are an example. Variable costs include not only the cost of a particular option but also how it will affect the costs of the rest of the system. Variable costs often are calculated by using a production-cost computer program that simulates the operation of the region's plants.

Fixed costs are independent of how a particular plant is run. They include interest, depreciation, profits, taxes, and so forth. In the United States, fixed and variable costs are often added and expressed as revenue requirements, the amount that customers will have to pay if the utility selects a particular option. To choose among several options, the present worths of their revenue requirements are compared.

Planners also assess the effects of load forecast uncertainty, generator forced outages, possible variations in fuel price, uncertainty in performance of new or rebuilt units, construction cost overruns, and, particularly in less-developed countries, uncertainty in construction time.

While an independent power producer (IPP) can be a generation-planning option from the perspective of a utility, it adds uncertainty and competes with utility-owned generators. It is evaluated much like a utility-owned plant in terms of its effect on cost of electricity, system reliability, and so forth. If the IPP's cost or performance is uncertain, the utility must ensure that it will be allowed to pass cost variations to its customers through tariffs. If there is uncertainty with regard to whether the IPP will be built, the utility must have contingency plans to replace it. If the IPP is viewed as a competitor, the utility attempts to design and operate its own generators at a cost lower than the price the IPP can afford to charge.

Planning is done from an entirely different perspective by IPPs or other entities without guaranteed cost recovery and without an obligation to serve. Decisions are based on market analyses, where load growth is only one of the critical drivers. Other drivers include the costs of existing and possible future generating units with which the IPP will compete. The IPP also considers regulations and market rules, which can affect its operation and its profitability. This type of planning is different in detail but not in concept from the development planning that any commercial organization must make. Planning is complicated by the high initial cost and long economic lives of generating units. Utilities are happy to recover their investments over 40 years. Most investors want less than 10 years.

Transmission planning. A transmission system consists of current-carrying hardware (lines, transformers, structures, and so forth), control and communication equipment (computers, fuses, relays, and so forth), and practices and procedures. The transmission system is coupled to the generation system; together, they form an integrated energy conversion machine. The transmission system is not a transportation system like a gas pipeline, and it cannot be analyzed independent of the generation system.

One transmission planning option is construction of new alternating-current or direct-current lines, as discussed above. Other options include upgrading existing lines, installing transformers or compensation devices (capacitors and reactors), and introducing new operating procedures. Advances in semiconductor devices have greatly increased the variety of control techniques available. Collectively called FACTS (flexible alternating-current transmission system), these control devices allow utilities to utilize more fully the capability of the transmission system and to improve reliability. *See* STATIC VAR COMPENSATOR.

The final decision between two technically equivalent options is based on what each would cost the ratepayer, but the system must meet reliability criteria.

Reliability is assessed with computer programs that simulate network flows by using complex mathematical models of the generation and transmission system. The most sophisticated of these models simultaneously solve tens of thousands of nonlinear equations, with tens of thousands of unknowns; solution techniques are so powerful that this takes a few seconds on a personal computer.

Engineers often plan the transmission system to withstand the sudden loss of any single generating unit or network element under peak-load conditions; this is called the first-contingency criterion. Since this criterion requires the system to withstand the most severe contingency, the system is presumably more than adequate for less serious failures, and can generally withstand more than one of these.

Strategic and integrated resource planning. Strategic planning is concerned with long-range decisions affecting the fundamental nature of an enterprise. Options include acquisitions, divestitures, mergers, and diversification.

Least-cost planning or integrated resource planning (the terms are synonymous) is strategic planning for energy. It differs from tactical generation and transmission planning in three ways. First, integrated resource planning considers more options, with unbiased treatment of supply (generation) options and demand options such as conservation and load shifting. Second, integrated resource planning seeks to reconcile the objectives of a variety of stakeholders. Finally, integrated resource planning deals in a more fundamental way with risk and uncertainty.

Options. Numerous demand options are available, as discussed more fully below. Some are active; the utility can switch off loads when system conditions require it. Others are passive, relying on price signals to influence customers to reduce or shift demand. Others are both active and passive: in demand subscription service, for example, customers are paid to join a program requiring them to reduce load to their subscription level if the utility requests it.

Objectives. While generation and transmission planning concentrate on two main objectives (reliability and ratepayer cost), integrated resource planning also recognizes costs and benefits perceived by the investors, the regional or national economy, neighbors concerned with the environment, utility employees, and so forth. These costs and benefits cannot all be reduced to monetary terms, and they represent objectives that often conflict. For instance, a clean-burning premium fuel is good for the neighbors but expensive for the ratepayer.

The traditional approach for resolving conflicting objectives is to create a "utility function" by simply adding all of the objectives, each multiplied by an appropriate weighting factor. The utility function is then optimized. An extensive body of theory has been developed to support this approach. The basic difficulty is that objectives often are truly incommensurate. Even something as obviously economic as reliability cannot be expressed precisely in monetary terms.

A newer approach is to express all objectives in their natural units and to seek solutions that give the best trade-offs or compromises among these conflicting objectives.

Uncertainties and risk. An uncertainty is a parameter whose value is unknown. Uncertainties can be modeled probabilistically or as unknown-but-bounded variables, depending on the circumstances. Both types of models have their place. A common error is to insist on using a probability model, because it appears to be more elegant mathematically, in circumstances where it is not valid.

For hundreds of years mathematicians and businessmen have studied the problem of risk. Some business risks are quantified in economic terms.

Nonetheless, many risks cannot be so quantified and managed. It is not reasonable to cover risks of terrorism, for example, by simply buying an insurance policy. The risks are not purely economic. Similarly, the risks of a blackout cannot be monetarized.

The approach to risk described here does not require that risks be reduced to dollar terms.

Risk is the hazard to which an entity is exposed because of uncertainties. Measures of risk include the likelihood or conditions under which a particular decision will be regrettable, and the magnitude of the possible regret. A decision that is not likely to be regretted, or where the possible regret is low, is said to be robust.

Integrated resource planners assess risk and develop hedges to manage it. Hedges include building small plants rather than big ones, buying futures contracts to protect against price volatility, doing pilot demand-management studies before selecting a demand-side option, and designing expansion so that it can be staged.

For example, one utility considered two major planning options: building a large coal-fired plant near the load, and building a subarctic hydroelectric plant with a long transmission line. The life-cycle costs were about the same, with the hydroelectric plant perhaps a bit cheaper, but the utility opted for the coal plant to avoid the risks associated with subarctic construction and a long line.

Regulatory changes. The regulatory compact is an unwritten agreement that permits electric utilities to enjoy monopoly protection and guaranteed cost recovery and profits. In exchange they accept an obligation to serve all comers and eschew monopoly profits. This century-old compact is being changed in a variety of ways in much of the world. Some feel that the industry is now mature and does not need special protection, and that a competitive market can do better than a regulated monopoly to meet customer needs at the lowest price.

Changes in the regulatory structure of the industry create important technical options and uncertainties. One key issue concerns the handling of transmission access and power wheeling, the provision of transmission services to energy buyers and sellers who are not directly interconnected. There are difficult engineering and equity problems associated with this issue, and they have not been resolved completely.

Planning in developing countries. Electric power is less available in the developing world because (1) its cost is beyond the means of many families and businesses, and (2) capital is not available for the massive investments needed to build generation and distribution facilities. These barriers can be overcome in part by planning and building the power system differently. Reliability and other quality-of-service standards in Addis Ababa should not be the same as in Tokyo. Furthermore, economic analyses

must use high discount rates (cost of capital) to reflect unavailability of capital to utilities and consumers. This means that minimizing first cost is more important than minimizing life-cycle costs. Equipment and development should be designed to be cheap initially, even if they must be replaced sooner rather than later. Hyde M. Merrill

Power System Reliability

An electric power system serves the basic function of supplying customers, both large and small, with electrical energy as economically and as reliably as possible. The reliability associated with a power system is a measure of its ability to provide an adequate supply of electrical energy for the period of time intended under the operating conditions encountered. Modern society, because of its pattern of social and working habits, has come to expect the power supply to be continuously available on demand. This availability, however, is not physically possible in reality due to random system failures which are generally outside the control of power system engineers, operators, and planners.

The probability of customers getting disconnected can be reduced by increased investment during the planning phase, the operating phase, or both. Overinvestment can lead to excessive operating costs, which must be reflected in the tariff structure. Consequently, the economic constraints can be violated even though the system may be highly reliable. On the other hand, underinvestment can lead to the opposite situation. Design, planning, and operating criteria and techniques have been developed over the past three or four decades in an attempt to resolve and satisfy the dilemma between the economic and reliability constraints.

The criteria and techniques first used in practical applications were basically deterministic (rule-of-thumb) ones; for instance, installed generating capacity equals the expected maximum load demand plus a fixed percentage of the expected maximum demand. The essential weakness of these methods is that they do not account for the probabilistic and stochastic nature of system behavior, customer load demands, and component failures. Such aspects can be considered only through probabilistic criteria.

It is important to conjecture at this point what can be done regarding reliability assessment and why it is necessary. Failures of components, plant, and systems occur randomly; the frequency, duration, and impact of failures vary from one year to the next. Generally, all utilities record details of the events as they occur, and produce a set of performance measures, such as system availability, estimated unsupplied energy, number of incidents, and number of hours of interruption. These measures are valuable since they identify weak areas needing reinforcements and modifications, establish chronological trends in reliability performance, establish existing indices which serve as a guide for acceptable values in future reliability assessments, enable previous predictions to be compared with actual operating experience, and monitor the response to system design changes.

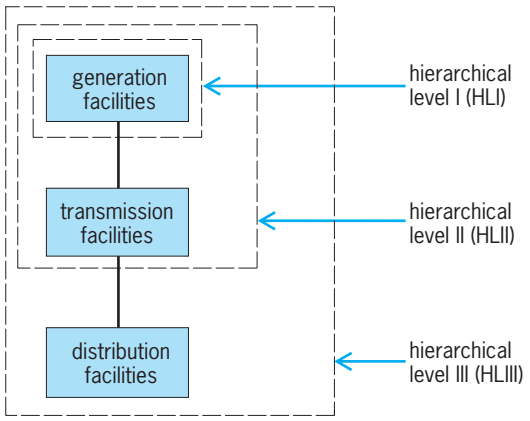


Fig. 9. Hierarchical levels in reliability evaluation.

Hierarchical levels. A modern power system is complex, highly integrated, and usually very large. Analyzing the entire system as a whole, apart from being extremely difficult, may provide results that are so vast that meaningful interpretation will be difficult, if not impossible. The power system can be categorized, for purposes of planning, organization, control, monitoring, operation, maintenance, and so forth, into three distinct zones of generation, transmission, and distribution. While this subdivision of the system may seem somewhat simplistic, most power utilities are divided into these zones or segments, and perhaps some smaller subzones such as subtransmission. The three zones perform quite different functions but must act in unison to supply the customer with reliable electric energy. These zones can also be combined to form three distinct hierarchical levels (HL) [**Fig. 9**], and reliability evaluation can be performed at one or more of these levels. There are two main approaches that can be used to evaluate the reliability of power systems regardless of the hierarchical level: analytical and simulation.

Adequacy and security. The concept of power system reliability, the overall ability of the system to satisfy the customer load requirements economically and reliably, is extremely broad. For the sake of simplicity, it can be divided into the two basic aspects of system adequacy and system security. Adequacy relates to the existence of sufficient facilities within the system to satisfy customer load demands. These include the facilities to generate power, and the associated transmission and distribution facilities required to transport the generated energy to the load points. Adequacy therefore relates to static system conditions. Security pertains to the response of the system to the perturbations or disturbances to which it is subjected. These may include conditions associated with local and widespread disturbances and loss of major generation or transmission. Most of the techniques presently available are in the domain of adequacy assessment.

Reliability cost and reliability worth. Due to the complex and integrated nature of a power system, failures in any part of the system can cause interruptions which range from inconveniencing a small number of local residents to a major and widespread

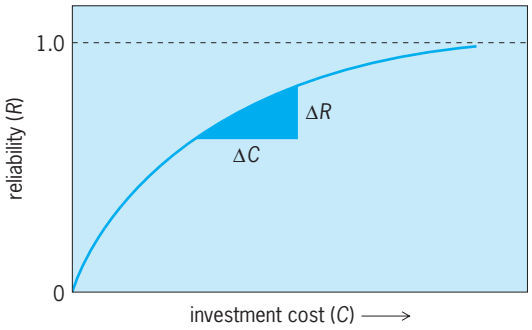


Fig. 10. Incremental cost of reliability.

catastrophic disruption of supply. The economic impact of these outages is not necessarily restricted to loss of revenue by the utility or loss of energy utilization by the customer but, in order to estimate the true costs, should also include indirect costs imposed on customers, society, and the environment due to the outage. For instance, 84% of the total costs of the 1977 New York blackout were attributed to indirect costs. In order to reduce the frequency and duration of these events, it is necessary to invest in the design phase, the operating phase, or both. This involves answering a few difficult questions—how much should be spent? is it worth spending any money? should the reliability be increased, maintained at existing levels, or allowed to degrade? who should decide, the utility, a regulator, the customer? on what basis should the decision be made?

The first step in answering the above questions is illustrated in **Fig. 10**, which shows how the reliability of a product or system is related to the investment cost; that is, increased investment is required in order to improve reliability. This figure clearly shows the general trend that the incremental cost ΔC to achieve a given increase in reliability ΔR increases as the reliability level increases. Alternatively, a given increase in investment produces a decreasing increment in reliability as the reliability is increased. In either case, high reliability is expensive to achieve.

HLI evaluation. The total problem can be divided into two conceptually different areas designated as static and operating capacity requirements. The static capacity area relates to the long-term evaluation of this overall system requirement. The operating capacity area relates to the short-term evaluation of the actual capacity required to meet a given load level.

One practice that has developed among utilities over many years is to measure the adequacy of both the planned and installed capacity in terms of the percentage reserve, which tends to compare the relative adequacy of capacity requirements provided for totally different power systems on the basis of peak loads experienced for each system over the same time period, and also does not attach any penalty to a power-generating unit because of size, unless the unit size exceeds the total capacity reserve.

During planning, it is necessary to determine how much capacity needs to be installed in order to satisfy the expected demand at some point in the future, and to provide sufficient reserve to perform corrective and preventive maintenance. The ability to move the generated energy to bulk supply points and customers is not considered at HLI.

Generating capacity reliability (HLI reliability) is defined in terms of the adequacy of the installed generating capacity to meet the system load demand. Outages of generating units in excess of the estimates could result in "loss of load," that is, the available capacity (installed capacity minus capacity on outage) being inadequate to supply the load. In general, this condition requires emergency assistance from neighboring systems and emergency operating measures such as system voltage reduction and voluntary load curtailment. Depending on the shortage of the available capacity, load shedding may be initiated as the final measure after the emergency actions.

Probabilistic criteria and indices used in HLI studies are:

1. Loss of load probability (LOLP), the probability that the load will exceed the available generation. Its weakness is that it defines the likelihood of encountering trouble (loss of load) but not the severity.

2. Loss of load expectation (LOLE), defined as the average number of days (or hours) per period on which the load is expected to exceed the available capacity.

3. Loss of energy expectation (LOEE), the expected energy that will not be supplied (in kilowatt-hours or megawatt-hours per period) due to those occasions when the load exceeds the available generation. It is presently less used than LOLE but is a more appealing index since it encompasses severity of the deficiencies as well as their likelihood.

4. Frequency and duration (F&D) indices, which measure how often a power outage occurs (frequency, in occurrences per period), and how long the outage lasts (duration, in hours per occurrence) each time it occurs.

The adequacy of the generating capacity in a power system is normally improved by interconnecting the system to another power system. The actual interconnection benefits depend on the installed capacity in each system, the total tie-line capacity, the load levels, and the type of agreement in existence between the systems. A loss-of-load situation is considered to arise in a system when the available assistance through the interconnections cannot offset a capacity deficiency arising due to capacity outages and load demands in that system.

HLII evaluation. An important element in the planning process, which is not considered in HLI studies, is the development of a suitable transmission network to transport the generated energy to the customer load points. The transmission network can be divided into the two general areas of bulk transmission and distribution facilities. The distinction between these two areas cannot be made strictly on a voltage basis but must include the function of the facility within the system.

In addition to providing the means to move the generated energy to the terminal stations, the bulk transmission facilities must be capable of

maintaining adequate voltage levels and loadings within the thermal limits of individual circuits, and also of maintaining system stability limits. The total problem of assessing the adequacy of the generation and bulk power transmission systems in regard to providing a dependable and suitable supply at the terminal stations can be designated as composite system reliability evaluation. Quantitative assessment of the adequacy of a composite system can be performed using a contingency evaluation approach. The basic procedure involves the selection and evaluation of contingencies (component failures), the classification of each contingency according to related failure criteria, and the accumulation of adequacy indices. Two types of indices exist: load point (bus) indices and system indices. Load point indices reflect the adequacy of individual bulk supply points, while the system indices reflect the overall adequacy of the entire system.

HLIII evaluation. HLIII includes all three functional zones in an actual assessment of customer load point reliability. The HLIII reliability performance evaluation is extremely complex. It can, however, be performed (under certain specified conditions) by utilizing the major load bus indices at HLII as input to the distribution networks.

Over the past few decades, distribution systems have received considerably less attention in terms of reliability modeling and evaluation as compared to the generating systems. The main reasons are that the generating stations are individually very capital intensive and that the generation inadequacy can have widespread catastrophic consequences for both society and its environment. Consequently, greater emphasis has been placed on ensuring the adequacy and meeting the needs of this part of the power system. A distribution system, on the other hand, is relatively cheap and outages have very localized effect. Therefore less effort has been devoted to quantitative assessment of various alternative designs and reinforcements. Analysis of customer failure statistics, however, indicates that for many utilities the distribution system makes the greatest contribution to the unavailability of supply to the customer.

The three basic reliability indices of average failure rate (number of failures per period of time), average outage time (duration of failure in hours per occurrence), and average annual outage time (hours per year of power interruptions) are calculated in HLIII studies. Although the three primary indices are fundamentally important, they do not always give a complete representation of the system behavior and response. For example, the same indices will be evaluated irrespective of whether 1 or 100 customers were connected to the load point or whether the average load at a load point was 10 kW or 100 MW. In order to reflect the severity of a system outage, additional reliability indices can be, and frequently are, evaluated: System average interruption frequency index SAIFI, customer average interruption frequency index (CAIFI), system average interruption duration index (SAIDI), customer average interruption duration index (CAIDI), average service availability index (ASAI), and average service unavailability index (ASUI). (ASAI indicates the percentage ratio of customer-hours of service provided to customer-hours requested.) ASUI indicates the percentage ratio of customer-hours of service not provided to customer-hours requested. *See* RELIABILITY, AVAILABILITY, AND MAINTAINABILITY. Lalit Goel

Demand-Side Management

Electricity is produced and distributed in direct response to the demands of residences, business, and industry. These demands come from the billions of electrical devices and appliances which are switched on and off. The demand which each device puts on the electrical system varies by its electrical characteristics, by the way it is used, and by the time of day. Power-system designs are based on plans in which power-system engineers try to forecast the future pattern and amount of the demand for electricity. That prediction is based on a complex set of formulas which consider a number of critical components. These considerations include the electrical demand from the appliances and devices that consume electricity, including predictions of how their population and efficiency may change over time; how new devices may become popular in the future; how people's behavior, or the way people use the devices, may change over time; how economic activity will allow business and industry to flourish and permit enough disposable income to cause consumers to purchase electricity; and the price of electricity and of competing energy forms (such as natural gas).

The role of electric power has grown steadily in both scope and importance over the past century. Developments in key technologies—including electric lighting, motors, computers, and telecommunications—have continuously reshaped daily life and increased the productivity of its commercial and industrial foundations. But the nature of end-use devices is changing dramatically. New uses are emerging from the convergence of developments in energy, telecommunications, transportation, the Internet, and electronic commerce, thus affecting the overall use of electricity.

These new devices require advanced, digital-quality service capabilities. In the United States, the electricity supply system was designed decades ago to nominally maintain 99.99% (5 sigma) reliability, while effectively ignoring outages lasting less than 5 minutes. This was generally acceptable in the twentieth century's analog, electromechanical economy. However, this traditional reliability standard is no longer relevant in today's digital world, where a power interruption or distortion of as little as a quarter of a cycle (0.004 second) can shut down a microprocessor-controlled end-use electrical device or process. This sensitivity infers a systemic (combined supply network plus end-use device) reliability requirement of the order of 10 sigma plus. By comparison, the current system performance results in an electricity unreliability cost to the U.S. economy that has grown to about a 50 surcharge on every dollar of electricity purchased, with no

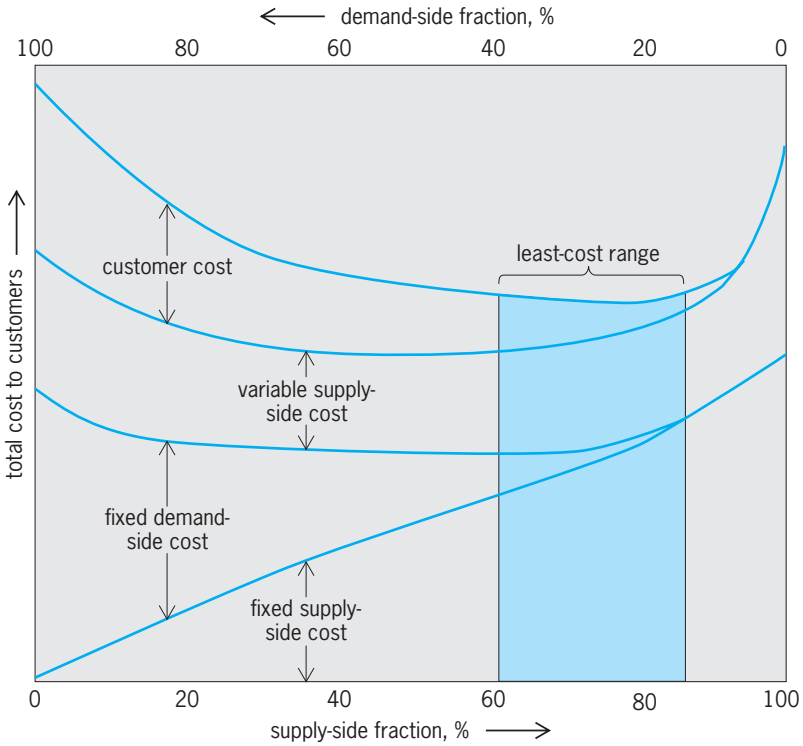


Fig. 11. Balance between supply-side and demand-side management activities.

end in sight until the system is modernized and its reliability transformed. This required performance improvement can most effectively be met by a "systems" approach involving a combination of electricity supply and end-use performance improvements.

Meeting the energy requirements of an increasingly digital society will require applying a combination of advanced technologies—including power generation (for example, central and distributed energy resources), interface devices, and less sensitive end-use equipment and circuits. In addition, the electricity infrastructure must be modernized to incorporate real-time communications, sensors, and computational capability.

The future demand for electricity, estimated based on assumptions about energy-consuming devices and their use, has traditionally been treated as a predetermined quantity by utility planners. Their responsibility has been to forecast that quantity and plan the addition of new generating units accordingly. By changing the efficiency of an appliance, or by changing the way it is used, a utility can alter the amount of future demand or alter the pattern of use in order to change the kind of new electrical generation required. However, traditionally, electrical utilities were not involved in customer programs to accomplish these objectives.

However, new technology is needed if society is to leverage the ever-expanding opportunities of networks such as the Internet and if electric utilities' natural connectivity to consumers is to be effectively utilized to revolutionize both the role of the rapidly changing electricity service industry and the way consumers may be connected to electricity markets in the future. This electricity energy service transformation will also open the door for significant efficiency improvements throughout the energy supply chain while facilitating greater use of relatively clean renewable and distributed energy resources.

In order to increase efficiency, control cost, and meet increasing demand for quality, electrical utilities have begun to address the increasingly complex needs of consumers and to shift from supply-side management, which considers only generating capacity, toward demand-side management. The latter promotes the efficient and lowest-cost utilization of facilities and augments the choices available to utilities and their customers. A compromise between supply-side and demand-side management can produce a range of low-cost resources called the least-cost range (**Fig. 11**).

Definition. Demand-side management is the planning and implementation of those utility activities designed to influence customer use of electricity in ways that will produce desired changes in the utility's load shape, that is, in the time pattern and magnitude of a utility's load. Utility programs in the category of demand-side management include load management, new uses, energy efficiency, electrification, customer generation, and adjustments in market share. Although this broad definition of demand-side management has many merits, in practice the term is generally used more narrowly to encompass increased efficiency, load management, and conservation, all with an emphasis on reducing the need for electrical energy or generation capacity.

Although there is an infinite combination of load-shape-changing possibilities, six changes in two categories can illustrate the range of possibilities: (1) load management including peak clipping, valley filling, load shifting, strategic load growth, and flexible load shape; and (2) energy efficiency. These changes are not mutually exclusive and may frequently be employed in combinations.

Load management. The first step in the shift toward demand-side management was the introduction and evolution of load management. Load-management programs focus on reducing customer use strategically at the time of high utility system loads. The goal is to avoid construction of generation facilities that would be operated for relatively few hours per year or costly power purchases when customer loads can be shifted or displaced at a lesser cost. There are several important components of load-management activities, including direct load control, pricing-based options (interruptible or curtailable rates), and thermal energy storage. The main costs to the utility for load-management programs are rate reductions or bill credits that are used to promote customer participation.

Energy efficiency. Efficiency-type demand-side management programs include various means by which a utility encourages the installation and use of high-efficiency appliances and devices. Programs often involve rebates for buying efficient appliances, bill credits, special rates, advertising, and promotion.

Programs have been focused on efficient, building insulation, heating and air-conditioning equipment installation, water heater wraps, high-efficiency lighting, and efficient refrigerators. Often appliance dealers, wholesalers, electricians, and others are involved in the utility efforts.

Consumer portal. Developments in electric power systems are stimulating the integration of a communications system throughout much of the power system. Once that system is present, connectivity to the ultimate consumers and facilitation of demand-side management can be enhanced with communications. This enhancement will allow three new areas of functionality: one that relates directly to electricity services (for example, added billing information or real-time pricing and demand-side management), one that involves services related to electricity (for example, home security or appliance monitoring); and a third that involves what are more generally thought of as communications services (for example, data services).

The integration of electric energy services and communications with consumer facilities and equipment opens a wide variety of opportunities to enhance consumer energy services, and to provide the consumer with access to a variety of new value-added services while enhancing the performance and operation of the energy system. Since consumer actions create the demand for energy that drives the operation of the energy system, communication with consumer equipment has the potential to profoundly impact the future of the energy system as well as to improve energy efficiency.

A consumer portal (CP) represents the technology, data standards, and protocols that enable the two-way communication of electronic messages with consumer-owned networks and intelligent equipment, and provide a pathway to the full development and implementation of a variety of demand-side management and energy-related services.

Energy saving through electricity. Because of the energy conversion which occurs in producing electricity, it has been argued that a unit of electricity saved is equivalent to preserving three units of primary energy. While this is often the case, concerns about electricity efficiency have sometimes extended to an assumption that electricity use is inefficient and adversely impacts the environment. However, a full understanding of energy efficiency requires a focus on total resource requirements. Electricity offers unique precision and control in its application, enhancing its efficiency. In addition, electricity offers benefits in comfort and environmental advantages. Because of these attributes, in almost every case new electric appliances and devices require less total resources than comparable natural gas or oil-fired systems. Increasing applications of electricity or electrification have been fundamental in increasing national productivity.

As originally defined, demand-side management not only is measured in terms of power-systems savings but also includes full-cycle energy efficiency (total resource use), environmental benefits, and productivity. As such demand-side programs can include not only load management and energy efficiency but new technologies which can expand electricity use.

Full-cycle energy efficiency. Reduced electricity consumption is not necessarily the goal of policy aimed at energy efficiency. A more comprehensive goal is to reduce total consumption of energy resources. For example, the total resource efficiency of an advanced high-efficiency air-source electric heat pump may be compared with the total resource efficiency of the best available gas furnace. By tapping solar energy, the heat pump delivers 3.4 or more units of heat for every unit of electricity produced. An older production and delivery system supplying the heat pump might lose 68% of the fuel in power-plant generation (0.32 factor) and 8% in transmission and distribution (0.92 factor), still yielding 100% efficiency. By contrast, gas has only an 86% efficiency, without considering the requirement for an electric fan. *See* HEAT PUMP.

Other examples where electricity appears to have an advantage in terms of total resource efficiency include ground-source heat pumps, plug-in hybrid electric vehicles, and a variety of advanced industrial processes. In these cases, utility policy that is aimed at energy efficiency may imply increased electricity use.

Environmental benefits. The key environmental benefit stems from efficiency-related demand-side management. By obtaining the same end-use service with less electricity, there is a reduction in power-plant operation and fuel usage, implying a reduction in central-station emissions. These environmental benefits are an important figure of merit in measuring the success of demand-side management. However, most electric appliances and devices are far cleaner at the point of end use. A comprehensive demand-side management strategy must examine the comparative environmental benefits of alternative devices and systems.

Productivity. Application of automation and electric technologies can have important productivity implications. The benefits come in form of faster processing, more precise process control, reduced material and process waste, and higher product quality. The net impact is often reduced process cost, implying reduced product cost and increased competitiveness. Enhanced productivity is an important measure of success, even if these productivity gains require greater electricity usage.

Demand-side management practice. System expansion can be delayed or eliminated and the use of critical resources reduced by significantly reducing energy and peak consumption through demand-side management. Indeed, demand-side management can help achieve a broad range of operational objectives by merely changing the systems load shape. Numerous industries have found that changing the pattern of the demand for their product can be profitable. For example, telephone utilities have long offered reduced evening rates to shift demand and to encourage off-peak usage, airlines offer night coach fares, and movie theaters offer reduced matinee

prices—all examples of deliberate attempts to change the demand pattern for a product or service.

In an electric utility, the physical plant is designed to serve the projected demand for electricity in the least-cost manner, given a specified level of desired quality and reliability. If the load shape is not fixed but may be altered, the cost of serving the load can be reduced still further. Cost reduction due to the changes in the load shape arise primarily from three attributes: reduction in the requirements for new assets or energy, higher utilization of facilities, and more efficient operation of facilities.

Higher utilization of existing and planned facilities can also be achieved through a program of load growth. While such programs increase total costs due to higher fuel and other operating expenses, they reduce unit costs by spreading the fixed costs (debt service and dividends) over more units of energy sold.

Clark W. Gellings

Bibliography. P. M. Anderson and A. A. Fouad, *Power System Control and Stability*, 2d ed., 2003; R. Billinton and R. N. Allan, *Reliability Evaluation of Power Systems*, 2d ed., 1996; Edison Electric Institute, *Statistical Yearbook of the Electrical Utility Industry*, annually; D. G. Fink and H. W. Beaty (eds.), *Standard Handbook for Electrical Engineers*, 14th ed., 2001; J. D. Glover and M. S. Sarma, *Power System Analysis and Design*, 3d ed., 2001; T. Gonen, *Modern Power System Analysis*, 1988; J. J. Grainger and W. D. Stevenson, *Power System Analysis*, 1994; E. Kahn, *Electric Utility Planning and Regulation*, 2d ed., 1991; P. Kundur, *Power System Stability and Control*, 1994; R. H. Miller and J. H. Malinowski, *Power System Operation*, 3d ed., 1994; North American Electric Reliability Council, *Electricity Supply and Demand* (10-year forecast), annually; North American Electric Reliability Council, *Reliability Assessment* (10-year forecast), annually; F. Saccomanno, *Electric Power System: Analysis and Control*, 2003; *Transmission Line Reference Book, 200 kV and Above*, Rep. EL-2500, Electric Power Research Institute, 3d ed., 2005; A. J. Wood and B. F. Wollenberg, *Power Generation, Operation, and Control*, 2d ed., 1996.

Electric power transmission

The transport of generator-produced electric energy to loads. An electric power transmission system interconnects generators and loads and generally provides multiple paths among them. Multiple paths increase system reliability because the failure of one line does not cause a system failure. Most transmission lines operate with three-phase alternating current (ac). The standard frequency in North America is 60 Hz; in Europe, 50 Hz. The three-phase system has three sets of phase conductors, one to four conductors per phase, and system voltage is defined as the root-mean-square voltage between any two of the phases. Long-distance energy transmission occasionally uses high-voltage direct-current (dc) lines. *See* ALTERNATING CURRENT; DIRECT CURRENT; DIRECT-CURRENT TRANSMISSION.

Overhead transmission lines are used in open areas such as interconnections between cities. However, environmentally preferable underground cables are used in congested areas within cities. The transmission lines generate electric and magnetic fields. The electric fields produce harmless but annoying sparks. For example, a worker standing on a wooden ladder may experience mild sparks when a coworker, standing on the ground, hands over a tool. The magnetic fields have become a source of concern, as discussed below. The use of underground cables eliminates the electric fields completely and reduces the magnetic fields in most cases. The sight of a large transmission line is not pleasing for most people. A further problem is the possibility of electrocution due to traffic accidents or accidental contact. The use of cable eliminates both problems.

The electric power system can be divided into the distribution, subtransmission, and transmission systems. With operating voltages less than 34.5 kV, the distribution system carries energy from the local substation to individual households, uses both overhead and underground lines, and is generally divided into high-voltage distribution (above 1000 V) and low-voltage distribution (110–480 V). *See* ELECTRIC DISTRIBUTION SYSTEMS.

With operating voltages of 69–138 kV, the subtransmission system distributes energy within an entire district and regularly uses overhead lines. For example, it may interconnect large substations located outside a town with the distribution substations inside the town. *See* ELECTRIC POWER SUBSTATION.

With operating voltage exceeding 230 kV, the transmission system interconnects generating stations and large substations located close to load centers by using overhead lines; high-voltage underground cables are used for special purposes only. Examples are the supply of large urban substations or underwater cable connections.

Requirements

The energy transportation system must be reliable and highly efficient. Both transmission-line capacity and supplied-power quality must meet variable load requirements.

Reliability. Transmission lines must transport energy continuously to customers in spite of being subject to the environment, that is, pollution, rain, storms, and lightning. System operation can itself produce disturbances which may interrupt line operation. Present utility practices require that the number of transmission line outages be less than one per year per 100 mi (160 km). The most frequent cause of line outages is the single-phase short circuit caused by lightning strikes, pollution-initiated flashover of the insulator, or trees touching the line during a storm.

Safety. Human activity near or under transmission lines may result in short circuits or accidents. A frequent cause of accidents is contact between the line and an operating crane. Transmission lines must

meet strict safety standards. Most important are strict clearances between the conductor and ground, and between the conductor and supporting structures. The electric and magnetic fields which are in an inevitable by-product of system operation have also caused concern, as discussed below.

Capacity. The maximum power that can be transmitted through a line depends on the size of the conductor, voltage drop, and system stability. The current-carrying capacity of a conductor limits the maximum power for short lines, while voltage drop and system stability are the determining factors for long lines.

Economy. Economic operation of the line is determined by system losses, operation costs, maintenance costs, and investment expenses. The dominant losses are the resistance (I^2R) losses in the conductors.

Power quality. Transmission-line voltage depends on the load. Load current produces voltage drop on the line impedance, reducing line voltage, and line capacitance increases voltage at the line end. Voltage at the receiving end of the line is usually low at heavy load during peak hours, and high at light load during night hours. Proper operation of the electrical system requires maintenance of the line voltage within narrow limits. Typical line voltage is kept between ±5%, and regulation is achieved by switched shunt capacitors and shunt reactances.

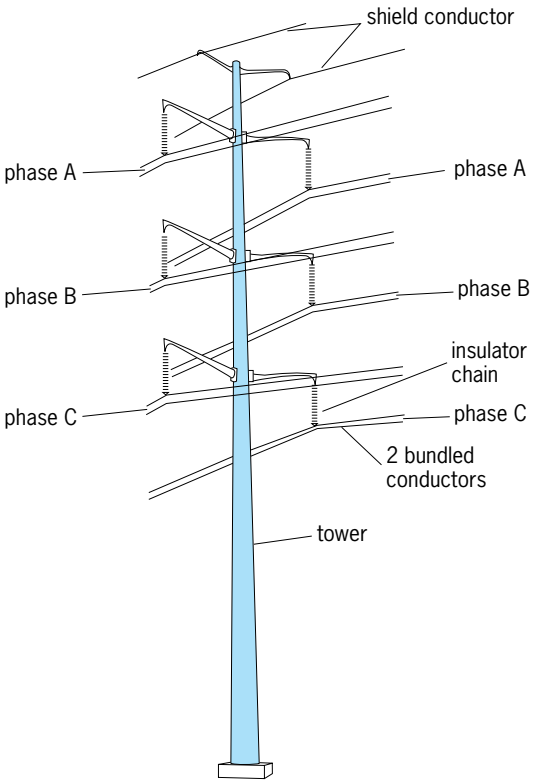


Fig. 1. Double-circuit, high-voltage transmission line.

Overhead Alternating-Current Transmission

Overhead transmission lines distribute the majority of the electric energy in the transmission system. A typical high-voltage line has three (**Fig. 1**) phase conductors to carry the current and transport the energy, and two grounded shield conductors to protect the line from direct lightning strikes. The usually bare conductors are insulated from the supporting towers by insulators attached to grounded towers or poles. The normal distance between the supporting towers is a few hundred feet.

Routes. Most low- and medium-voltage transmission lines are located along highways, roads, or streets. High-voltage lines are placed in transmission-line corridors to reduce interference with the public. Several lines usually share the same corridor.

Hardware. The major components of a transmission line are conductors, insulators, and towers or poles.

Conductors. Transmission lines use ACSR (aluminum cable, steel reinforced) and ACAR (aluminum cable, alloy reinforced) conductors. In an ACSR conductor, a stranded steel core carries the mechanical load, and layers of stranded aluminum surrounding the core carry the current. An ACAR conductor is a stranded cable made of an aluminum alloy with low resistance and high mechanical strength. ACSR conductors are usually used for high-voltage lines, and ACAR conductors for subtransmission and distribution lines. Ultrahigh-voltage (UHV) and extrahigh-voltage (EHV) lines use bundle conductors. Each phase of the line is built with two, three, or four conductors connected in parallel and separated by about 1.5 ft (0.5 m). Bundle conductors reduce corona discharge. *See* CONDUCTOR (ELECTRICITY).

Insulators. Lower-voltage lines use post insulators, while the high-voltage lines are built with insulator chains or long-rod composite insulators. A post insulator is made of either porcelain or composite material and bolted to the cross arm of the tower. The conductor is tied to the top of the insulator. A typical composite post insulator (**Fig. 2***a*) consists of a fiberglass core covered by a silicone or ethylene propylene diene monomer (EPDM) rubber weathershed. The fiberglass provides high mechanical strength, and the rubber assures long leakage distance and good performance in bad weather.

An insulator chain is typically built with ball-and-socket glazed porcelain insulators (Fig. 2*b*). Under typical operating conditions, voltage is around 8–12 kV per unit. Wind swings the vertical chains and reduces clearances to the tower, a movement that is eliminated by a V arrangement.

In a suspension-type composite insulator (Fig. 2*c*), high-strength glass fibers bonded by epoxy or polyester resin form a fiberglass rod that is protected by an injection-molded weathershed made of silicone or EPDM rubber elastomer with hydrated aluminum filler. A corrosion-resistant iron or steel end fitting is glued with epoxy and wedged, crimped, or compressed to the fiberglass rod.

Industrial pollution covers insulators with a contamination layer that becomes conductive when wet by fog or dew, reducing insulator flashover voltage. Pollution-caused flashover results in many service interruptions. In polluted environments, composite

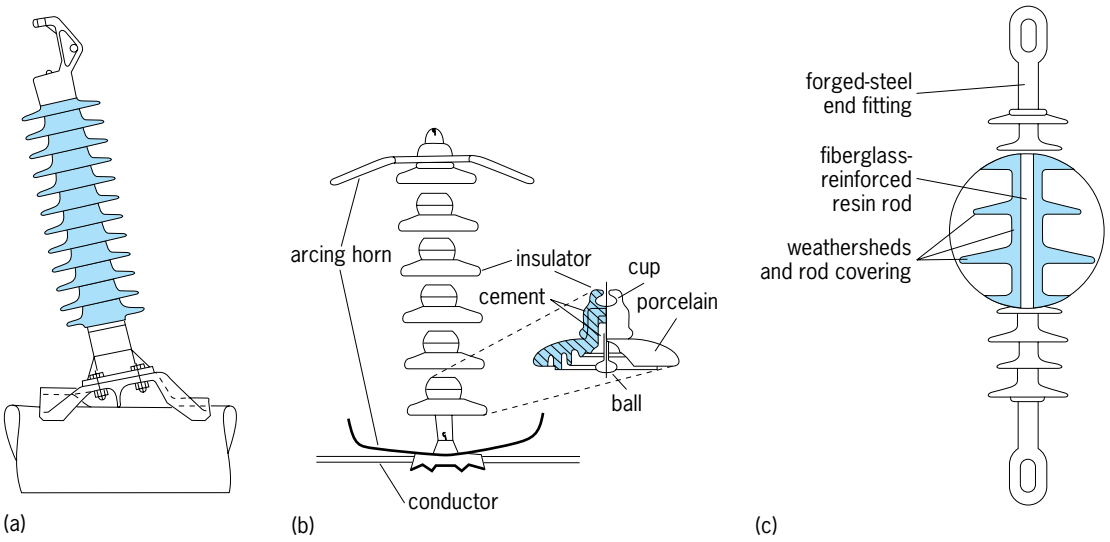


Fig. 2. Transmission-line insulators. (*a*) Composite post insulator. (*b*) Insulator chain. (*c*) Suspension-type composite insulator. (*Sediver, Inc.*)

insulators usually perform better than porcelain insulators. The performance of porcelain insulators can be improved by covering the insulators with silicone grease or a thin composite (silicone rubber) layer. In extremely severe conditions, insulators are washed regularly. Fog-type porcelain insulators are built with deep corrugation on the lower side to increase leakage distance. *See* ELECTRIC INSULATOR; ELECTRICAL INSULATION.

Towers and poles. A transmission line is supported by poles or towers. Distribution lines use wooden poles treated with chemicals (**Figs. 3***a*, *b*). In some areas, wooden poles are replaced by concrete or tapered metal structures, whose higher mechanical strength permits longer span length, reducing the number of towers required for a line. These towers also have longer life spans and better appearance.

Some double-circuit high-voltage lines use lattice

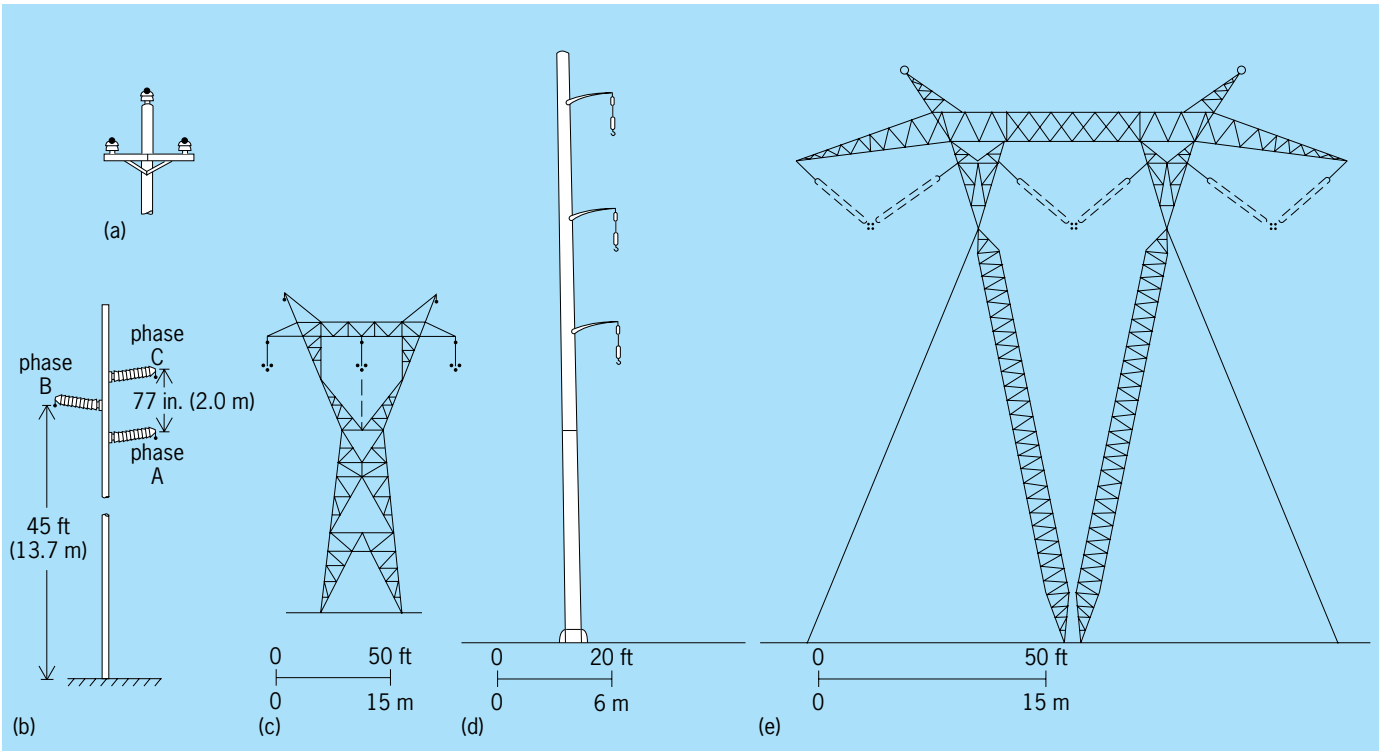


Fig. 3. Transmission-line towers and poles. (*a*) Distribution wood pole with porcelain post insulators. (*b*) Distribution wood pole with composite post insulators. (*c*) Steel tower with lattice structure and insulator chains. (*d*) Tapered, steel-tube tower with insulator chains. (*e*) Guyed V aluminum tower. (*After D. G. Fink and H. W. Beaty, eds., Standard Handbook for Electrical Engineers, 13th ed., Mc Graw-Hill, 1993*)

towers (Fig. 3*c*). Other high-voltage lines built with tubular steel poles (Fig. 3*d*) require less right-of-way and have a more pleasing appearance. Some extrahigh-voltage lines use guided towers in a V arrangement, supported by steel wires (Fig. 3*e*).

A transmission line is divided into sections, each terminated by a stronger dead-end tower with nearly horizontal insulators. From 10 to 25 suspension or tangent towers with vertical insulators are installed between the two dead-end towers, which have only transverse and vertical strength. Dividing the line assures that mechanical failures such as the breakage of conductors will affect only a small section of the line between dead ends. Stronger angle towers are used when lines change direction.

Mechanical performance. Transmission lines are subject to environmental adversities, including wide variations of temperature, high winds, and ice and snow deposits. Typically designed to withstand environmental stresses occurring once every 50–100 years, lines are intended to operate safely in adverse conditions.

Tension and sag. Conductor temperature varies with air temperature and power load, affecting sag and tension. Increased temperature reduces conductor tension and increases sag, reducing ground clearance (**Fig. 4**), while lower temperature increases tension of transmission lines.

Wind and ice loading. Extreme winter weather may coat conductors with ice, which increases conductor weight, as well as tension and sag. High wind produces a horizontal force on the conductor that increases tension and swings the conductor to an inclined position, reducing clearance between the conductor and the tower. Fortunately, high wind, ice loading, and minimum temperature do not occur simultaneously. Combined mechanical loads should not produce tension exceeding the conductor's tensile yield strength, and clearance should be greater than the minimum safety clearance.

Vibration. Nonuniform wind pressure produces conductor vibration. This is a low-amplitude, audio-frequency mechanical oscillation that bends the conductor at the point where it is clamped to the insulator. Stress from repeated bending may cause fatigue breakage of individual wires in the stranded conductor. Ultimately, complete breakage of the conductor may result in service interruptions. Dampers with weights are installed a few feet from the insulators to eliminate vibration-caused conductor breakage.

Galloping. The sudden drop of ice loading in a span causes galloping, a low-frequency, large-amplitude, mechanical oscillation that generates high mechanical stress and reduces the distance between phases in vertical conductors. The danger of galloping can be reduced by melting ice with high currents, using a low-voltage auxiliary transformer to drive the current through a short-circuited line.

Electrical performance. Variable weather affects line operation. Extreme weather reduces corona inception voltage, leading to an increase in audible noise, radio noise, and telephone interference. Load

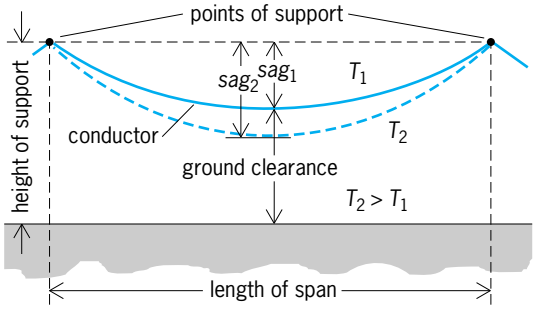


Fig. 4. Catenary curve showing the effect of temperature increase from T_1 to T_2.

variation requires regulation of line voltage. A short circuit generates large currents, overheating conductors and producing permanent damage.

Line parameters. Transmission-line parameters are conductor resistance, inductance, and capacitance. Each phase has different inductance if the distance between conductors is not equal, and different capacitance if conductor heights are different. Longer lines are transposed by cyclic changes of conductor positions, permitting the use of average inductance and capacitance.

Line inductance and capacitance of a single-circuit, three-phase line can be calculated from the distances between the phases and the radius and geometry of the phase conductors. Conductor alternating-current resistances can be obtained from tables. Alternating-current resistance differs slightly from direct-current resistance because of the skin effect, and is also dependent on temperature. Use of a bundle conductor reduces both line reactance and resistance and increases shunt capacitance. *See* CAPACITANCE; ELECTRICAL RESISTANCE; INDUCTANCE; REACTANCE; SKIN EFFECT (ELECTRICITY).

Equivalent circuit. The parameters (inductance, resistance, and capacitance) are distributed along the line. For calculation of the relationship between sending- and receiving-end voltages and currents, distributed parameters are concentrated to an equivalent resistance, inductance, and capacitance (**Fig. 5**). Equivalent values for a short line are calculated by multiplication with the line length; for a long line, more complicated equations are used. An equivalent circuit represents only one phase of a three-phase system; thus, one-third of the three-phase power flows from the sending end (supply) to the receiving end (load). Voltages in this single-phase equivalent circuit are line-to-ground. The relationship between

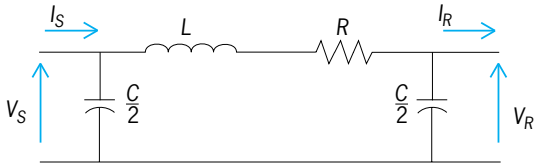


Fig. 5. Single-phase equivalent circuit for a transmission line. Line capacitance (*C*) is divided into two equal parts. V_S, I_S = voltage and current at sending end; V_R, I_R = voltage and current at receiving end.

sending- and receiving-end voltages and currents can be calculated by the Kirchhoff equations. *See* KIRCHHOFF'S LAWS OF ELECTRIC CIRCUITS.

Typically, in no-load or open-circuit conditions, receiving-end voltage is higher than sending-end voltage by as much as 5–25%, and can produce insulator flashover, voltage transformer fault, and so forth. The receiving-end voltage may be limited by shunt inductance, which loads the line. This inductance may be switched off when the load increases.

Increased load produces a voltage drop on the line and may reduce receiving-end voltage below sending-end voltage. This voltage drop can be compensated for by switched capacitors connected in series with the line. Capacitor reactance is negative and subtracted from line reactance, reducing the voltage drop and increasing receiving-end voltage for inductive load. Continuous control of series compensation can be achieved by shunting the series capacitor with a thyristor-controlled reactance. Delay of thyristor firing controls the time in which the inductance is connected in parallel with the capacitor in each half cycle, reducing the effect of capacitance. *See* SEMICONDUCTOR RECTIFIER.

Line power-transfer capacity. The power that a line can transport is limited by the line's electrical parameters. Voltage drop is the most important factor for distribution lines; where the line is supplied only from one end, the permitted voltage drop is about 5%.

The high-voltage line interconnects two buses in the electric system. Each of these buses is directly or indirectly connected to a voltage source. The amplitude and phase-angle difference between bus voltages regulates the direction and amount of power transferred through the line. The power transfer in a line with negligible losses is given by the equation below. Here, the bus voltages at the receiving and

$$P = \frac{V_S V_R}{X} \sin \delta$$

sending ends are V_R and V_S respectively; the phase angle between V_R and V_S is $\delta = \delta_S - \delta_R$, where δ_R and δ_S are the voltage phase angles at the receiving and sending ends; and the line reactance is X. Maximum power transfer occurs when $\delta = 90^\circ$. Stability of this operation requires that δ be not more than 40–60°. Maximum transferred power can be increased by reduction of series impedance by using series compensation. Voltage at both ends can be supported by shunt capacitors or static var compensators, which continuously generate and regulate reactive power. *See* STATIC VAR COMPENSATOR.

Thermal loading limit. Conductor temperature must be lower than the temperature which causes permanent elongation. A typical maximum steady-state value for ACSR is 212°F (100°C), but in an emergency temperatures 10–20% higher are allowed for a short period of time (10 min to 1 h). Steady-state temperature is calculated from the heat balance of the conductor. The sum of resistance (I^2R) losses and solar-radiation-generated heat is equal to the sum of convection- and radiation-produced heat losses. The line is typically designed for 104°F (40°C) ambient, 212°F (100°C) conductor temperature, and 2 ft/s (0.6 m/s) wind speed. Maximum current is reduced by increased altitude and affected by weather conditions. The typical design value for a 500-kV line with two conductors in a bundle is 2400 A.

Corona. This discharge is generated when the electric field at the surface of the conductor becomes larger than the breakdown strength of the air. This high electric field accelerates free electrons to high speed, and the collision of these electrons with air molecules produces ionization. Ionization generates visible light, and fast ion movement produces losses and an audible hissing noise. The oscillatory nature of the discharge generates high-frequency, short-duration current pulses, the source of corona-generated radio and television interference. Because the electric field decreases rapidly with distance, the corona discharge occurs only in the vicinity of the conductor. Surface irregularities such as water droplets cause local field concentration, enhancing corona generation. Thus, during bad weather, corona discharge is more intense and losses are much greater. *See* ELECTRICAL INTERFERENCE.

Corona discharge generates audible noise with two components: a broad-band, high-frequency component, which produces crackling and hissing, and a 120-Hz pure tone. The loudest noise is generated during rain, but low-level noise during fog or wet conditions is equally disturbing; fair-weather noise is inconsequential. *See* CORONA DISCHARGE.

Electric field. Transmission-line conductors are surrounded by an electric field which decreases as distance from the line increases, and depends on line voltage and geometry. Near the conductors, this field generates corona discharge. At ground level, it induces current and voltage in grounded bodies, causes corona in grounded objects, and can induce fuel ignition. Current induced in the human body results in unpleasant sensations such as hair stimulation, and the field produces a secondary spark discharge when an ungrounded person touches a grounded object under the line. Utilities limit the electric field at the perimeter of right-of-ways to about 1000 V/m. *See* ELECTRIC FIELD.

Magnetic field. Load current produces an ac magnetic field around the transmission line that decreases with distance from the line. For example, the maximum magnetic field at ground level under a line which carries 2000 A is 176 milligauss (17.6 microtesla), and decreases to 9.6 milligauss (0.96 microtesla) at 330 ft (100 m) from the line. The field depends on line geometry and current, but is independent of line voltage. The Earth produces a dc magnetic field around 500 mG (50 μT).

Lightning protection. Lightning strikes produce high voltages and traveling waves on transmission lines, causing insulator flashovers and interruption of operation. Steel grounded shield conductors at the tops of the towers (Fig. 1) significantly reduce, but do not eliminate, the probability of direct lightning strikes to phase conductors. The ground resistance of the towers is important. A direct lightning strike to the shield wires or tower drives high current

to the ground. The ohmic voltage drop on the tower's ground resistance increases the tower's potential and may cause insulator flashover, referred to as backflash. Reduction of tower footing resistance reduces the probability of backflash. Proper line design results in 0.5–1 fault per year per 100 mi (160 km). *See* LIGHTNING AND SURGE PROTECTION.

Switching surges. The operation of circuit breakers causes switching surges that can result in interruption of inductive current, energization of lines with trapped charges, and single-phase ground fault. A switching surge is a unidirectional or oscillatory impulse with a 250–500-microsecond rise time and 2–10-μs time to half value. These surges can have peak voltages 2–3.5 times the peak value of the 60-Hz voltage. A switching surge is most dangerous in extrahigh voltage lines (500 kV and above). Modern circuit breakers, operating in two steps, reduce switching surges to 1.5–2 times the 60-Hz voltage. First, the breaker inserts a resistance into the circuit. Then the breaker interrupts the already reduced current, or shorts the resistance and completes the line energization. *See* CIRCUIT BREAKER.

Telephone interference. Line current induces a disturbing voltage in telephone lines running parallel to transmission lines. Because the induced voltage depends on the mutual inductance between the two lines, disturbance can be reduced by increasing the distance between the lines and shielding the telephone lines. Telephone interference has also been reduced by the use of fiber optics and digital telephone techniques. A short circuit may induce dangerous voltages in neighboring telephone lines. *See* ELECTRICAL SHIELDING; INDUCTIVE COORDINATION.

Underground and Submarine Power Transmission

Most cities use underground cables to distribute electrical energy. These cables virtually eliminate negative environmental effects and reduce electrocution hazards. However, they entail significantly higher construction costs, typically 6–10 times those of transmission lines with similar capacity for medium-voltage distribution, and about 20 times higher for high-voltage transmission. High-voltage power cables are also used to transmit energy through wide deep rivers and bays where building transmission lines is difficult.

Cable construction. Underground cables are divided into two categories: distribution cables (less than 69 kV) and high-voltage power-transmission cables (69–500 kV).

Distribution cables. Extruded solid dielectric cables dominate in the 15–33-kV urban distribution system. In a typical arrangement (**Fig. 6***a*), the stranded copper or aluminum conductor is shielded by a semiconductor layer, which reduces the electric stress on the conductor's surface. Cross-linked polyethylene (XLPE) or ethylene-propylene-rubber (EPR) is used most frequently for solid dielectric insulation. The insulation is protected by a semiconducting metallic tape shield which terminates the electric field. A poly(vinyl chloride) [PVC] jacket provides mechanical protection and prevents moisture penetration.

The major problem with these cables is trees generated by electric fields and water penetration. Trees are channels in the dielectric produced by electric discharges. In most cases, they start at the conductor and progress slowly through the insulation, ultimately producing dielectric breakdown and cable failure.

Oil-impregnated paper-insulated cables are used for higher voltages and in older installations (Fig. 6*b*). The stranded copper conductor is shielded by a metallized paper tape, which alleviates the unevenness caused by stranding. The conductors are insulated with oil-impregnated paper tape. Each insulated conductor is covered by metallized paper tape, which produces a concentric electric field. Oil-impregnated paper fillers are used to create a round cable assembly, held together by a metallized paper binder. The assembly in protected from moisture aggression by a lead or lead-alloy shield. Cables designed for insulation in a duct are protected by an extruded PVC sheet. Cables designed for direct burial are covered by a fibrous material (jute) impregnated by bituminous compound and double steel-tape armoring.

High-voltage power-transmission cables. Cable temperatures vary with load changes, and cyclic thermal expansion and contraction may produce voids in the cable. High voltage initiates corona in the voids, gradually destroying cable insulation. Low-pressure oil-filled cable construction reduces void formation. A single-phase concentric cable has a hollow conductor with a central oil channel. Three-phase cables have three oil channels located in the filler. Channels are supplied by low-pressure oil which penetrates the paper insulation and fills the voids

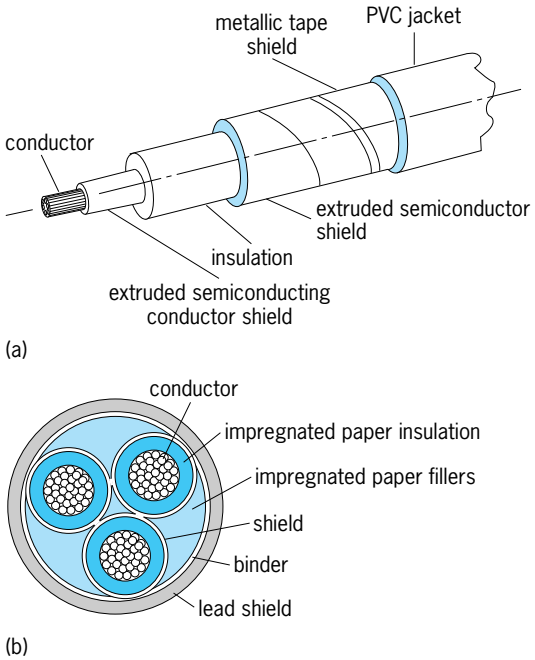


Fig. 6. Underground distribution cables. (*a*) Extruded solid dielectric cable. (*b*) Three-phase oil-impregnated paper-insulated cable. (*After R. Bartnikas Srivastava, Elements of Cable Engineering, Sandford Educational Press, 1980*)

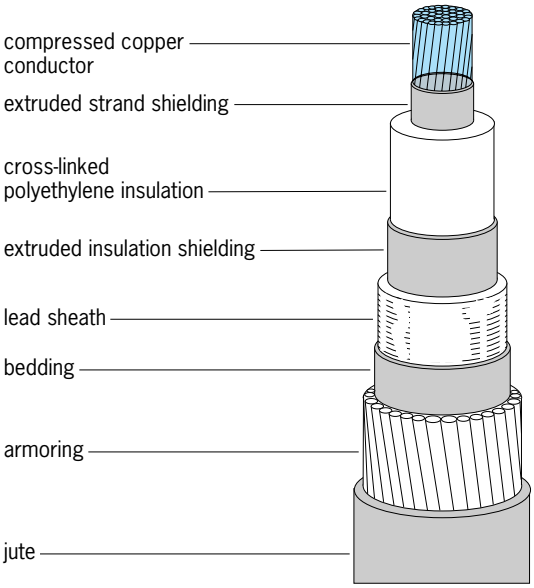


Fig. 7. Solid-dielectric submarine cable. (*ASEA Kabel*)

produced by thermal expansion. Oil is degassed before installation, and the system is hermetically sealed to avoid gas penetration during operation. These cables need complicated joints and oil tanks alongside them, which can cause difficulties. They are used between 69 and 500 kV, although 750-kV cable has been developed.

For extrahigh voltage, high-pressure oil-filled pipe cables are used. Three single-conductor shielded cables are installed in a steel pipe. The pipe is filled with high-pressure (200-lb/in.^{2} or 1.4-megapascal) low-viscosity oil that penetrates the insulation and prevents corona generation. Oil pressure and temperature are maintained by automatically operated pumping stations.

Submarine cables. High-voltage cables are frequently used for crossing large bodies of water. Water provides natural cooling, and pressure reduces the possibility of void formation. A typical submarine cable (**Fig. 7**) has cross-linked polyethylene insulation, and corrosion-resistant aluminum alloy wire armoring that provides tensile strength and permits installation in deep water.

Cable parameters. Cables have significantly larger phase-to-ground capacitance and smaller inductances than overhead lines. Large capacitance results in high charging current, which reduces the power transfer capacity. At critical cable length, charging current is equal to load current. Typical critical length is in the range of 15 mi (25 km) at 500 kV and 45 mi (70 km) at 138 kV. Charging current can be compensated by shunt reactors or static var compensators. Because of the small inductance, reactive components of the voltage drop dominate in a cable system, and the voltage drop is significantly less than in an overhead line.

Cable power-transfer capability is dependent on the temperature of the dielectric, which depends on the heat generated by resistive power loss and on the thermal resistance. In a three-phase system, the unbalanced load current induces a circulating current in the cable shield, which heats the cable and reduces power-transfer capacity.

Cable power-transfer capacity can be improved by externally cooling the cables. One method used in large substations is installing pipes cooled by circulating water near the cables.

Installation. In urban areas, cables are installed in underground concrete ducts requiring access holes every 1000–1500 ft (300–500 m). The ducts facilitate location of faulted cables and provide protection against accidental cutting of the cable during road construction.

In suburban areas, cables may be buried in trenches. Concrete slabs or bricks are often placed on top of the cable for protection and marking. The disadvantages of this method are difficulty in fault location and a lack of mechanical protection.

Cables are manufactured in sections 1000–1500 ft (300–500 m) long, with joints connecting the sections for urban cable. These joints, which are located in the access holes, frequently incur faults, and their preparation requires highly skilled workers. Bending and pulling may damage cables during installation, and proper installation requires expensive equipment and skilled labor. *See* ELECTRIC POWER SYSTEMS; TRANSMISSION LINES.

George G. Karady

Bibliography. Electric Power Research Institute, *Transmission Line Reference Book: 345 kV and Above*, EL 2500, 1987; D. G. Fink and H. W. Beaty (eds.), *Standard Handbook for Electrical Engineers*, 14th ed., 1999; J. A. Williams and P. L. Ostermann, *Underground Transmission Systems Reference Book*, Electrical Power Research Institute, 1992.

Electric protective devices

A particular type of equipment applied to electric power systems to detect abnormal and intolerable conditions and to initiate appropriate corrective actions. These devices include lightning arresters, surge protectors, fuses, and relays with associated circuit breakers, reclosers, and so forth.

From time to time, disturbances in the normal operation of a power system occur. These may be caused by natural phenomena, such as lightning, wind, or snow; by falling objects such as trees; by animal contacts or chewing; by accidental means traceable to reckless drivers, inadvertent acts by plant maintenance personnel, or other acts of humans; or by conditions produced in the system itself, such as switching surges, load swings, or equipment failures. Protective devices must therefore be installed on power systems to ensure continuity of electrical service, to limit injury to people, and to limit damage to equipment when problem situations develop.

Protective devices, like any type of insurance, are applied commensurately with the degree of protection desired or felt necessary for the particular system. Thus, application of these devices varies widely.

Protective relays. Protective relays are compact analog or digital networks, connected to various points of an electrical system, to detect abnormal conditions occurring within their assigned areas. They initiate disconnection of the trouble area by circuit breakers. These relays range from the simple overload unit on house circuit breakers to complex systems used to protect extrahigh-voltage power transmission lines. They operate on voltage, current, current direction, power factor, power, impedance, temperature, and so forth, as well as combinations of these. In all cases there must be a measurable difference between the normal or tolerable operation and the intolerable or unwanted condition. System faults for which the relays respond are generally short circuits between the phase conductors, or between the phases and grounds. Some relays operate on unbalances between the phases, such as an open or reversed phase. A fault in one part of the system affects all other parts. Therefore relays and fuses throughout the power system must be coordinated to ensure the best quality of service to the loads and to avoid operation in the nonfaulted areas unless the trouble is not adequately cleared in a specified time. *See* FUSE (ELECTRICITY); RELAY.

Zone protection. For the purpose of applying protection, the electric power system is divided into five major protection zones: generators; transformers; buses; transmission and distribution lines; and motors (**Fig. 1**). Each block represents a set of protective relays and associated equipment selected to initiate correction or isolation of that area for all anticipated intolerable conditions or trouble. The detection is done by protective relays with a circuit breaker used to physically disconnect the equipment. For other areas of protection *see* GROUNDING; UNINTERRUPTIBLE POWER SUPPLY.

Fault detection. Fault detection is accomplished by a number of techniques. Some of the common methods are the detection of changes in electric current or voltage levels, power direction, ratio of voltage to current, temperature, and comparison of the electrical quantities flowing into a protected area with the quantities flowing out. The last-mentioned is known as differential protection.

Differential protection. This is the most fundamental and widely used protection technique (**Fig. 2**). The inputs to the relays are currents from a current transformer. Current through the relay (nominally 5 or less amperes) is proportional to the high-power current of the main circuit. For load through the equipment or for faults either to the right or left of the current transformers, the secondary currents all sum to essentially zero (Fig. 2*a*). However, for internal trouble, they add up to flow through the relay (Fig. 2*b*).

A typical three-line diagram for a two-circuit differential system is shown in **Fig. 3**. For increased sensitivity, differential relays have restraint and operating windings. On through conditions (Fig. 2*a*) the restraint winding currents desensitize the operating winding, while full sensitivity is available for the internal faults (Fig. 2*b*). Thus for a fault

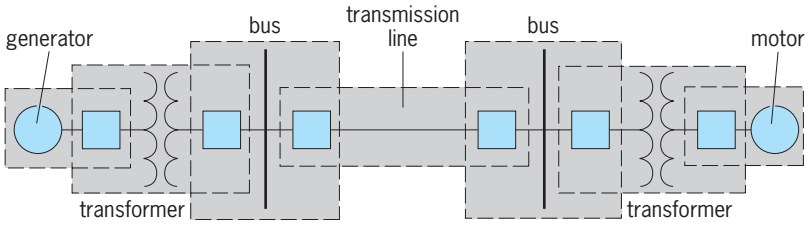


Fig. 1. Zones of protection on simple power system.

between the current transformers, secondary current flows through the restraining windings (with reversed phase in one of the coils) and in the operating coils to operate the relay, trip the two circuit breakers, and isolate the fault or damaged equipment from the rest of the power system.

The equipment to be protected can be a generator, transformer, bus, motor, or line. In the last case, with many miles between the circuit breakers, telephone-type low-energy pilot wires, optical fibers, radio frequency on the power line, or a microwave channel is used. The quantities at the line terminals are thus differentially compared via the channel, in what is known as pilot relaying.

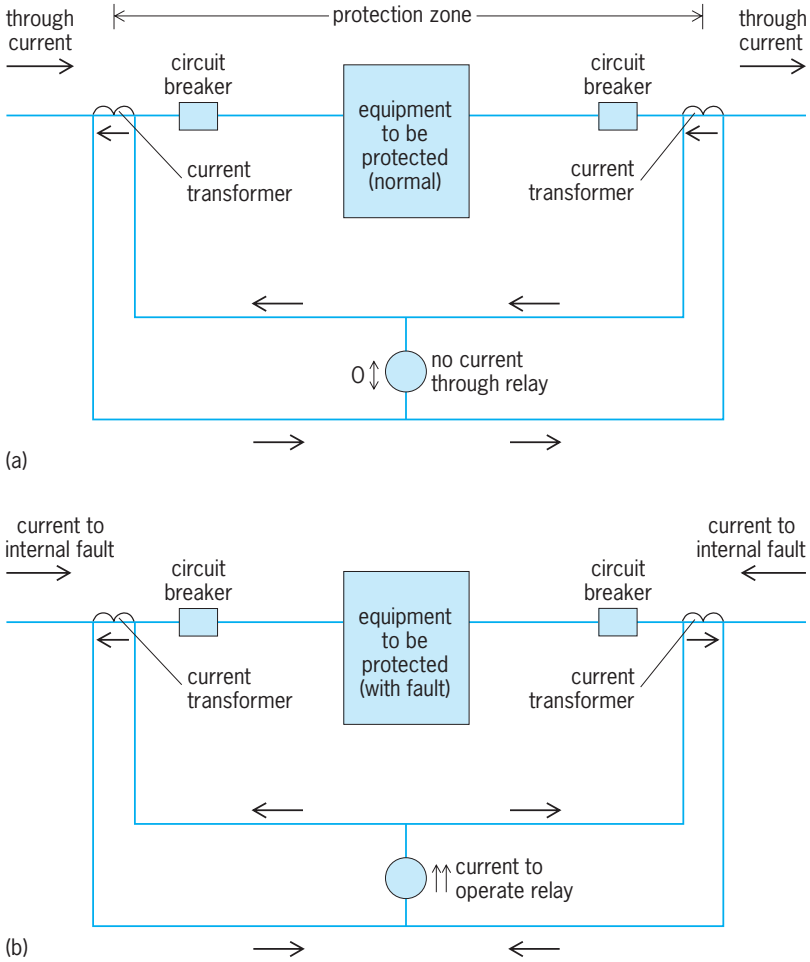


Fig. 2. Simple differential protection scheme. The system compares currents to detect faults in the protection zone. Current transformers reduce the primary current to a small secondary value. (*a*) Normal operation. (*b*) With fault.

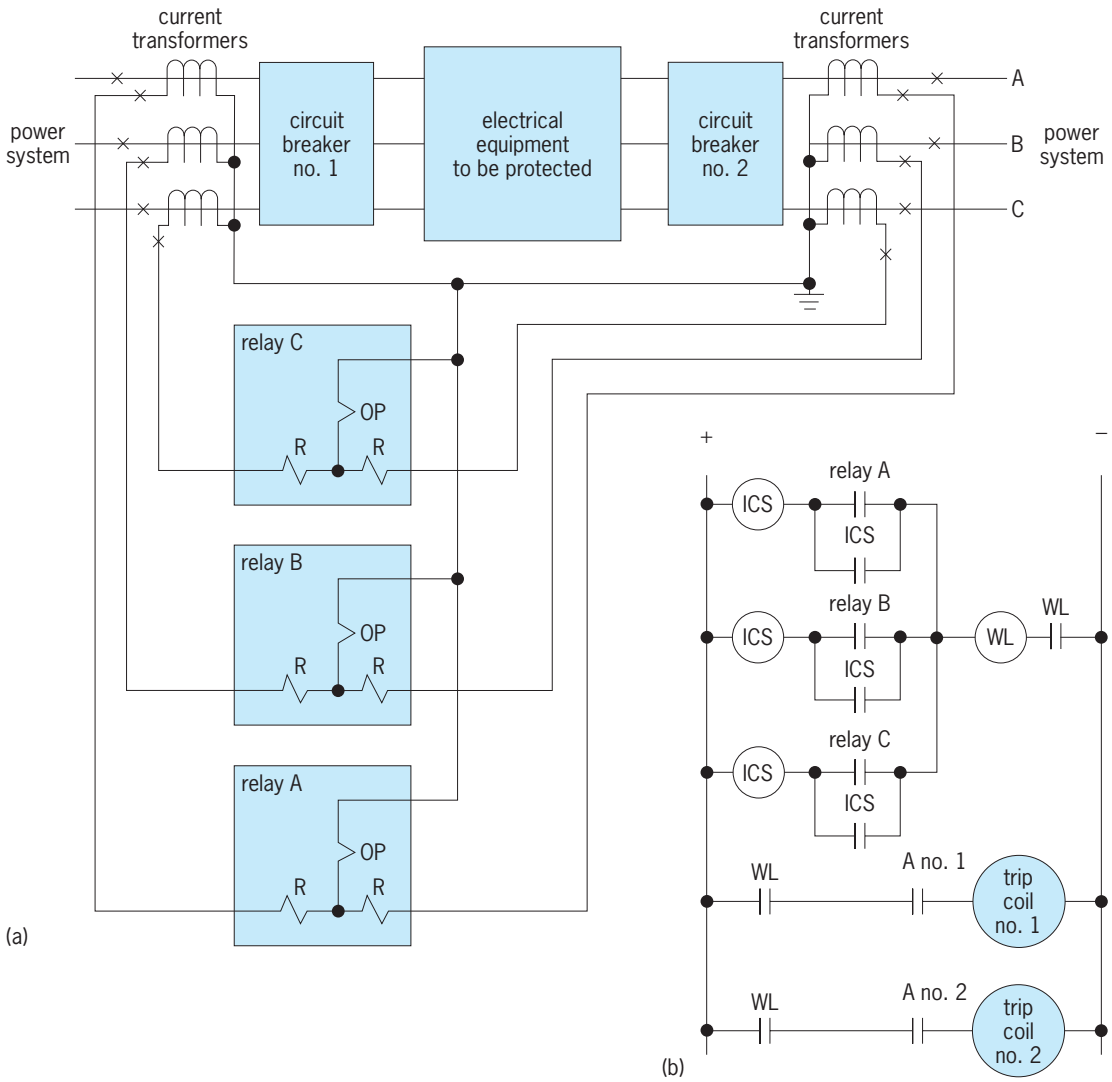


Fig. 3. Differential protection. (*a*) Typical arrangement for electrical equipment. Relays A, B, and C are differential relays. (*b*) Trip circuit. A = auxiliary out-off breaker; ICS = operation indicator and circuit seal-in auxiliary relay; OP = operating coil; R = restraining coil; WL = auxiliary hand-reset lock-out relay.

Overcurrent protection. This must be provided on all systems to prevent abnormally high currents from overheating and causing mechanical stress on equipment. Overcurrent in a power system usually indicates that current is being diverted from its normal path by a short circuit. In low-voltage, distribution-type circuits, such as those found in homes, adequate overcurrent protection can be provided by fuses that melt when current exceeds a predetermined value.

Small thermal-type circuit breakers also provide overcurrent protection for this class of circuit. As the size of circuits and systems increases, the problems associated with interruption of large fault currents dictate the use of power circuit breakers. Normally these breakers are not equipped with elements to sense fault conditions, and therefore overcurrent relays are applied to measure the current continuously. When the current has reached a predetermined value, the relay contacts close. This actuates the trip circuit of a particular breaker, causing it to open and thus isolate the fault. *See* CIRCUIT BREAKER.

Overcurrent relays are of two basic types: instantaneous and inverse-time. Instantaneous units have no intentional time delay and operate in the order of 10 to 50 ms. Inverse-time units operate quickly for high currents and with increasing time for smaller currents.

Overcurrent protection is applied in all parts of the power system. A typical application for the protection of a line between two stations R and S is shown in **Fig. 4***a*. The overcurrent relays are connected to the current transformers to the left of breaker *E* and to the right of breaker *F* to operate for all faults in the primary protection zone between the two stations R and S.

If there is a significant magnitude difference between the close-in fault F_1 and the far bus fault F_2, instantaneous overcurrent relays are set to operate for faults F_1 out to n_1 but not to operate for fault F_2. The operating zone is indicated by the horizontal line from the relays at breaker *E* to n_1, and for the relays at breaker *F* to n_2. These lines represent the

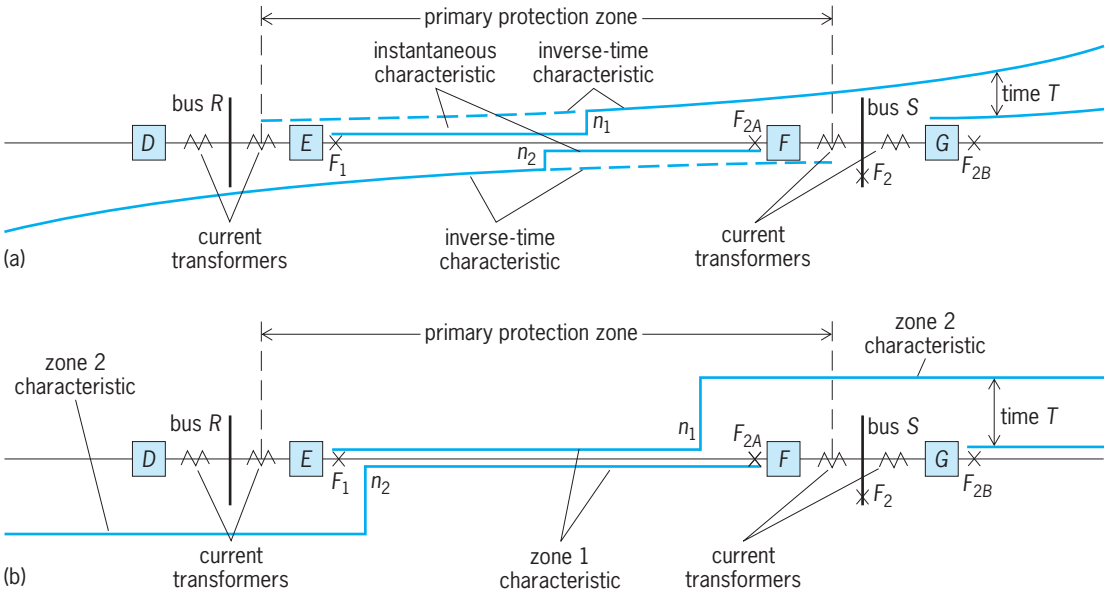


Fig. 4. Typical relay applications and settings. (*a*) Instantaneous and inverse-time overcurrent relays *D*, *E*, *F*, and *G* are circuit breakers. For the relay characteristics shown, operating times are shown vertically. Instantaneous relay time is instantaneous from the relay current transformers to points n_1 and n_2. (*b*) Distance relay. Zone 1 relay characteristic is instantaneous. Zone 2 relay characteristic is instantaneous with operation delayed by time *T*.

maximum coverage. When the fault level decreases, less of the line will be protected.

The current magnitude through the relays at breaker E for a fault at either F_2, F_{2A}, or F_{2B} is the same. Thus it is not possible for the relays at breaker E to determine if fault F_2 is at F_{2A} for which it should operate, or at F_{2B} where it should be cleared by the relays at breaker G. Inverse-time overcurrent relays with directional sensing into the line are applied at both breakers E and G. The relays at E are set to operate as fast as possible for faults at F_1 but to delay for a time interval T for the F_2 faults. Unfortunately this arrangement delays tripping for the internal fault F_{2A}, but it does provide time for the relays and breaker at G to clear the faults F_{2B} and faults to the right of G. If, however, either the relays or breaker G fails to operate for faults F_{2B}, then in time T breaker E will be opened to remove the fault current from the left side. This is backup protection by the relays and breaker at E for faults on the line to the right of breaker G. This process is known as relay coordination or selective settings.

Distance protection. Distance-type relays operate on the combination of reduced voltage and increased current occasioned by faults. They are widely applied for the protection of higher voltage lines. An application is illustrated in Fig. 4*b*. A major advantage is that the operating zone is determined by the line impedance and is almost completely independent of current magnitudes. Typical instantaneous settings n_1 and n_2 are 90% of the total line impedance.

The same difficulty of distinguishing internal faults F_{2A} from external faults F_{2B} at breaker E exists, and so a second distance unit is applied. It is set well into the adjacent lines as shown and is operated through a timer T. This permits the G relays and breaker an opportunity to clear faults F_{2B} and faults to the right before E relays and breaker can operate as backup.

Overvoltage protection. Lightning in the area near the power lines can cause very short-time overvoltages in the system and possible breakdown of the insulation. Protection for these surges consists of lightning arresters connected between the lines and ground. Normally the insulation through these arresters prevents current flow, but they momentarily pass current during the high-voltage transient to limit overvoltage. Overvoltage protection is seldom applied elsewhere except at the generators, where it is part of the voltage regulator and control system. In the distribution system, overvoltage relays are used to control taps of tap-changing transformers or to switch shunt capacitors on and off the circuits. *See* LIGHTNING AND SURGE PROTECTION.

Undervoltage protection. This must be provided on circuits supplying power to motor loads. Low-voltage conditions cause motors to draw excessive currents, which can damage the motors. If a low-voltage condition develops while the motor is running, the relay senses this condition and removes the motor from service.

Undervoltage relays can also be used effectively prior to starting large induction or synchronous motors. These types of motors will not reach their rated speeds if started under a low-voltage condition. Relays can be applied to measure terminal voltage, and if it is below a predetermined value the relay prevents starting of the motor.

Underfrequency protection. A loss or deficiency in the generation supply, the transmission lines, or other components of the system, resulting primarily from faults, can leave the system with an excess of load. Solid-state and digital-type underfrequency relays are connected at various points in the system to detect this resulting decline in the normal system

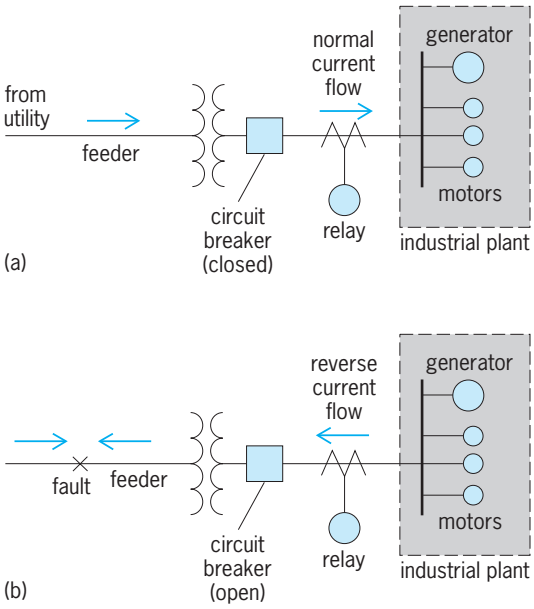


Fig. 5. Reverse-current protection. (*a*) Normal load conditions. (*b*) Internal fault condition; relay trips circuit breaker under reverse-current condition.

frequency. They operate to disconnect loads or to separate the system into areas so that the available generation equals the load until a balance is reestablished.

Reverse-current protection. This is provided when a change in the normal direction of current indicates an abnormal condition in the system. In an ac circuit, reverse current implies a phase shift of the current of nearly 180° from normal. This is actually a change in direction of power flow and can be directed by ac directional relays.

A common application of reverse-current protection is shown in **Fig. 5**. In this example, a utility supplies power to an industrial plant having some generation of its own. Under normal conditions, current flows from the utility to the plant (Fig. 5*a*). In the event of a fault occurring on the utility feeder (Fig. 5*b*) the current reverses direction and flows from the plant to the fault location. The relay operates and trips the circuit breaker, isolating the plant from the utility, thus preventing an excessive burden on the plant generator. Usually in these cases, the plant generator cannot carry the plant load, so that underfrequency relays are used to shed noncritical load. When the utility tie is restored, the shed loads then can be reconnected for full plant service.

Phase unbalance protection. This protection is used on feeders supplying motors where there is a possibility of one phase opening as a result of a fuse failure or a connector failure. One type of relay compares the current in one phase against the currents in the other phases. When the unbalance becomes too great, the relay operates. Another type monitors the three-phase bus voltages for unbalance. Reverse phases will operate this relay.

Reverse-phase-rotation protection. Where direction of rotation is important, electric motors must be protected against phase reversal. A reverse-phase-rotation relay is applied to sense the phase rotation. This relay is a miniature three-phase motor with the same desired direction of rotation as the motor it is protecting. If the direction of rotation is correct, the relay will let the motor start. If incorrect, the sensing relay will prevent the motor starter from operating.

Thermal protection. Motors and generators are particularly subject to overheating due to overloading and mechanical friction. Excessive temperatures lead to deterioration of insulation and increased losses within the machine. Temperature-sensitive elements, located inside the machine, form part of a bridge circuit used to supply current to a relay. When a predetermined temperature is reached, the relay operates, initiating opening of a circuit breaker or sounding of an alarm. J. Lewis Blackburn

Bibliography. J. L. Blackburn, *Protective Relaying: Principles and Applications*, 2d ed., 1997; W. A. Elmore (ed.), *Protective Relaying: Theories and Applications*, 1994; S. H. Horowitz, *Protective Relaying for Power Systems*, 2 vols., 1981, 1992; Institute of Electrical and Electronics Engineers, *IEEE Guides and Standards for Protective Relaying Systems*, rev. ed., 1991; National Association of Relay Manufacturers, *Engineers Relay Handbook*, 4th ed., 1992.

Electric rotating machinery

Any form of apparatus, having a rotating member, which generates, converts, transforms, or modifies electric power. Essentially all of the world's electric power is produced by rotating electrical generators, and about 70% of this energy is consumed in driving electric motors. Therefore, in a large sense, modern technology is built on the foundation of electric machinery. Electric machines are electromechanical energy converters; generators convert mechanical energy into electrical energy and motors convert electrical energy into mechanical energy. The basic laws of physics on which electric machinery is based were developed in the late nineteenth century.

Physical laws. It is possible to construct electric machines that are based on the energy stored in electric or magnetic fields. Those using electric fields are not efficient except in very small sizes, and as a result most machines are based on magnetic attraction or repulsion. The operation of electric machines can be explained in terms of two laws of physics: Ampère's force law and Faraday's law of induction. When a conductor is placed in a magnetic field, it experiences a force **F** that is proportional to the current i and to the strength of the magnetic field **B**. This force is in a direction at right angles to both the current and the field. This is known as Ampère's force law and can be expressed mathematically as Eq. (1), where **l** is the length of the conductor. Ampère's law has numerous equivalent statements, from any of which the production of torque

$$\mathbf{F} = i\mathbf{l} \times \mathbf{B} \tag{1}$$

can be explained. The second law is Faraday's law, which states that an electromotive force **E** will be induced in a circuit when that circuit experiences a change in flux linkage λ. The magnitude of **E** is proportional to the time rate of change of flux linkage. The electromotive force will be directed so that it produces currents that will oppose the change of flux linkage. Faraday's law can be written as Eq. (2).

$$\mathbf{E} = \frac{d\lambda}{dt} \tag{2}$$

See ELECTRIC FIELD; ELECTROMAGNETIC INDUCTION.

Principles of operation. An electric machine can be constructed on the principle that a magnet will attract a piece of permeable magnetic material such as iron or magnetic steel. In **Fig. 1***a*, a pole structure is shown along with a magnetic block that is allowed to rotate. The magnetic block will experience a torque tending to rotate it counterclockwise to the vertical direction. This torque, called a reluctance or saliency torque, will be in the direction to minimize the reluctance of the magnetic circuit. In Fig. 1*b*, a winding is added to the rotor (the part which is allowed to rotate). In this case there is an additional torque on the rotor in the counterclockwise direction produced by the attraction of opposite poles. This torque will be approximately proportional to the sine of the angle θ. While the magnets in Fig. 1 are electromagnets, permanent magnets could be used with the same effect. *See* MAGNET; WINDINGS IN ELECTRIC MACHINERY.

In these examples, if the rotor were allowed to move under the influence of the magnetic forces, it would eventually come to rest at an equilibrium position, $\theta = 0$. Since most applications require continuous motion and constant torque, it is necessary to keep the angle between the rotor magnetic field and the stator magnetic field constant. Thus, in the above examples, the stator magnetic field must rotate ahead of the rotor. In a polyphase machine this is accomplished by distributing the winding so as to produce a rotating magnetic field. In the case of a three-phase machine the windings are distributed

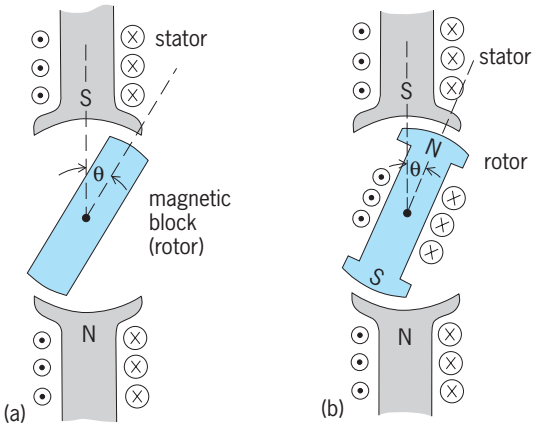


Fig. 1. Devices illustrating principles of electric machines. (*a*) Permeable rotor and stator with magnetic pole structure. (*b*) Device with magnetic pole structures on both stator and rotor.

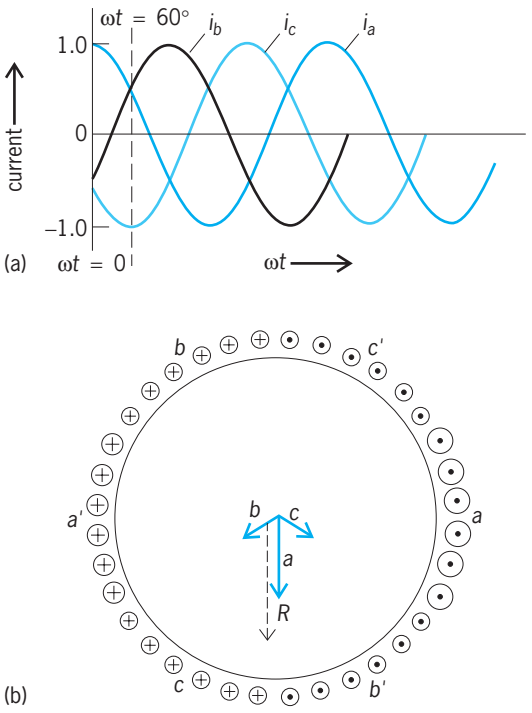


Fig. 2. Three-phase machine. (*a*) Three-phase balanced currents with angular frequency ω. (*b*) Distribution of currents and fields at $\omega t = 60^\circ$.

in space so that their magnetic axes lie 120° apart. During normal operation the currents in the three windings (i_a, i_b and i_c in phases *a*, *b*, and *c*; **Fig. 2***a*) have the same magnitude but are 120° out of phase. (Such currents are called balanced currents.) At time $t = 0$ the current in phase *a* is maximum, and the currents in phases *b* and *c* are $-\frac{1}{2}$ of their peak values. The magnetic field *R* due to the combined three-phase system lies in the same direction that phase *a* current alone would produce (Fig. 2*b*). At a later time, $\omega t = 60^\circ$, where ω is the angular frequency of the currents, the current in phase *c* is at its negative peak, while the currents in phases *a* and *b* are $\frac{1}{2}$ of their peak value. The resultant field is along the (negative) magnetic axis of phase *c*, the total distribution having rotated 60° in space. It can be demonstrated that the fundamental spatial component of the field has constant magnitude and rotates uniformly at angular frequency ω. In this way it is possible to keep a constant angle between the stator and rotor magnetic fields and achieve constant torque. *See* ALTERNATING CURRENT.

In single-phase machines, there is no starting torque, and a device such as a capacitor or a short-circuiting ring is used to simulate the effect of a rotating distribution at starting. The single-phase winding with sinusoidal current produces two rotating pole distributions, one rotating in the forward direction and one in the backward direction. In single-phase alternating-current machines, once the rotor is rotating (in either direction) there is a net magnetic torque in the direction of rotation for motor operation and opposite the direction of rotation for generator operation.

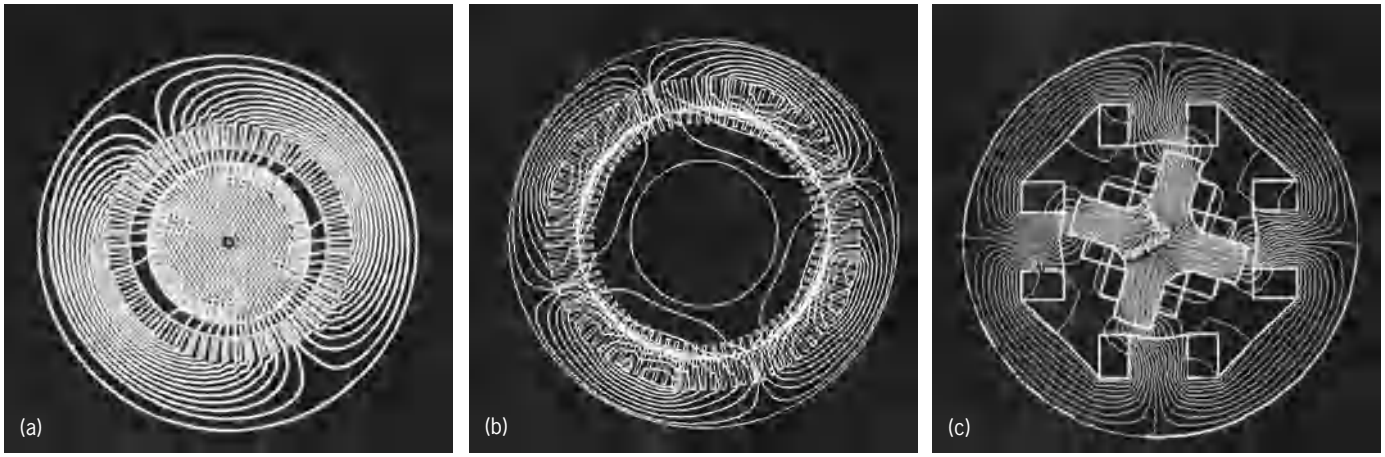


Fig. 3. Magnetic flux lines in rotating electric machines operating under load. (*a*) Three-phase, round-rotor synchronous generator (*Rensselaer Polytechnic Institute*). (*b*) Three-phase, squirrel-cage induction motor (*Rodolpho Palma, Rensselaer Polytechnic Institute*). (*c*) Direct-current motor (*MAGSOFT Corp.*).

Types. Although there are many variations, the three basic machine types are synchronous, induction, and direct-current machines. These machines may be used as motors or as generators, but the basic principles of operation remain the same.

Synchronous machines. The synchronous machine runs at a constant speed determined by the line frequency. There is an alternating-current winding (normally on the stator) and a direct-current winding (normally on the rotor). **Figure 3***a* shows the magnetic field lines for a round-rotor synchronous generator operating under load. *See* ALTERNATING-CURRENT GENERATOR; ALTERNATING-CURRENT MOTOR; SYNCHRONOUS MOTOR.

Induction machines. The induction machine is another alternating-current machine which runs close to synchronous speed. The alternating-current winding of the stator is similar to that of the synchronous machine. The rotor may have an insulated winding (wound rotor) but more commonly consists of uninsulated bars embedded in a laminated structure and short-circuited at the end (squirrel cage). There is normally no voltage applied to the rotor. The voltages are produced by means of Faraday's law of induction. In an induction motor the stator-produced flux-density wave rotates slightly faster than the rotor during normal operation, and the flux linkages on the rotor therefore vary at low frequency. The rotor currents induced by these time-varying flux linkages produce a magnetic field distribution that rotates at the same speed as the stator-produced flux wave. The flux distribution of a squirrel-cage induction motor under load is shown in Fig. 3*b*. *See* INDUCTION MOTOR.

Direct-current machines. In a direct-current motor, direct current is applied to both the rotor and the stator. The stationary poles on the stator produce a stationary magnetic field distribution. Since the angle between the stator-produced poles and rotor-produced poles must remain constant, the direct-current machine uses a device known as a commutator which switches the current from one rotor circuit to another so that the resulting field is stationary. A direct-current motor operating under load is depicted in Fig. 3*c*. *See* DIRECT-CURRENT GENERATOR; DIRECT-CURRENT MOTOR.

Additional systems. Besides the magnetic system that produces the energy conversion in rotating electric machines, a number of other systems operate in these devices.

Electric circuits. The windings of the machines may be any combination of series or parallel circuits, possibly with multiple turns in each circuit using one or many phases. The geometrical layout of these circuits is an important consideration in the electrical design because it affects the efficiency and performance of the device. *See* CIRCUIT (ELECTRICITY).

Thermal system. Heat is generated in the windings because of Joule heating and in the magnetic circuit by eddy currents and hysteresis. When the rotor is turning, there are also frictional and aerodynamic losses. The heat associated with all of these losses must be removed. Ventilation of the machine to limit the temperature to an acceptable level is often the limiting factor in machine design. The most commonly used cooling medium is air. Large machines may use high-pressure hydrogen or purified water as a coolant. *See* AIR COOLING; CORE LOSS; EDDY CURRENT; ELECTRICAL RESISTANCE.

Insulation system. Machines vary in voltage levels from a fraction of a volt to many thousands of volts. A dielectric material is used to coat the windings and insulate them from each other and from other parts of the machine. Deterioration of insulation with age and with increased temperature is one factor that limits the lifetime of machines. Mechanical forces on insulation within windings and between windings and other parts are also a limit on the design. *See* DIELECTRIC MATERIALS; ELECTRICAL INSULATION.

Mechanical system. The rotor on a rotating machine must be free to turn. This means that a system of bearings (often with a lubricating system) must be provided. The rotor windings must be supported

against the centrifugal forces due to rotation. The electromagnetic torque that is applied to the rotor is also applied, in an opposite direction, to the stator, which must be suitably supported. If vibrations are not limited to acceptable levels, excessive audible noise or fatigue failures may occur. *See* ACOUSTIC NOISE; ANTIFRICTION BEARING; MECHANICAL VIBRATION; METAL, MECHANICAL PROPERTIES OF.

With these different modes of operation and design considerations, thousands of variations of motors and generators have been developed for industrial and domestic use. *See* GENERATOR; MOTOR.

Sheppard Salon

Bibliography. B. Chalmers and A. Williamson, *A. C. Machines, Electromagnetics and Design*, 1991; A. E. Fitzgerald, C. Kingsley, and S. D. Umans, *Electric Machinery*, 6th ed., 2003; I. L. Kosow, *Electric Machinery and Transformers*, 2d ed., 1991; S. A. Nasar and I. Boldea, *Electric Machines: Dynamics and Control*, 1992; M. G. Say, *Alternating Current Machines*, 5th ed., 1983; M. G. Say, *Direct Current Machines*, 2d ed., 1986.

Electric spark

A transient form of gaseous conduction. This type of discharge is difficult to define, and no universally accepted definition exists. It can perhaps best be thought of as the transition between two more or less stable forms of gaseous conduction. For example, the transitional breakdown which occurs in the transition from a glow to an arc discharge may be thought of as a spark.

Electric sparks play an important part in many physical effects. Usually these are harmful and undesirable effects, ranging from the gradual destruction of contacts in a conventional electrical switch to the large-scale havoc resulting from lightning discharges. Sometimes, however, the spark may be very useful. Examples are its function in the ignition system of an automobile, its use as an intense short-duration illumination source in high-speed photography, and its use as a source of excitation in spectroscopy. In the second case the spark may actually perform the function of the camera shutter, because its extinction renders the camera insensitive. *See* SPECTROSCOPY; STROBOSCOPIC PHOTOGRAPHY.

Mechanisms. The spark is probably the most complicated of all forms of gaseous conduction. It is exceedingly difficult to study, because it is a transient and because there are so many variables in the system. Some of these variables are the components of the gaseous medium, the gas pressure, the chemical form of the electrodes, the physical shape of the electrodes, the microscopic physical surface structure, the surface temperature, the electrode separation, the functional dependence of potential drop on time, and the presence or absence of external ionizing agents. One or more of these conditions may change from one spark to the next. Because of the great complexity, it will be impossible to do more than touch on some of the main features in this article.

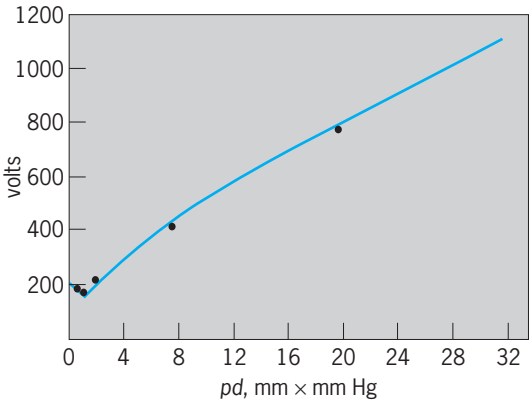


Fig. 1. Dependence of sparking potential on *pd* for a sodium cathode in hydrogen gas. (*After L. B. Loeb and J. M. Meek, The Mechanism of the Electric Spark, Stanford University Press, 1941*)

The dependence of breakdown, or sparking, potential on pressure p and electrode separation d is considered first. It was shown experimentally by F. Paschen and theoretically by J. S. Townsend that the sparking potential is a function of the product pd and not of p or d separately (**Fig. 1**). Further, there is a value of pd for which the sparking potential is a minimum. Thus, if it is desired to prevent sparking between two electrodes, the region may be evacuated or raised to a high pressure. The latter method is used in accelerators of the electrostatic generator variety. Here the entire apparatus is placed in a pressurized tank.

Qualitatively, one of the aspects of a spark is that the entire path between electrodes is ionized. It is the photon emission from recombination and decay of excited states which gives rise to the light from the spark. Further, if the spark leads to a stable conduction state, the cathode must be capable of supplying the needed secondary electrons, and the conduction state produced must permit the discharge of the interelectrode capacitance at the very minimum. *See* ARC DISCHARGE; ELECTRICAL CONDUCTION IN GASES; GLOW DISCHARGE.

In a consideration of the mechanism involved in the spark, the time required for the breakdown of the gas in a gap is an important element. L. B. Loeb pointed out that this time is often less than that required for an electron to traverse the gap completely. This implies that there must be some means of ionization present other than electron impact and that the velocity of propagation of this ionizing agent or mechanism must be much greater than the electron velocity. It seems definitely established that this additional method must be photoionization. In the intense electric field which is necessary for the spark, the initial electron will produce a heavy avalanche of cumulative ionization. Light resulting from the decay processes will produce ionization throughout the gas and electrons at the surfaces by the photoelectric

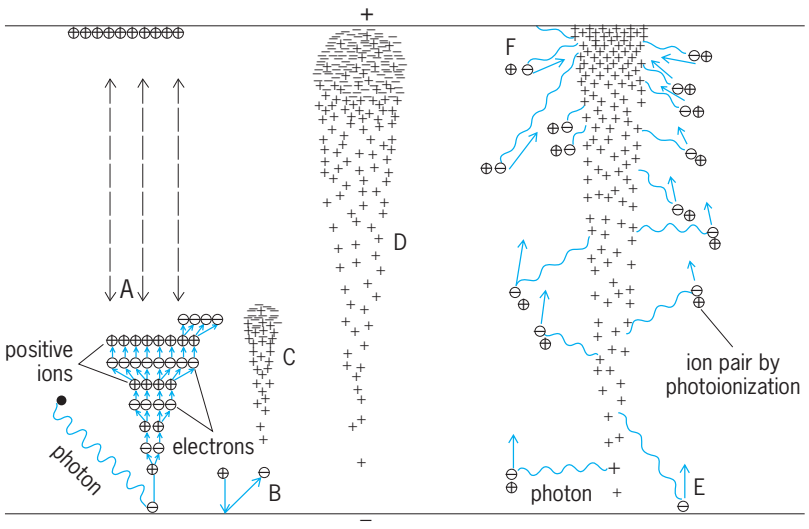


Fig. 2. Electron multiplication and avalanching during an electric spark discharge. (*After L. B. Loeb and J. M. Meek, The Mechanism of the Electric Spark, Stanford University Press, 1941*)

effect (**Fig. 2**). The electrons resulting from this will in turn produce further avalanches through the entire region, so that in a time of the order of 10^{-8} s the entire path becomes conducting. If the pressure is approximately atmospheric, the spark will be confined to a relatively narrow region, so that the conducting path, while not straight, will be a well-defined line. If the external circuit can supply the necessary current, the spark will result in an arc discharge. At lower pressure the path becomes more diffuse, and the discharge takes on either a glow or arc characteristic.

Figure 2 shows A, the electron multiplication of electrons by the cumulative ionization of a single electron liberated from the cathode by a photon; B, a secondary electron emitted from the cathode by a positively charged ion; C, the development and structure of an avalanche, with positively charged ions behind electrons at the tip; D, the avalanche crossing the gap and spreading by diffusion; and F, an older avalanche when electrons have disappeared into the anode. A positive space-charge boss appears on the cathode at F. Ion pairs out from the trail indicate the appearance of photoelectric ion pairs in the gas produced by photons from the avalanche. E shows a photoelectron from the surface of the cathode produced by the avalanche.

Theory. Mathematically, the theory of Townsend predicts that the current in a self-sustained discharge of the glow variety will follow Eq. (1), where I is

$$I = I_0 \frac{e^{\alpha x}}{1 - \gamma e^{\alpha x}} \tag{1}$$

the current with a given plate separation x, I_0 is the current when x approaches zero, and α and γ are constants associated with the Townsend coefficients. This equation represents the case where the electrode separation is varied while the ratio of electric field to pressure is held constant. The condition for a spark is that the denominator approach zero, which may be stated as in Eq. (2). Loeb indi-

$$\gamma = e^{-\alpha x} \tag{2}$$

cated that this criterion must be handled carefully. Townsend's equation really represents a steady-state situation, and it is here being used to explain a transient effect. If the processes which are involved are examined more carefully, it appears that there should be a dependence on I_0 as well. Glenn H. Miller

Bibliography. M. N. Hirsh and H. J. Oskam (eds.), *Gaseous Electronics*, 1978; Y. P. Raizer and J. E. Allen, *Gas Discharge Physics*, 1997.

Electric susceptibility

A dimensionless parameter measuring the ease of polarization of a dielectric. In a vacuum, the electric flux density $\mathbf{D}$ (measured in coulombs/m^2) and electric field strength $\mathbf{E}$ (volts/m) are related by Eq. (1) where ϵ_0 is the permittivity of free space,

$$\mathbf{D} = \epsilon_0 \mathbf{E} \tag{1}$$

having the approximate value 8.854×10^{-12} farad/m. In a dielectric material, polarization occurs, and the flux density has the value given by Eq. (2), where

$$\mathbf{D} = \epsilon_0 \epsilon_r \mathbf{E} = \epsilon_0 \mathbf{E} + \mathbf{P} \tag{2}$$

ϵ_r is the relative permittivity of the material and $\mathbf{P}$ is the polarization flux density. This can be written as Eq. (3), where $\chi_e = \epsilon_r - 1$ is known as the electric

$$\mathbf{P} = \epsilon_0(\epsilon_r - 1)\mathbf{E} = \epsilon_0 \chi_e \mathbf{E} \tag{3}$$

susceptibility of the dielectric material. It is a measure of that part of the relative permittivity which is due to the material itself.

The electric susceptibility can be related to the polarizability α by expressing the polarization in terms of molecular parameters. Thus Eqs. (4) hold, where

$$\mathbf{P} = N\langle \mu \rangle_{\text{avg}} = N\alpha \mathbf{E}_L$$
$$\chi_e = \frac{N\alpha \mathbf{E}_L}{\epsilon_0 \mathbf{E}} \tag{4}$$

N is the number of molecules per unit volume, $\langle \boldsymbol{\mu}_{\text{avg}} \rangle$ is their average dipole moment, and $\mathbf{E}_L$ is the local electric field strength at a molecular site. At low concentrations, $\mathbf{E}_L$ approaches $\mathbf{E}$, and the susceptibility is proportional to the concentration N. For a discussion of the properties and measurement of electric susceptibility *see* PERMITTIVITY; POLARIZATION OF DIELECTRICS. Robert D. Waldron; A. Earle Bailey

Electric switch

A device that makes, breaks, or changes the course of an electric circuit. Basically, an electric switch consists of two or more contacts mounted on an

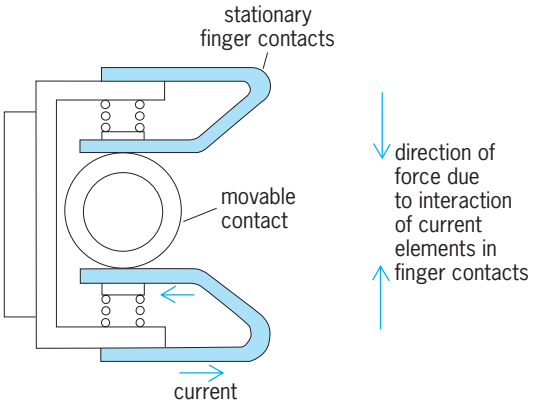


Fig. 1. With paired fingers for fixed contacts, short-circuit forces tend to hold contacts closed.

insulating structure and arranged so that they can be moved into and out of contact with each other by a suitable operating mechanism. *See* CIRCUIT (ELECTRICITY).

The term switch is usually used to denote only those devices intended to function when the circuit is energized or deenergized under normal manual operating conditions, as contrasted with circuit breakers, which have as one of their primary functions the interruption of short-circuit currents. Although there are hundreds of types of electric switches, their application can be broadly classified into two major categories: power and signal. *See* CIRCUIT BREAKER.

Power switches. In power applications, switches function to energize or deenergize an electric load. On the low end of the power scale, wall switches are used in homes and offices for turning lights on and off; dial and push-button switches control power to electric ranges, washing machines, and dishwashers. On the high end of the scale are load-break switches and disconnecting switches in power systems at the highest voltages (several hundred thousand volts).

For power applications, when closed, switches are required to carry a certain amount of continuous current without overheating, and in the open position they must provide enough insulation to isolate the circuit electrically. The latter function is particularly important in high-voltage circuits because it is the practice in the electrical industry to forbid people from working on electrical equipment unless it is isolated from the electrical supply system by a visible break in air. As an added precaution, a grounding switch is often used to connect the equipment to the ground before permitting any direct contact by a worker.

Load-break switches are required also to have the capability of interrupting the load circuit. Although this requirement is easily met in low-voltage and low-current applications, for high-voltage and high-current circuits, arc interrupters, similar to those used in circuit breakers, are needed.

In medium-voltage applications the most popular interrupter is the air magnetic type, in which the arc is driven into an arc chute by the magnetic field produced by the load current in a blowout coil. The chute may be made either of ceramic materials or of organic materials such as plexiglass. For voltages higher than 15,000 V, gaseous interrupters are sometimes used. Special gases such as sulfur hexafluoride (SF_6) have been found to have good interrupting capability for load currents without a high-speed flow. For short-circuit currents, even with SF_6, a high-speed flow is needed. Vacuum interrupters have been produced with high-voltage load-breaking duties. *See* BLOWOUT COIL; SHORT CIRCUIT.

Some load-break switches may also be required to have the capability of holding the contacts in the closed position during short-circuit conditions so that the contacts will be blown open by electromagnetic forces when the circuit breaker in the system interrupts the short-circuit current. For most contact configurations, the interaction between the short-circuit current and the magnetic field tends to force the contacts to part. Special configurations such as finger contacts are sometimes used to overcome this problem (**Fig. 1**).

Signal switches. For signal applications, switches are used to detect a specified situation that calls for some predetermined action in the electrical circuit. For example, thermostats detect termperature; when a certain limit is reached, contacts in the thermostat energize or deenergize another electrical switching device to control power flow. Centrifugal switches prevent overspeeds of motors. Limit switches prevent cranes or elevators from moving beyond preset positions. Many different types of switches perform such signaling purposes in communication systems, computers, and control systems for industrial processes. *See* ALTERNATING-CURRENT MOTOR; SWITCHING SYSTEMS (COMMUNICATIONS); THERMOSTAT.

Switches for signaling purposes are often required to have long life, high speed, and high reliability. Contaminants and dust must be prevented from interfering with the operation of the switch. For this purpose, switches are usually enclosed and are sometimes hermetically sealed.

Among the many different arrangements for contacts, the knife switch (**Fig. 2**), because of its

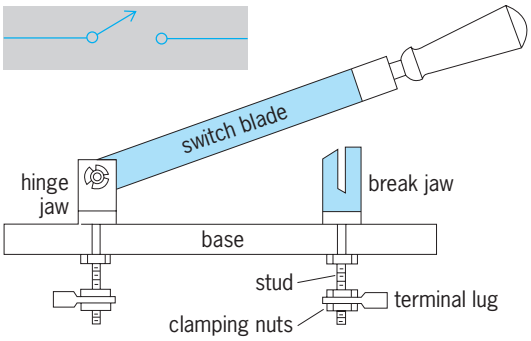


Fig. 2. Early knife switch with schematic symbol of single switch shown at top.

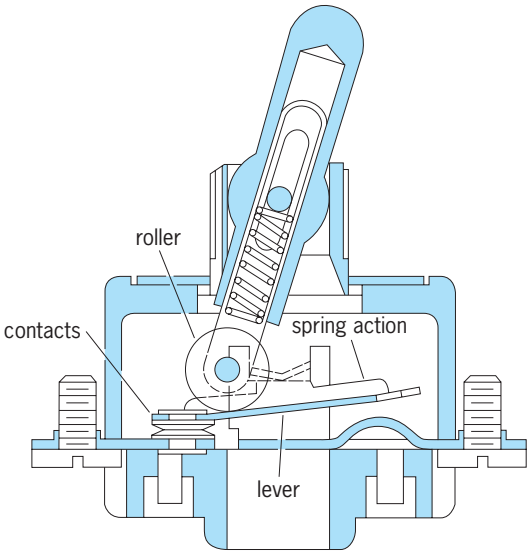


Fig. 3. Toggle mechanism actuates butt-contact switch.

early widespread use, is the basis for the elemental switch symbol. In the knife switch, a metal blade is hinged to a stationary jaw at one end and contacts a similar jaw at the other end.

In the leaf-spring switch, parallel strips of spring metal are sandwiched between blocks of insulation at one end and pushed into or out of contact at the other end. The sliding contact switch takes the form of a dial or drum with metal segments engaging contact fingers that wipe over the dial or drum surface. A butt-contact switch (**Fig. 3**) consists of one movable contact in the form of a disk or bar and one or two stationary contacts of similar form. In the mercury switch, a pool of mercury is made either to bridge the contacts or to separate from them (**Fig. 4**).

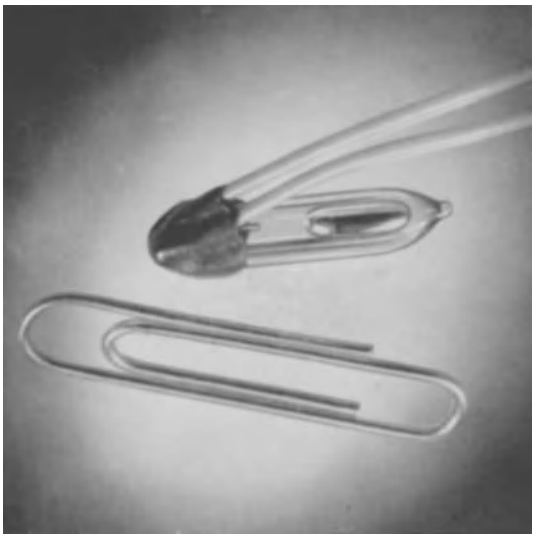


Fig. 4. Miniature mercury element may be used with a variety of tilting mechanisms to cause switch action. (*Minneapolis-Honeywell Regulator Co.*)

Switches frequently are composed of many single circuit elements, known as poles, all operated simultaneously or in a predetermined sequence by the same mechanism. Switches used in complex machines, such as computers, may have a large number of poles. Switches used in power circuits usually have from one to four poles, depending on the kind of circuit. Switches are often typed by the number of poles and referred to as single-pole or double-pole switches, and so on. It is common to express the number of possible switch positions per pole, such as a single-throw or double-throw switch.

Thomas H. Lee

Bibliography. D. G. Fink and H. W. Beaty (eds.), *Standard Handbook for Electrical Engineers*, 14th ed., 1999.

Electric transient

A temporary component of current and voltage in an electric circuit which has been disturbed. In ordinary circuit problems, a stabilized condition of the circuit is assumed and steady-state values of current and voltage are sufficient. However, it often becomes important to know what occurs during the transition period following a circuit disturbance until the steady-state condition is reached. Transients occur only in circuits containing inductance or capacitance. In general, transients accompany any change in the amount or form of energy stored in the circuit. After the following Introduction, transients in direct-current (dc) and alternating-current (ac) circuits with lumped elements are discussed. Transients of a more complex nature occur on distributed-parameter circuits, such as transmission lines. Such phenomena are elaborated upon in a separate section on transients in power systems. Finally, the modeling of transients in digital simulation is explained.

Introduction

An electric circuit or system under steady-state conditions of constant, or cyclic, applied voltages or currents is in a state of equilibrium. However, the circuit conditions of voltage, current, or frequency may change or be disturbed. Also, circuit elements may be switched in or out of the circuit. Any change of circuit condition or circuit elements causes a transient readjustment of voltages and currents from the initial state of equilibrium to the final state of equilibrium. In a sense the transient may be regarded as superimposed on the final steady state, so that Eq. (1) applies.

$$\begin{pmatrix}\text{Instantaneous}\\ \text{condition}\end{pmatrix} = \begin{pmatrix}\text{final}\\ \text{condition}\end{pmatrix} + \begin{pmatrix}\text{transient}\\ \text{terms}\end{pmatrix} \quad (1)$$

Furthermore, since the instantaneous condition at the first instant of disturbance (time zero) must be

the initial condition, it may be described by Eq. (2).

$$\begin{pmatrix}\text{Initial}\\ \text{condition}\end{pmatrix} = \begin{pmatrix}\text{final}\\ \text{condition}\end{pmatrix} + \begin{pmatrix}\text{transient terms}\\ \text{at time zero}\end{pmatrix} \quad (2)$$

A great deal of information may be obtained from these two "word equations" without recourse to mathematics. For example, if the weight on the end of a vertical spring is suddenly increased, its final displacement can be determined. Since the spring-weight combination is known to be an oscillating system, the amplitude of the transient oscillation follows from Eq. (2) as the difference between the initial and final displacements.

The nature of an electric transient is determined by three things: (1) the circuit or network itself—the interconnections of its elements and the circuit parameters (resistances R, inductances L, capacitances C, mutual inductances M); (2) the initial conditions of voltages, currents, charges, and flux linkages at the start of the transient; and (3) the nature of the disturbance which initiated the transient.

The circuit, or network, is usually defined by a diagram of connections showing all of the interconnections, junctions, meshes, circuit parameters, voltage and current sources and their polarities, and switches. Corresponding to the network a differential (or integral-differential) equation, or a set of such equations, may be written. These equations may also be written in terms of operational calculus, or as Laplace transforms, or in any other suitable mathematical equivalents. They are established in accordance with Kirchhoff's laws: (I) The sum of the currents i at a junction is equal to zero: $\Sigma i = 0$. (II) The sum of the voltages e around a mesh is equal to zero: $\Sigma e = 0$. *See* KIRCHHOFF'S LAWS OF ELECTRIC CIRCUITS.

The voltage drops associated with a resistance, inductance, mutual inductance, and capacitance, respectively, are given by Eqs. (3). In applying

$$\begin{aligned} e_R &= Ri \qquad & e_L &= L\frac{di}{dt} \\ e_i &= M_{12}\frac{di_2}{dt} & e_c &= \frac{1}{C}\int i\,dt \end{aligned} \quad (3)$$

Kirchhoff's law to a closed mesh, it is merely necessary (with due regard for signs and polarities) to equate the sum of all the voltage sources (such as generators and batteries) to the sum of all the voltage drops in the mesh, as in Eq. (4), where m refers

$$\sum_m e_m = \sum_m \sum_n \left(R_{mn} i_n + L_{mn}\frac{di_n}{dt} + \frac{1}{C_{mn}}\int i_n\,dt \right) \quad (4)$$

to any branch of the mesh in question, and n to any current causing a voltage drop in that branch (a branch may carry currents from meshes other than the mesh in question, or be mutually coupled with other branches). Equation (4) formulates the differential equation of the circuit in terms of the voltages and currents. It is sometimes convenient to make use of charges q and fluxes ϕ by the substitutions of Eqs. (5).

$$i = \frac{dq}{dt} \quad \text{or} \quad q = \int i\,dt \quad \text{and} \quad N\phi = Li \quad (5)$$

Since Eq. (4) contains an integral, it is convenient to eliminate it either by differentiating once or by substituting Eq. (5). Then Eqs. (6) follow.

$$\begin{aligned} \sum_m \frac{de_m}{dt} &= \sum_m \sum_n \left(R_{mn}\frac{di_n}{dt} + L_{mn}\frac{d^2 i_n}{dt^2} + \frac{i_n}{C_{mn}} \right) \\ \sum_m e_m &= \sum_m \sum_n \left(R_{mn}\frac{dq_n}{dt} + L_{mn}\frac{d^2 q_n}{dt^2} + \frac{q_n}{C_{mn}} \right) \end{aligned} \quad (6)$$

See COUPLED CIRCUITS.

It is customary to factor out the current from Eq. (4) and write the equation in terms of a generalized impedance operator Z_{mn} defined by Eq. (7). Since an equation of this type can be

$$\begin{aligned} \sum_m e_m &= \sum_m \sum_n \left(R_{mn} + L_{mn}\frac{d}{dt} + \frac{1}{C_{mn}}\int dt \right) i_n \\ &= \sum_m \sum_n Z_{mn} i_n \end{aligned} \quad (7)$$

written for each of the N-meshes of the network, the totality of the differential equations for the entire network constitutes a system of simultaneous differential equations in the form of Eqs. (8).

$$\begin{aligned} e_1 &= Z_{11}i_1 + Z_{12}i_2 + \cdots + Z_{1N}i_N \\ e_2 &= Z_{21}i_1 + Z_{22}i_2 + \cdots + Z_{2N}i_N \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ e_N &= Z_{N1}i_1 + Z_{N2}i_2 + \cdots + Z_{NN}i_N \end{aligned} \quad (8)$$

In general, the process of solution of such a set of equations for any current leads to a differential equation of order $2N$; there will be $2N$ integration constants associated with it, and these integration constants must be determined from the initial conditions at the first instant of the disturbance.

Equation (9) is the solution of an ordinary

$$\begin{pmatrix}\text{Complete}\\ \text{solution}\end{pmatrix} = \begin{pmatrix}\text{particular}\\ \text{integral}\end{pmatrix} + \begin{pmatrix}\text{complementary}\\ \text{solution}\end{pmatrix} \quad (9)$$

differential equation with constant coefficients and is in two parts.

The particular integral depends on the form of the applied voltage, and represents the final steady-state solution.

The complementary solution is independent of the form of the applied voltages, but depends on the initial conditions in the circuit, and represents the transient terms. Thus Eq. (9) is the mathematical equivalent of Eq. (1).

The opening of a switch may be simulated by superimposing an equal and opposite current and thereby canceling the current through the switch. Likewise, the closing of a switch may be simulated by canceling the voltage across its terminals by an equal and opposite voltage. The effect of a cancellation current (or voltage) may be superimposed on the currents and voltages already existing in the system, as in Eq. (10).

$$\begin{pmatrix}\text{Resultant}\\ \text{voltages}\\ \text{and}\\ \text{currents}\end{pmatrix} = \begin{pmatrix}\text{those which}\\ \text{would exist if}\\ \text{switches were}\\ \text{not operated}\end{pmatrix} + \begin{pmatrix}\text{those due to}\\ \text{“cancellation”}\\ \text{voltages or}\\ \text{currents}\end{pmatrix} \quad (10)$$

In many practical cases transients are one of three types:

1. Single-energy transients, in which only one form of energy storage (electromagnetic or electrostatic) is present; the transient exhibits simple exponential decay from the initial to the final conditions.
2. Double-energy transients, in which both forms of energy storage are present; the transient is either aperiodic or a damped sinusoid.
3. Combination of 1 and 2.

DC Circuit with Lumped Elements

Transients are initiated in dc circuits either when a switch is closed on a dc voltage, or when a switch is opened and the detached circuit is permitted to discharge its energy sources. The assumption is made that the suddenly applied dc voltage is sustained at a constant value for an indefinite period after the switching operation, and time is counted from the instant, $t = 0$, at which the switch is operated. It is further usually assumed, where possible, that all circuit parameters (resistances R, inductances L, and capacitances C) are constant, so that the resulting differential equations are linear.

The energy sources in dc systems comprise dynamoelectric machines, mercury-arc rectifiers, vacuum tubes, and electric batteries. Energy W is stored in a circuit in inductances ($W = 1/2Li^2$) and capacitances ($W = 1/2Ce^2$). Energy is dissipated in circuit resistances ($W = Ri^2$).

Resistance and inductance in series. A lumped parameter circuit comprising a dc voltage source E, a resistance R, and an inductance L is shown in **Fig. 1**. When the switch S is closed, the voltage E is suddenly impressed on the series circuit. By Kirchhoff's second law the differential equation is given by Eq. (11).

$$E = Ri + L\frac{di}{dt} \quad (11)$$

Equating the right-hand side of Eq. (11) to zero, the complementary solution is found by assuming that Eq. (12), in which I and a are constants to be

$$i = I\epsilon^{at} \quad (12)$$

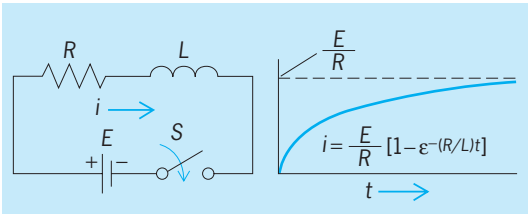


Fig. 1. Resistance and inductance in series.

determined, applies. Upon substitution in Eq. (11), Eq. (13) results, from which follows $a = -R/L$.

$$0 = RI\epsilon^{at} + aLI\epsilon^{at} = (R + aL)i \quad (13)$$

The particular integral is the final steady-state solution or $i(\infty) = E/R$. The complete solution, Eq. (14), is the sum of the complementary solution and the particular integral. But this expression contains the unknown integration constant I which must be found from the initial conditions. At $t = 0$, the instant at which the switch is closed, there can be no current in the circuit, since it is impossible to store energy in an inductance instantaneously. Therefore $i(0) = 0$ and when this is put in Eq. (14), Eq. (15) is obtained.

$$i = \frac{E}{R} + I\epsilon^{-(R/L)t} \quad (14)$$

$$i(0) = 0 = \frac{E}{R} + I \text{ or } I = -\frac{E}{R} \quad (15)$$

The complete solution by Eq. (14) therefore is given by Eq. (16), which is in the form of Eq. (1).

$$i = \frac{E}{R} - \frac{E}{R}\epsilon^{-(R/L)t} \quad (16)$$

Here E/R is the final (steady-state) condition and $-(E/R)\epsilon^{-(R/L)t}$ is the transient term. Note also that at $t = 0$ this equation gives $i(0) = 0$, agreeing with Eq. (2). The graph of the transient is shown in Fig. 1, where the current starts at zero, rises exponentially, and approaches its steady-state value of E/R as t approaches infinity. *See* ELECTRICAL RESISTANCE; INDUCTANCE.

Resistance and capacitance in series. A resistance-capacitance series circuit is shown in **Fig. 2**. A residual charge q_0 is on the capacitor just prior to closing the switch S.

Equation (17) is the differential equation of the

$$E = Ri + \frac{1}{C}\int i\,dt \quad (17)$$

circuit. This equation may be solved as it is, or

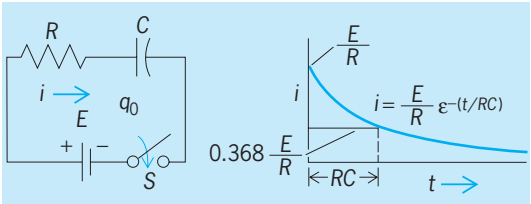


Fig. 2. Resistance and capacitance in series.

converted to a linear differential equation, Eq. (18),

$$0 = R\frac{di}{dt} + \frac{i}{C} \tag{18}$$

by differentiation or restated in terms of the instantaneous charge on the capacitor by substituting $i = dq/dt$, giving Eq. (19).

$$E = R\frac{dq}{dt} + \frac{q}{C} \tag{19}$$

The solution to this equation, following precisely the steps of the previous paragraph, is given by Eq. (20), in which Q is the integration constant. But

$$q = Q\epsilon^{-(t/RC)} + CE \tag{20}$$

initially, at $t = 0$, there was a residual charge q_0 on the capacitance; therefore the relationship given by Eq. (21) pertains, and upon the substitution of Eq. (21) in Eq. (20), Eq. (22) is obtained. This

$$q(0) = q_0 = Q + CE \text{ or } Q = q_0 - CE \tag{21}$$

$$q = CE - (CE - q_0)\epsilon^{-(t/RC)} \tag{22}$$

equation is in the form of Eq. (1), in which CE is the final (steady-state) condition and the other term is the transient. Initially, at $t = 0$, $q = q_0$ in agreement with Eq. (2). The current is given by Eq. (23).

$$\begin{aligned} i = \frac{dq}{dt} &= \frac{CE - q_0}{RC}\epsilon^{-(t/RC)} \\ &= \left(\frac{E}{R} - \frac{q_0}{RC}\right)\epsilon^{-(t/RC)} \end{aligned} \tag{23}$$

Equations (22) and (23) have been plotted in Fig. 2 for the case $q_0 = 0$. *See* CAPACITANCE.

Resistance, inductance, and capacitance in series. This circuit is shown in **Fig. 3**. The capacitor is assumed to have an initial charge q_0, or an initial voltage $V = q_0/C$. The differential equation of the circuit is given by Eq. (24).

$$E = Ri + L\frac{di}{dt} + \frac{1}{C}\int i\,dt \tag{24}$$

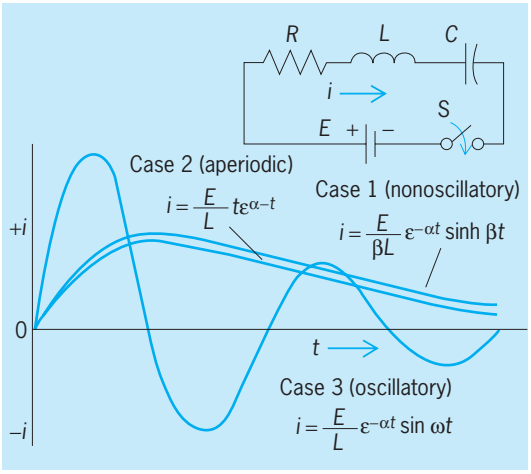


Fig. 3. Resistance, inductance, and capacitance in series for three cases.

Differentiating once to clear the integral, Eq. (25)

$$0 = R\frac{di}{dt} + L\frac{d^2i}{dt^2} + \frac{i}{C} \tag{25}$$

is obtained. Equation (25) is a second-order linear differential equation with constant coefficients; since it is equal to zero, it will possess only a complementary solution. Equation (26) is assumed, which, upon substitution in Eq. (25), yields Eq. (27).

$$i = A\epsilon^{at} \tag{26}$$

$$0 = aRA\epsilon^{at} + a^2LA\epsilon^{at} + (A/C)\epsilon^{at} \tag{27}$$

Canceling the constant A and the exponential, there results a quadratic in a whose solution gives the two possible values shown in Eqs. (28).

$$\begin{aligned} a_1 &= \frac{-RC + \sqrt{R^2C^2 - 4LC}}{2LC} \\ a_2 &= \frac{-RC - \sqrt{R^2C^2 - 4LC}}{2LC} \end{aligned} \tag{28}$$

Associating the integration constant A_1 with a_1 and A_2 with a_2, the solution takes the form of Eq. (29).

$$i = A_1\epsilon^{a_1 t} + A_2\epsilon^{a_2 t} \tag{29}$$

The voltage across the capacitor, Eq. (30), is obtained from Eq. (24).

$$\begin{aligned} e_C &= \frac{1}{C}\int i\,dt = E - Ri - L\frac{di}{dt} \\ &= E - (R + a_1L)A_1\epsilon^{a_1 t} - (R + a_2L)A_2\epsilon^{a_2 t} \end{aligned} \tag{30}$$

Initially, at $t = 0$, the current must be zero because of the inductance, and the capacitor voltage is $e_C = V$. By the first of these conditions, $i(0) = 0$, in Eq. (29) it is seen that $A_2 = -A_1$. And by the second condition, $e_C(0) = V$, in Eq. (30), Eq. (31) is obtained.

$$\begin{aligned} V &= E - (R + a_1L)A_1 - (R + a_2L)A_2 \\ &= E - (a_1 - a_2)LA_1 \\ A_1 &= \frac{E - V}{(a_1 - a_2)L} = \frac{C(E - V)}{\sqrt{R^2C^2 - 4LC}} \end{aligned} \tag{31}$$

Then Eq. (32) is the complete solution.

$$i = \frac{C(E - V)}{\sqrt{R^2C^2 - 4LC}}(\epsilon^{a_1 t} - \epsilon^{a_2 t}) \tag{32}$$

There are three special cases of this solution, depending on the nature of the radical in Eq. (32).

Nonoscillatory case. $R^2C > 4L$. In this case the radical is positive, and the exponents a_1 and a_2 are real and negative, $a_1 = -\alpha + \beta$, $a_2 = -\alpha - \beta$, where $\alpha = R/2L$ and $\beta = \sqrt{R^2C^2 - 4LC}/2LC$. Then i is given by Eq. (33), which is shown in Fig. 3, where q_0 is

$$\begin{aligned} i &= \frac{C(E - V)}{\sqrt{R^2C^2 - 4LC}}\epsilon^{-\alpha t}(\epsilon^{+\beta t} - \epsilon^{-\beta t}) \\ &= \frac{2C(E - V)}{\sqrt{R^2C^2 - 4LC}}\epsilon^{-\alpha t}\sinh \beta t \end{aligned} \tag{33}$$

assumed to be 0 and $V = 0$.

Aperiodic case. $R^2C = 4L$. In this case $a_1 = a_2 = -R/2L$, and the radical in the denominator of Eq. (32) is zero, as is the numerator. The indeterminant is easily evaluated from Eq. (33) upon letting $\beta \to 0$, thus giving Eq. (34). This is the aperiodic or

$$i = \frac{2C(E-V)}{2LC\beta} e^{-\alpha t} \sinh \beta t \mid_{\beta \to 0}$$

$$= \frac{E-V}{L} e^{-\alpha t} t \tag{34}$$

critical case, and is illustrated in Fig. 3.

Oscillatory case. $R^2C < 4L$. In this case the radical in Eq. (28) becomes imaginary and the exponents take the form of Eqs. (35), and Eq. (32) becomes Eq. (36). This is a damped oscillation and is illustrated in Fig. 3.

$$a_1 = \frac{-R}{2L} + j\frac{\sqrt{4LC - R^2C^2}}{2LC} = -\alpha + j\omega$$

$$a_2 = \frac{-R}{2L} - j\frac{\sqrt{4LC - R^2C^2}}{2LC} = -\alpha - j\omega \tag{35}$$

$$i = \frac{C(E-V)}{2j\omega LC} e^{-\alpha t}(\epsilon^{j\omega t} - \epsilon^{-j\omega t})$$

$$= \frac{E-V}{\omega L} \epsilon^{-\alpha t} \sin \omega t \tag{36}$$

AC Circuit with Lumped Elements

Alternating-current transients differ from direct-current transients in two important respects: (1) The final condition, or steady state, is an alternating or cyclic one, and (2) the amplitudes of the transient terms depend on the point on the ac applied voltage wave at which the transient is initiated, and can therefore have many different values, or even change sign.

In the section on dc transients, solutions were carried out for *RL*, *RC*, and *RLC* circuits switched onto a voltage source. In the present section, the solution will be carried out in detail for the *RLC* circuit only; the others will be regarded as special cases of this general solution by putting $C = \infty$ and $L = 0$, respectively.

Resistance, inductance, and capacitance in series. Consider the circuit of **Fig. 4** in which an *RLC* circuit is suddenly switched onto an ac voltage $e = E \sin(\omega t + \gamma)$ at an electrical angle γ displaced from $\omega t = 0$. Assume that a current I is flowing in the circuit and a voltage V is across the capacitance at the instant the switch is closed on the alternator. The voltage equation and the initial conditions then are given by Eqs. (37) and (38).

$$E \sin(\omega t + \gamma) = Ri + L\frac{di}{dt} + \frac{1}{C}\int i\,dt \tag{37}$$

$$i = I \quad \text{and} \quad e_c = V \quad \text{at} \quad t = 0 \tag{38}$$

Differentiating Eq. (37) once to clear it of the integral gives Eq. (39). The complementary solution

$$\omega E \cos(\omega t + \gamma) = L\frac{d^2i}{dt^2} + R\frac{di}{dt} + \frac{i}{C} \tag{39}$$

of this equation is the same as in the dc case of

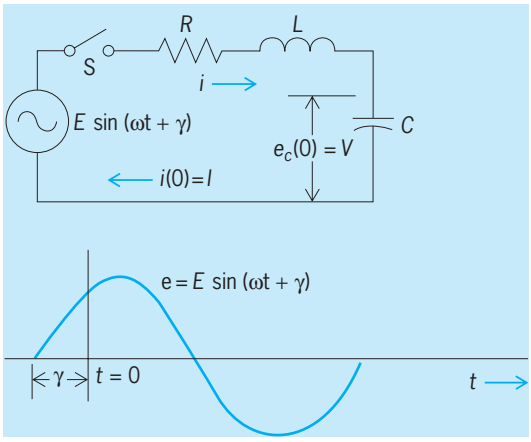


Fig. 4. *RLC* circuit with alternating current source.

Eq. (25), that is, Eq. (29), where a_1 and a_2 are given in Eq. (28).

The particular integral of Eq. (39) is its final steady-state current. This is most easily obtained as Eq. (40), the ordinary ac solution for the current, to which Eqs. (41) and (42) apply.

$$i_{ac} = \frac{E}{Z} \sin(\omega t + \gamma - \theta) \tag{40}$$

$$Z = \sqrt{R^2 + (\omega L - 1/\omega C)^2} \tag{41}$$

$$\tan\theta = \frac{\omega L - 1/\omega C}{R} = \frac{\omega^2 LC - 1}{\omega CR} \tag{42}$$

The complete solution then is given by Eq. (43),

$$i = \frac{E}{Z} \sin(\omega t + \gamma - \theta) + A_1\epsilon^{a_1 t} + A_2\epsilon^{a_2 t} \tag{43}$$

in which A_1 and A_2 are integration constants and a_1 and a_2 are given in Eq. (28).

The capacitor voltage is given by Eq. (44).

$$e_c = \frac{1}{C}\int i\,dt = -\frac{E}{\omega CZ} \cos(\omega t + \gamma - \theta)$$

$$+ \frac{A_1}{Ca_1}\epsilon^{a_1 t} + \frac{A_2}{Ca_2}\epsilon^{a_2 t} \tag{44}$$

Now applying the initial conditions of Eq. (38) at $t = 0$, there results from Eqs. (43) and (44), respectively, Eqs. (45) and (46).

$$i(0) = I = \frac{E}{Z} \sin(\gamma - \theta) + A_1 + A_2 \tag{45}$$

$$e_c(0) = V$$

$$= \frac{-E}{\omega CZ} \cos(\gamma - \theta) + \frac{A_1}{Ca_1} + \frac{A_2}{Ca_2} \tag{46}$$

Solving Eqs. (45) and (46) simultaneously for A_1 and A_2 and using Eq. (28), Eqs. (47)-(49) result.

$$A_1 = \frac{1}{\sqrt{R^2C^2 - 4LC}} \left\{ \frac{I}{a_2} - CV \right.$$

$$\left. - \frac{E}{Z}\left[\frac{1}{\omega} \cos(\gamma - \theta) + \frac{1}{a_2} \sin(\gamma - \theta) \right] \right\} \tag{47}$$

$$A_2 = \frac{-1}{\sqrt{R^2C^2 - 4LC}} \left\{ \frac{I}{a_1} - CV - \frac{E}{Z}\left[\frac{1}{\omega}\cos(\gamma - \theta) + \frac{1}{a_1}\sin(\gamma - \theta)\right]\right\} \quad (48)$$

$$a_1 = \frac{-R}{2L} + \sqrt{\left(\frac{R}{2L}\right)^2 - \frac{1}{LC}}$$
$$a_2 = \frac{-R}{2L} - \sqrt{\left(\frac{R}{2L}\right)^2 - \frac{1}{LC}} \quad (49)$$

These values, together with Eqs. (41) and (42), inserted in Eqs. (43) and (44) give the solution in a form suitable for the critically damped case. However, if $R^2C < 4L$, then the radicals in Eqs. (47)–(49) become imaginary and these expressions become thecomplex numbers given in Eqs. (50), whereupon

$$A_1 = {}^1/_2(M + jN) \qquad A_2 = {}^1/_2(M - jN)$$
$$a_1 = -\alpha + j\omega \qquad a_2 = -\alpha - j\omega \quad (50)$$

Eq. (43) takes the damped oscillatory form given in Eq. (51).

$$i = \frac{E}{Z}\sin(\omega t + \gamma - \theta) + \epsilon^{-\alpha t}(M\cos\omega t - N\sin\omega t) \quad (51)$$

It is evident from Eq. (43) or Eq. (51) that the final steady-state value, after the transient has died out, is the alternating current of Eq. (40). It is also clear from Eqs. (47) and (48) that the amplitudes of the transient terms depend upon the angle γ on the ac applied voltage wave at $t = 0$.

Resistance and inductance in series. This may be regarded as a special case of Eq. (43), in which $C = \infty$ (the capacitance short-circuited) and $V = 0$. Consequently, the solution is given by Eq. (52). Thus

$$i = \frac{E}{Z}\sin(\omega t + \gamma - \theta) + \left[I - \frac{E}{Z}\sin(\gamma - \theta)\right]\epsilon^{-(R/L)t} \quad (52)$$

the transient starting from an initial value I decays exponentially to its final ac steady-state value.

There will be no transient if the switch is closed at the angle γ on the voltage wave such that Eq. (53) applies.

$$I = \frac{E}{Z}\sin(\gamma - \theta) \quad (53)$$

Resistance and capacitance in series. This may be regarded as a special case of Eq. (43), in which $L = 0$ and $I = 0$. Under these conditions the solution is given by Eq. (54) and the capacitor voltage is,

$$i = \frac{E}{Z}\sin(\omega t + \gamma - \theta) - \left[\frac{V}{R} + \frac{E}{\omega CRZ}\cos(\gamma - \theta)\right]\epsilon^{-t/RC} \quad (54)$$

by Eq. (44), defined by Eq. (55). Thus the transient,

$$e_c = \frac{-E}{\omega CZ}\cos(\omega t + \gamma - \theta) + \left[V + \frac{E}{\omega CZ}\cos(\gamma - \theta)\right]\epsilon^{-t/RC} \quad (55)$$

starting with a voltage V, decays exponentially to its final ac steady-state value. If the switch is closed at an angle γ on the voltage wave such that Eq. (56)

$$V = -\frac{E}{\omega CZ}\cos(\gamma - \theta) \quad (56)$$

applies, there will be no transient. Loyal V. Bewley

Power Systems with Lumped and Distributed Elements

In an electric power system, a transient is a temporary disturbance leading to changes in the amounts of energy stored in the capacitances and inductances, causing variations of power system voltage and current waveforms. Such disturbances may also be referred to as electromagnetic transients to indicate that oscillatory variations are caused through the alternating conversion of electric energy of capacitors into magnetic energy of inductors. The frequency range typically extends from several hertz to several megahertz.

Classification. A classification of electric transients by cause is given in **Fig. 5**. Lightning strikes induce currents that cause electric transients in the form of traveling waves to propagate in forward and backward directions on lines.

Equipment switching events change the network topology, that is, the network configuration, and force redistribution of stored energy. Switching can also result in traveling waves on lines or cause energization and deenergization of lumped capacitances and inductances.

Conceptual analysis. The following study of the transient recovery voltage (TRV) serves as introduction to conceptual analysis of electric transients involving both harmonic oscillations and traveling waves. **Figure 6*a*** shows a line-to-ground short circuit on a transmission line. The fault current is supplied via a transformer and needs to be interrupted to clear the fault. The opening of a circuit breaker triggers electric transients due to a change in the

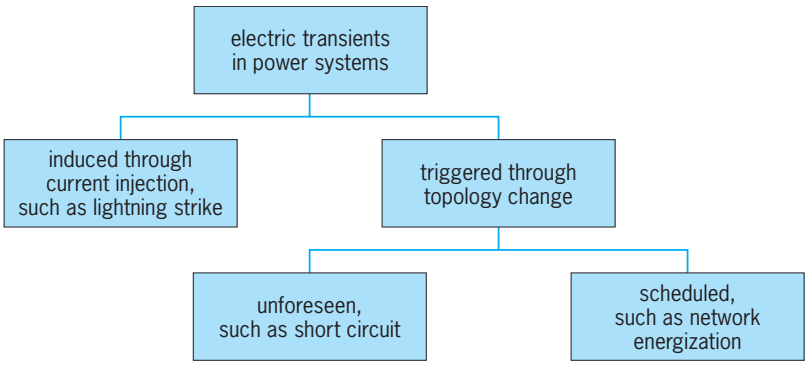


Fig. 5. Causes of electric transients in power systems.

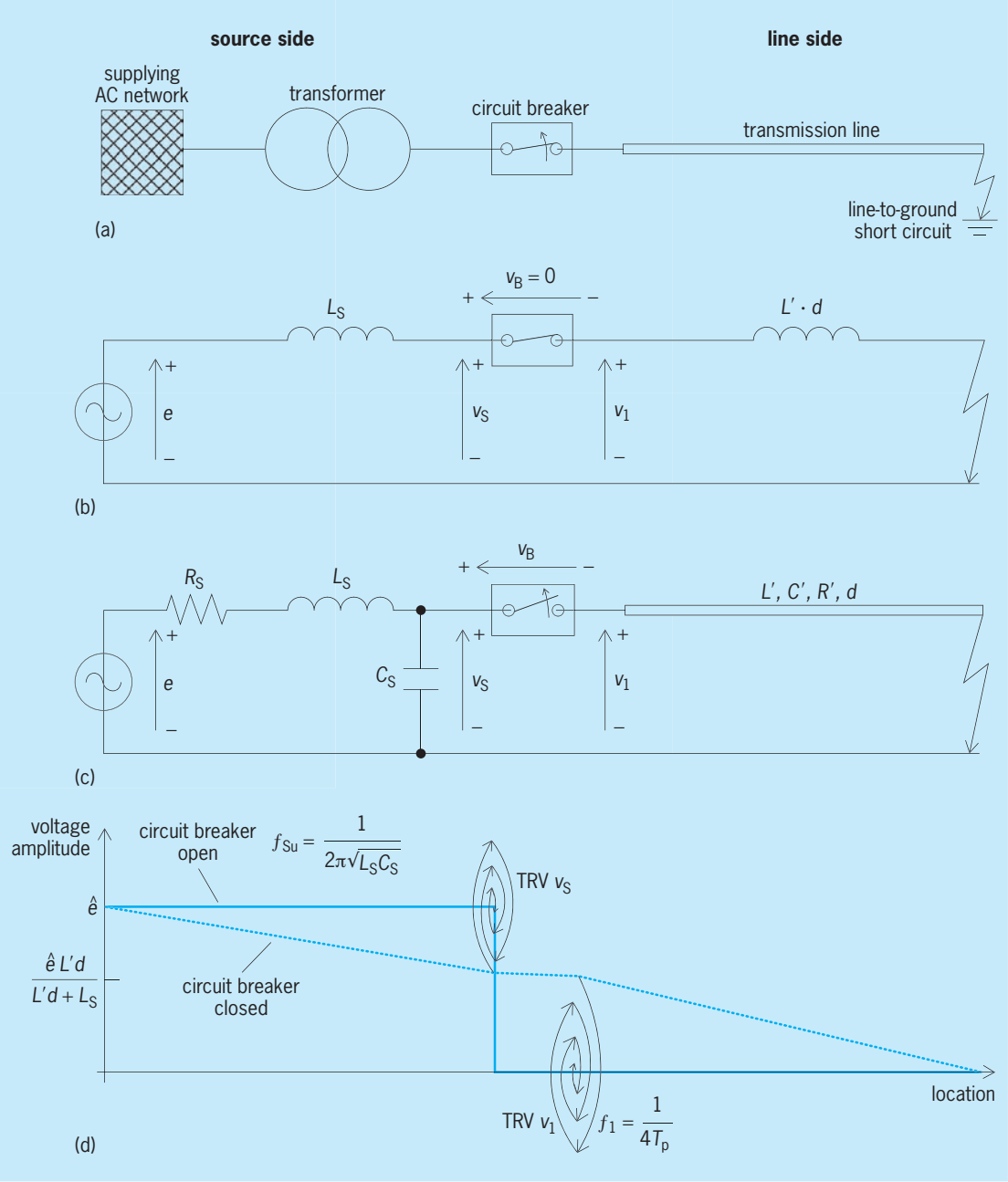


Fig. 6. Analysis of electric transients involving both harmonic oscillation and traveling waves. (*a*) Configuration of power system for study of transient recovery voltage (TRV). (*b*) Model for evaluating steady state before breaker opening. (*c*) Model for evaluating electric transients upon breaker opening. (*d*) Qualitative assessment of transients described by voltages.

topology of the system. In order to focus on the behavior of transients, only single-pole networks will be considered. (Multipole networks can be represented as single-pole equivalents.) *See* CIRCUIT BREAKER; SHORT CIRCUIT; TRANSMISSION LINES.

Before the circuit breaker opens, the voltage v_B across it is zero, as indicated in Fig. 6*b*, and hence $v_S = v_1$. The circuit of Fig. 6*b* models the configuration of Fig. 6*a* at source frequency f_e, which is 50 or 60 Hz in utility power systems depending on regional preferences. The purpose of this model is to estimate the voltage profile while the fault is still in place and the circuit breaker is closed. This allows the determination of the steady-state voltages giving initial conditions from which the transients will start. The line has inductance per unit length L', and the distance to the fault is d. The source side is modeled through an equivalent ac voltage source e behind a lumped inductive element L_S, which represents inductances of the transformer and the supplying ac network. Since the circuit breaker is to open at the instant when the fault current crosses zero to reduce stress, the voltages will reach their peak values at this instant. This is due to the inductive behavior of the line and the resulting phase shift of 90° between the fault current and the voltages v_S and v_1. The voltage profile with the circuit breaker closed is shown by a line in Fig. 6*d*, which gives the

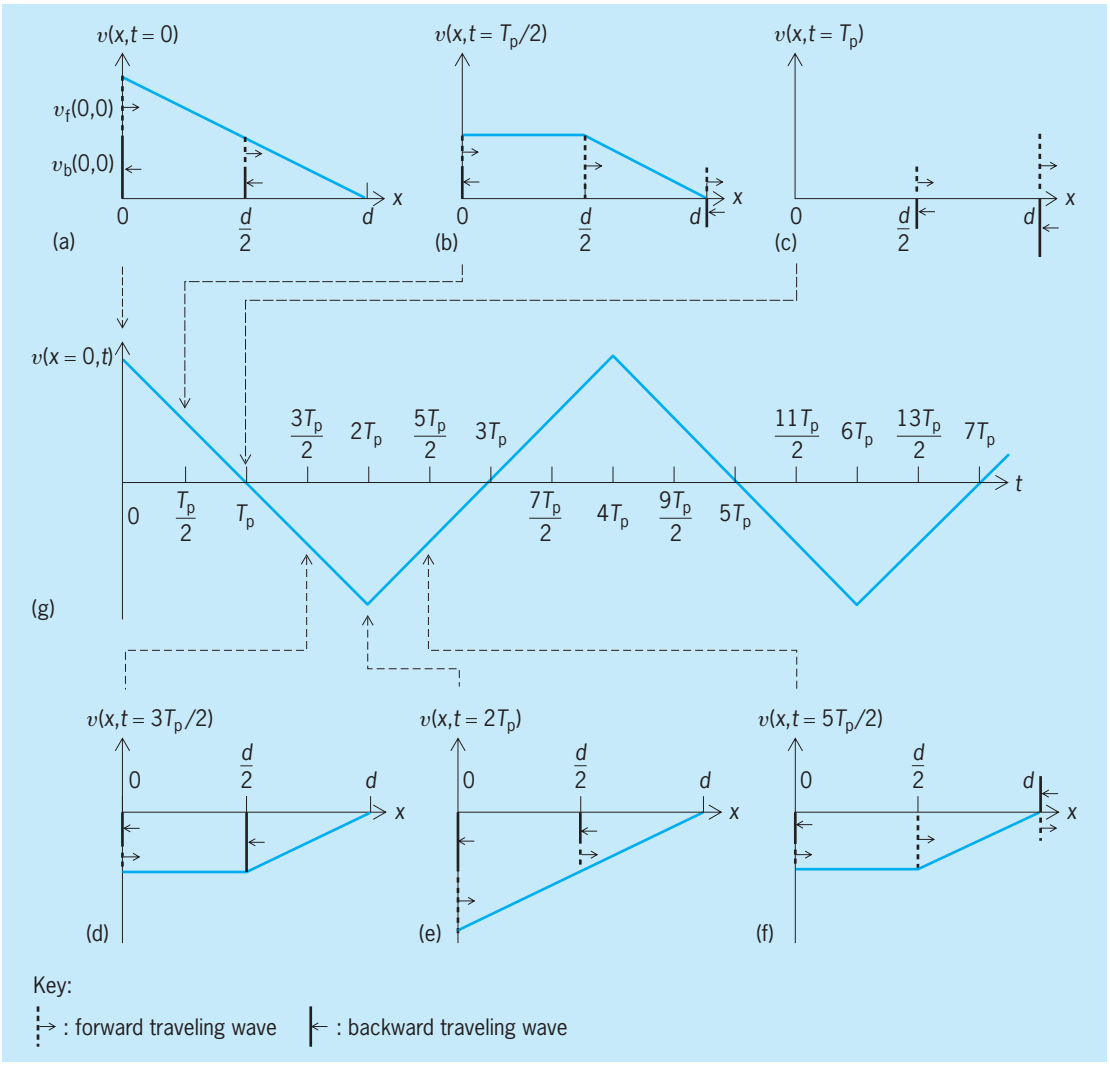


Fig. 7. Graphical interpretation of electromagnetic transients on transmission line. (*a–f*) Voltage profiles along line. (*g*) Derived voltage at beginning of line as a function of time.

amplitude as a function of location. The amplitude of the voltage source is denoted by $\hat{e}$. At the location of the short circuit, the voltage is zero. The amplitudes of v_S and v_l are obtained with the voltage divider expression given on the ordinate. Figure 6*d* also shows the steady-state voltage profile that is reached when the circuit breaker is open and the fault current is zero.

To assess the electric transients between the two steady states, the steady-state model of Fig. 6*b* is not adequate. To represent the redistribution of energy, both inductors and capacitors need to be modeled as indicated in Fig. 6*c*. The lumped capacitive element C_S represents the capacitance of transformer and bus bars on the source side, while C' gives the capacitance per unit length of the line. Losses are accounted for through R_S and R' on the source and line sides, respectively.

As the circuit breaker opens, the source and line sides become decoupled. The TRV v_S must first increase, and describes a damped harmonic oscillation around the new steady state (Fig. 6*d*). If R_S were neglected, the undamped frequency of this harmonic oscillation on the source side would be given by Eq. (57), but because of the presence of R_S the observed value is slightly lower. The exact trajectory of the source-side TRV can be calculated with the techniques developed for circuits with lumped parameters. At the entrance to the line, the TRV v_l will first decrease.

$$f_{Su} = 1/(2\pi\sqrt{L_S C_S}) \qquad (57)$$

The waveform of the TRV v_l is influenced by traveling waves which are investigated by means of **Fig. 7** for the case of a lossless line. In Fig. 7*a* the initial voltage profile along the line is redrawn from Fig. 6*d*. The entrance to the line is at $x = 0$, and the circuit breaker is opened at $t = 0$. The waves propagate in all possible directions. Since the line is one-dimensional, there exist a forward-traveling wave marked with subscript f and arrow in direction of increasing x, and a backward-traveling wave marked with subscript b and arrow in direction of decreasing x. At any point on the line, the voltage is given as the sum of forward and backward traveling waves.

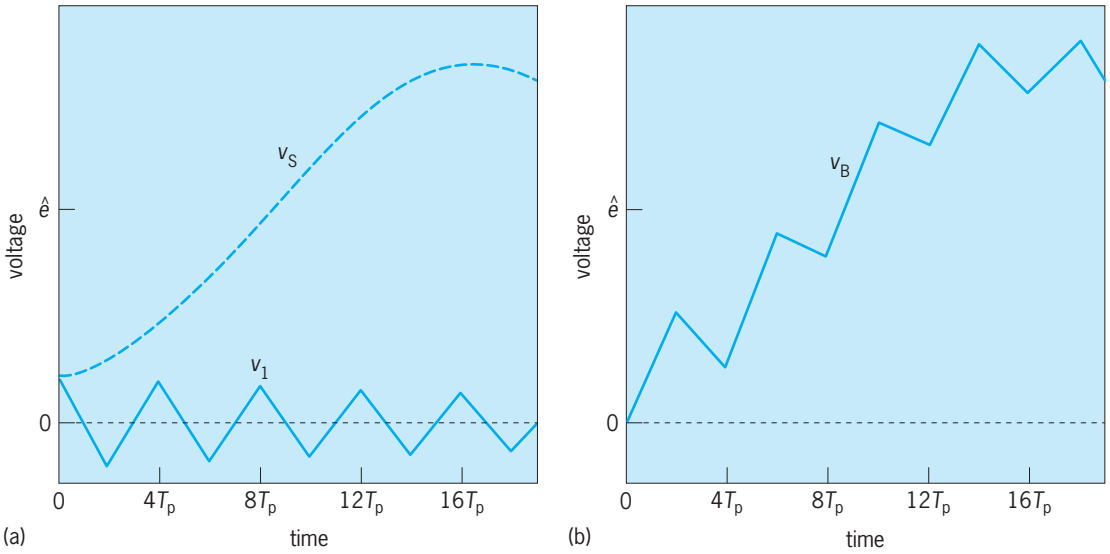


Fig. 8. Transient recovery voltages (TRVs). (*a*) TRVs on source side (v_S) and line side (v_l). (*b*) TRV across circuit breaker (v_B).

The propagation time that the waves need to travel from one end of the line to the other end is denoted by T_p.

Figure 7*b* illustrates how the original voltage profile of $t = 0$ has changed as $t = T_p/2$ is reached. The forward traveling wave of $x = 0$ at $t = 0$ traveled to $x = d/2$. The forward traveling wave of $x = d/2$ at $t = 0$ has moved by the same amount from $x = d/2$ to $x = d$. Likewise, the backward traveling wave of $x = d/2$ at $t = 0$ moved in opposite direction to reach $x = 0$. The waves are reflected at both ends. Since there is a short circuit at $x = d$, the voltage at $x = d$ must be zero at any time, that is, $v(d,t) = 0$, and therefore $v_f(d,t) = -v_b(d,t)$. At $x = 0$, there is an open circuit where the voltage reaches a maximum and $v_f(d,t) = v_b(d,t)$. Therefore, at the entrance of the line $v(0,t) = 2\, v_b(0,t)$.

Figure 7*c* to 7*f* shows the further evolution of the voltage profile in increments of $T_p/2$. Figure 7*g* shows the voltage $v_l = v(0,t)$ at the entrance to the line as a function of time derived from the changing voltage profiles. The voltage is of sawtooth shape with a period of $4T_p$, that is, the frequency of oscillation of v_l, is $f_l = 1/(4T_p)$. On overhead lines, where the insulating material is air, the waves propagate at the speed of light $c_0 = 3 \times 10^8$ m/s (1.68×10^5 mi/s). At a distance to the fault of $d = 1$ km (0.62 mi), this leads to $f_l = c_0/(4\mathrm{d}) = 75$ kHz. Such a fault is also known as a kilometric fault or short-line fault. For longer distances, the frequency is lower, but the voltage peak is higher as shown by the voltage divider in Fig. 6*d*. Thus, the stress on the circuit breaker due to the rate of voltage rise, which is directly related to the frequency, decreases, while the stress due to the reached voltage amplitude increases. For shorter distances, the opposite is the case. In the case of the kilometric fault, the combination of the stress due to rate of rise and amplitude of the voltage has been found to be particularly severe. Appropriate sizing of power system equipment depends on knowledge of the properties of electric transients.

In **Fig. 8***a*, the TRV waveforms on source and line sides are drawn in the same diagram. The source-side TRV (v_S) describes a harmonic oscillation in the form of a damped sinusoid which, as indicated in Fig. 6*d*, starts from its lowest point given by the voltage divider. In the steady state, after the transient has died out, v_S approximates the ac voltage source e. Since the frequency of the source-side TRV is much lower than the frequency of the line-side TRV, only the start of the source-side TRV is shown. On the line side, damping is taken into account (which is not done in Fig. 7), as evidenced by the decaying peak values of the sawtooth. The resulting TRV $v_B = v_S - v_l$ across the breaker is shown in Fig. 8*b*.

Digital Simulation

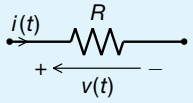
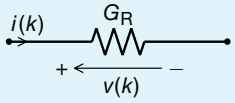
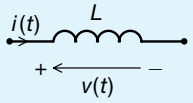
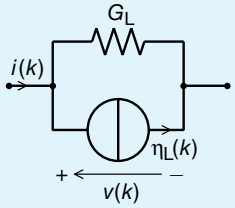
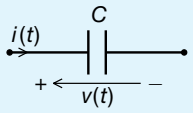
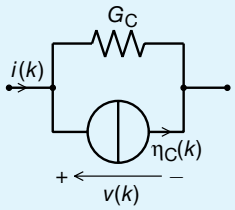
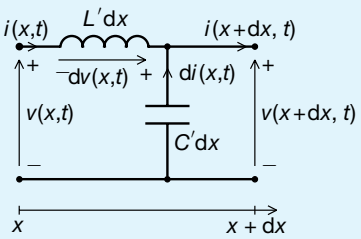
The conceptual analysis of the preceding section allows understanding of causes and consequences of electric transients. To the study specific cases and in particular those that involve complex circuits and systems, the use of digital simulation is required. Among the popular simulators are the Electromagnetic Transients Program (EMTP) and others of its type. Key to the success of EMTP is the companion modeling of network elements, where the differential equations characterizing a network element are approximated by a companion model consisting of a resistive circuit. A library of companion element models can so be established as shown in the **table** and derived hereafter.

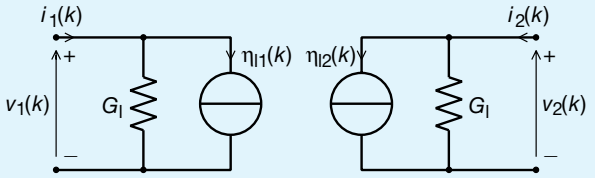
Companion modeling for lumped-parameter elements. In the continuous time domain, the lumped inductance is described through differential equation (58).

$$v(t) = L\frac{\mathrm{d}i(t)}{\mathrm{d}t} \tag{58}$$

For digital simulation, Eq. (58) must be discretized. Using the trapezoidal integration yields

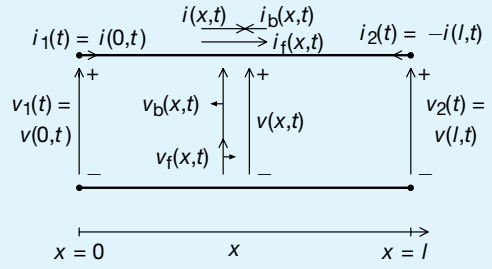
Library of companion element models

Element type	Representation in continuous-time domain	Companion element model in discrete-time domain
Lumped resistance	 $v(t) = i(t)R$	 $i(k) = G_R v(k)$ $G_R = 1/R$
Lumped inductance	 $v(t) = L\frac{di(t)}{dt}$	 $i(k) = G_L v(k) + \eta_L(k)$ $G_L = \frac{\tau}{2L}$ $\eta_L(k) = i(k-1) + G_L v(k-1)$
Lumped capacitance	 $i(t) = C\frac{dv(t)}{dt}$	 $i(k) = G_C v(k) + \eta_C(k)$ $G_C = \frac{2C}{\tau}$ $\eta_C(k) = -i(k-1) - G_C v(k-1)$
Distributed parameter, lossless line		$\frac{\partial v(x,t)}{\partial x} = -L'\frac{\partial i(x,t)}{\partial t}$ $\frac{\partial i(x,t)}{\partial x} = -C'\frac{\partial v(x,t)}{\partial t}$



$i_1(k) = G_I v_1(k) + \eta_{I1}(k)$ $\quad$ $i_2(k) = G_I v_2(k) + \eta_{I2}(k)$

$\eta_{I1}(k) = -i_2(k-\kappa) - G_I v_2(k-\kappa)$ $\quad$ $\eta_{I2}(k) = -i_1(k-\kappa) - G_I v_1(k-\kappa)$



$$G_I = \sqrt{C'/L'}$$

$$\tau = \frac{T_p}{\tau}$$

$$T_p = l\sqrt{L'C'}$$

Eq. (59), where τ is the time step size separating successive discrete time points at which solutions are calculated and k is the time step counter. The higher the frequencies of the electric transients to be studied, the smaller must be the time step size to allow for accurate tracking of the waveforms. Rearranging Eq. (59) yields the companion model of Eq. (60) [also shown in the table], where G_L and $\eta_L(k)$ are given by Eqs. (61) and (62).

$$\tfrac{1}{2}[v(k) + v(k-1)] = L\frac{[i(k) - i(k-1)]}{\tau} \tag{59}$$

$$i(k) = G_L v(k) + \eta_L(k) \tag{60}$$

$$G_L = \frac{\tau}{2L} \tag{61}$$

$$\eta_L(k) = i(k-1) + \frac{\tau}{2L}v(k-1) \tag{62}$$

The companion model simulates the inductance through a conductance and a current source, which provides the initialization for the solution at time step k based on the results of prior time step $k-1$. The same procedure is applied to derive the companion model for the capacitance. In the case of the resistance, no discretization is necessary.

Companion modeling for distributed-parameter elements. A lossless transmission line of length l can be

considered as being composed of an infinite number of incremental segments with incremental length dx for which Eqs. (63) and (64) apply, as indicated in the table.

$$\frac{\partial v(x,t)}{\partial x} = -L' \frac{\partial i(x,t)}{\partial t} \tag{63}$$

$$\frac{\partial i(x,t)}{\partial x} = -C' \frac{\partial v(x,t)}{\partial t} \tag{64}$$

As already established, the voltage at any point is given by the sum of forward and backward traveling voltage waves: $v(x,t) = v_f(x - ct) + v_b(x + ct)$ with c, given by Eq. (65), being equal to the speed of light

$$c = 1/\sqrt{L'C'} \tag{65}$$

in the case of overhead lines. A forward traveling current wave $i_f(x - ct) = v_f(x - ct)/Z_l$ is associated with $v_f(x,t)$, and a backward traveling current wave $i_b(x + ct) = v_b(x + ct)/Z_l$ is associated with $v_b(x,t)$, where the characteristic impedance Z_l, given by Eq. (66),

$$Z_l = \sqrt{L'/C'} \tag{66}$$

relates the associated waves. Using the sign conventions given in the table, the voltages and currents at each point of the line are given by Eqs. (67) and (68).

$$v(x,t) = Z_l i_f(x - ct) + Z_l i_b(x + ct) \tag{67}$$

$$i(x,t) = i_f(x - ct) - i_b(x + ct) \tag{68}$$

From these two equations follow Eqs. (69) and (70).

$$i_f(x - ct) = \frac{1}{2}\left(i(x,t) + \frac{1}{Z_l} v(x,t)\right) \tag{69}$$

$$i_b(x - ct) = \frac{1}{2}\left(-i(x,t) + \frac{1}{Z_l} v(x,t)\right) \tag{70}$$

The argument $(x - ct)$ of i_f is zero for $x = 0$, $t = 0$ and $x = d$, $t = T_p$. Furthermore, at terminal 1, $x = 0$, $i_1(t) = i(0,t)$, and at terminal 2, $x = l$, $i_2(t) = -i(l,t)$. Therefore, from Eqs. (69) and (70) follows Eq. (71), which can be rearranged to Eq. (72),

$$i_1(t - T_p) + \frac{1}{Z_1} v(t - T_p) = -i_2(t) + \frac{1}{Z_1} v_2(t) \tag{71}$$

$$i_2(t) = \frac{1}{Z_1} v_2(t) - \left(i_1(t - T_p) + \frac{1}{Z_1} v_1(t - T_p)\right) \tag{72}$$

Similarly, Eq. (73) holds.

$$i_1(t) = \frac{1}{Z_1} v_1(t) - \left(i_2(t - T_p) + \frac{1}{Z_1} v_2(t - T_p)\right) \tag{73}$$

Discretization readily leads to the companion element model of the lossless line shown in the table.

Companion modeling for the network. The companion element models are connected to the overall companion network model in accordance with the topology of the power system under study. For the companion network model, node equations are formulated to obtain the nodal equation system. *See* NETWORK THEORY.

Solution procedure. Once the nodal equation system is known, the time step counter k is set to zero and a simulation loop is entered. The solution to the nodal equation system is calculated, and the quantities of interest are provided as output. Then, the time step counter is incremented, the loop is reentered, and the operations are repeated. The simulation stops once a termination condition is met. Simulation results can be displayed graphically, as shown in Fig. 8. *See* NUMERICAL ANALYSIS; SIMULATION. Kai Strunz

Bibliography. J. B. Aidala and L. Katz, *Transients in Electric Circuits*, Prentice Hall, 1980; H. W. Dommel, *EMTP Theory Book*, Microtran Power System Analysis Corp., Vancouver, 1992; A. P. S. Meliopoulos, *Power System Grounding and Transients*, Marcel Dekker, New York, 1988; K. Strunz, *Numerical Methods for Real Time Simulation of Electromagnetics in AC/DC Network Systems*, VDI-Verlag, Düsseldorf, 2002; L. van der Sluis, *Transients in Power Systems*, Wiley, Chichester, 2001; N. Watson and J. Arrillaga, *Power Systems Electromagnetic Transients Simulation*, IEE, London, 2003.

Electric vehicle

A ground vehicle propelled by a motor that is powered by electrical energy from rechargeable batteries or other source onboard the vehicle, or from an external source in, on, or above the roadway. Examples are the golf cart, industrial truck and tractor, automobile, delivery van and other on-highway truck, and trolley bus. In common usage, electric vehicle refers to an automotive vehicle in which the propulsion system converts electrical energy stored chemically in a battery into mechanical energy to move the vehicle. This is classed as a battery-only-powered electric vehicle. The other major class is the hybrid-electric vehicle, which has more than one power source. *See* AUTOMOBILE; BUS; TRUCK.

History. Construction of the first electric vehicle is credited to the French inventor and electrical engineer M. Gustave Trouvé, who demonstrated a motorized tricycle powered by lead-acid batteries in 1881. In the United States, Andrew L. Riker is credited with building the first electric vehicle (also a tricycle) in 1890, and by 1891 William Morrison had built the first electric four-wheeler. In France in 1899, a four-wheel electric vehicle driven by Camille Jenatzy became the first car to break 60 mi/h (96 km/h). By then, production of battery-powered vehicles for use as personal transportation, commercial trucks, and buses had already begun.

Electric vehicles, with their instant starting, quiet running, and ease of operation, peaked in their challenge to steam- and gasoline-powered cars in 1912. The limited performance, range, and speed of electric vehicles, plus the need for frequent battery charging, restricted their usefulness and dampened their popularity. By the 1920s, the piston-type internal combustion engine had prevailed as the dominant automotive powerplant. Most production and development work on electric vehicles ended during

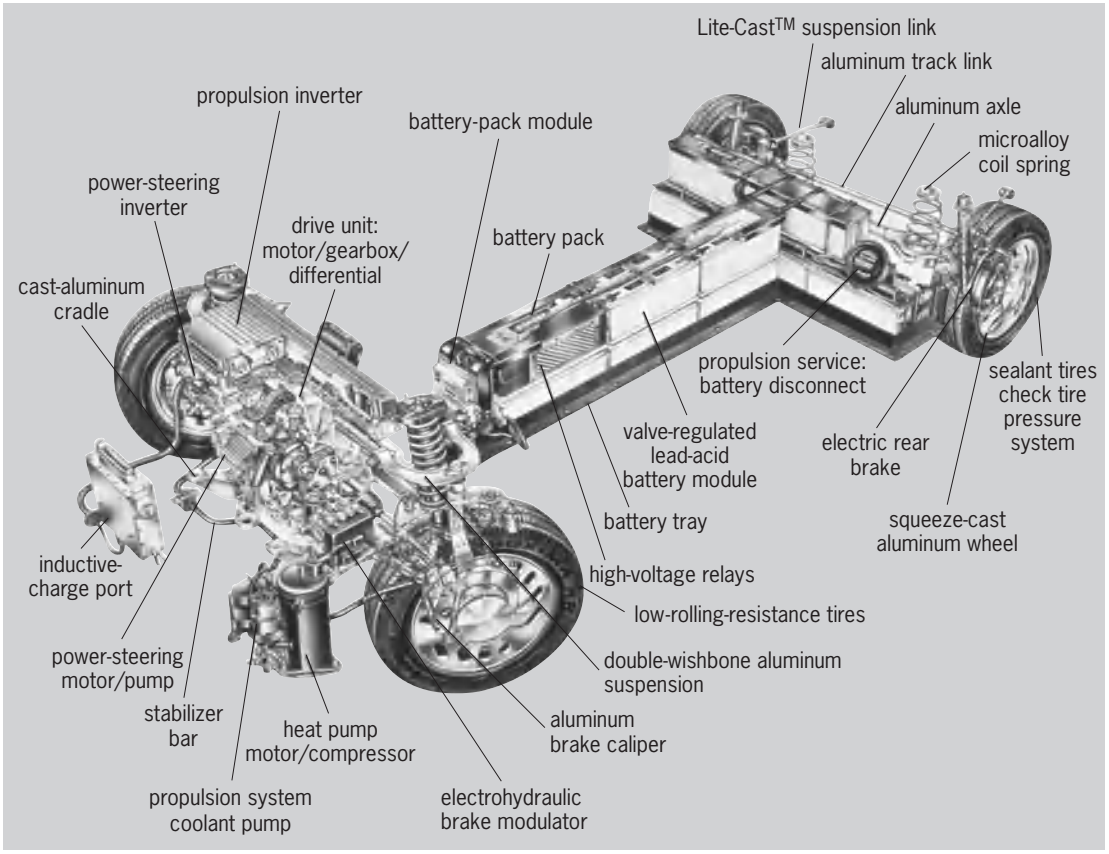


Fig. 1. EV1, a two-seat electric vehicle powered by lead-acid batteries. (*General Motors Corp.*)

the 1930s. *See* AUTOMOTIVE ENGINE; ENGINE; INTERNAL COMBUSTION ENGINE.

In the 1960s, interest revived in electric vehicles as a result of concern with diminishing petroleum reserves, rising cost of crude oil production, and air pollution from the automotive engine that burned gasoline which was refined from crude oil. Over the years, a few electric vehicles had been constructed, usually by converting small light cars and trucks into electric vehicles by removing the engine and fuel tank and installing an electric motor, controls, and batteries. However, during that time no major automotive manufacturer brought out an electric vehicle. *See* AIR POLLUTION; GASOLINE; PETROLEUM.

The Clean Air Act of 1963 and its amendments established limits on emissions from new vehicles sold in the United States. In 1990, the California Air Resources Board decided to further reduce air pollution by mandating (but later rescinding) that 2% of each automaker's sales must have zero emissions in the 1998 model year. This demand for a zero-emission vehicle (ZEV) could be met only by the electric vehicle, which typically was powered by lead-acid batteries. Used in an electric vehicle, lead-acid batteries have two major weaknesses: relatively high weight for the amount of energy stored, and reduced capacity in cold weather.

To help develop a better battery for electric vehicles, the U.S. Advanced Battery Consortium was formed in 1991. The purpose of this partnership among United States automakers and the electric utility industry was to develop advanced batteries capable of providing future generations of electric vehicles with significantly increased range and performance.

In 1996 General Motors began limited marketing of the electric vehicle EV1 (**Fig. 1**). The EV1 was the first specifically designed electric car produced by a major automaker since before World War II. Other automakers also have developed and tested electric vehicles and vehicle conversions powered by lead-acid or advanced batteries.

Battery-only power. The General Motors EV1 is a battery-only-powered vehicle containing 26 lead-acid batteries which are assembled into a T-shaped pack that provides a nominal voltage of 312 volts (**Fig. 2**). These batteries differ from 12-V automotive batteries primarily in duty cycle. In the electric vehicle, the batteries must supply all the energy needs. Except for regenerative braking, there is no onboard charging. The batteries provide the power to propel the vehicle, and to power the lights and all accessories such as air conditioning and radio. As a result, electric-vehicle batteries go through much deeper discharge cycles than an automotive battery, which seldom is discharged more than 5% of its rated capacity. *See* BATTERY.

Because of the increased duty cycle, electric-vehicle batteries can deliver 85% of their charge without damaging the batteries or shortening their useful life. Ideally, this could provide the EV1 and similar electric vehicles with a useful range of 70 mi

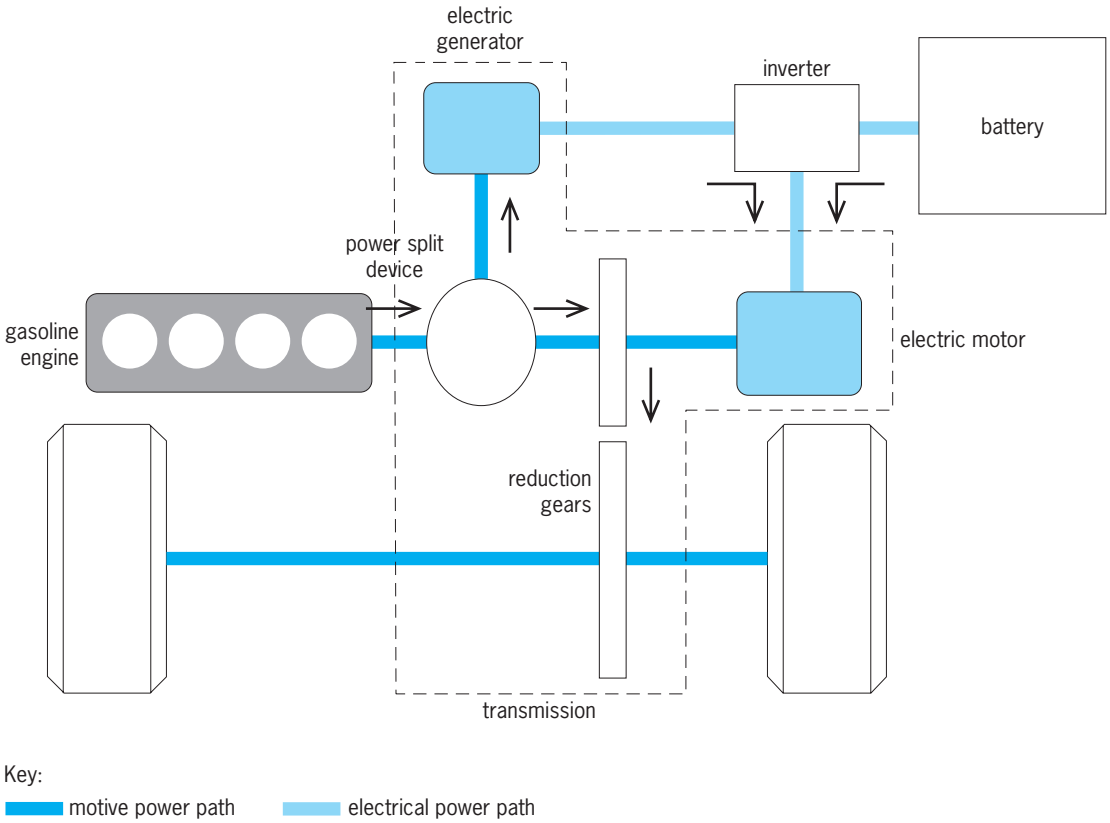


Fig. 2. Power flows through a hybrid vehicle which uses a gasoline engine to propel the vehicle and drive a generator that can operate the electric motor directly or charge the batteries, as necessary. (*Toyota Motor Sales, U.S.A., Inc.*)

(113 km) of city driving or 90 mi (145 km) of highway driving. A top speed of 80 mi/h (130 km/h) may be possible, but at a sacrifice in range, which also is shortened by hilly terrain and use of any electrical equipment on the vehicle.

In addition to lead-acid batteries, other batteries are used in electric vehicles. These include the newer nickel-metal hydride battery and the lithium-ion battery.

Hybrid power. A hybrid electric vehicle has more than one source of power. These sources can be different types of energy storage devices, power converters, and inverters. The first hybrid vehicle is credited to an Italian, Count Felix Carli. In 1894, Carli constructed an electric-powered tricycle that had a system of rubber springs which could release a short burst of additional power when needed.

Although power losses occur each time that energy is converted from one form to another, hybrid drive can be more efficient than a conventional automotive engine. However, having two or more power sources can increase the complexity, cost, and weight of a hybrid vehicle, as well as its manufacturing, safety, emissions, maintenance, and service problems. *See* AUTOMOTIVE ENGINE.

Internal combustion engine. Since the 1890s, major hybrid-vehicle research and development work has focused on adding a small internal combustion engine to an electric vehicle. Typically, the battery-powered motor drives the wheels, while the engine, usually running at constant speed, drives a generator that charges the batteries. Operating the engine at constant speed reduces fuel consumption and produces cleaner exhaust gas than if the engine were larger, operating at variable speed, and providing the sole source of vehicle power.

In some hybrid vehicles, engine power is split (**Fig. 2**). Part of the engine power propels the vehicle, while part drives the generator which can operate the motor directly or charge the batteries, as necessary. Reportedly, engine exhaust emissions of hydrocarbons, carbon monoxide, and oxides of nitrogen are about 10% that of a conventional gasoline-engine vehicle. In addition, fuel efficiency is doubled.

Fuel cell. In this electrochemical device, the reaction between a fuel, such as hydrogen, and an oxidant, such as oxygen or air, converts the chemical energy of the fuel directly into electrical energy. The fuel cell is not a battery and does not store energy, although the fuel cell also has two electrodes separated by an electrolyte. As long as fuel is supplied to one electrode of the fuel cell and oxygen or air to the other, a voltage is produced between the electrodes. When an external circuit connects the electrodes, electrons will flow through the external circuit. Since fuel-cell voltage is less than 1 V, stacks of fuel cells are connected together to provide the needed electrical energy.

When fuel cells are the primary power source in a

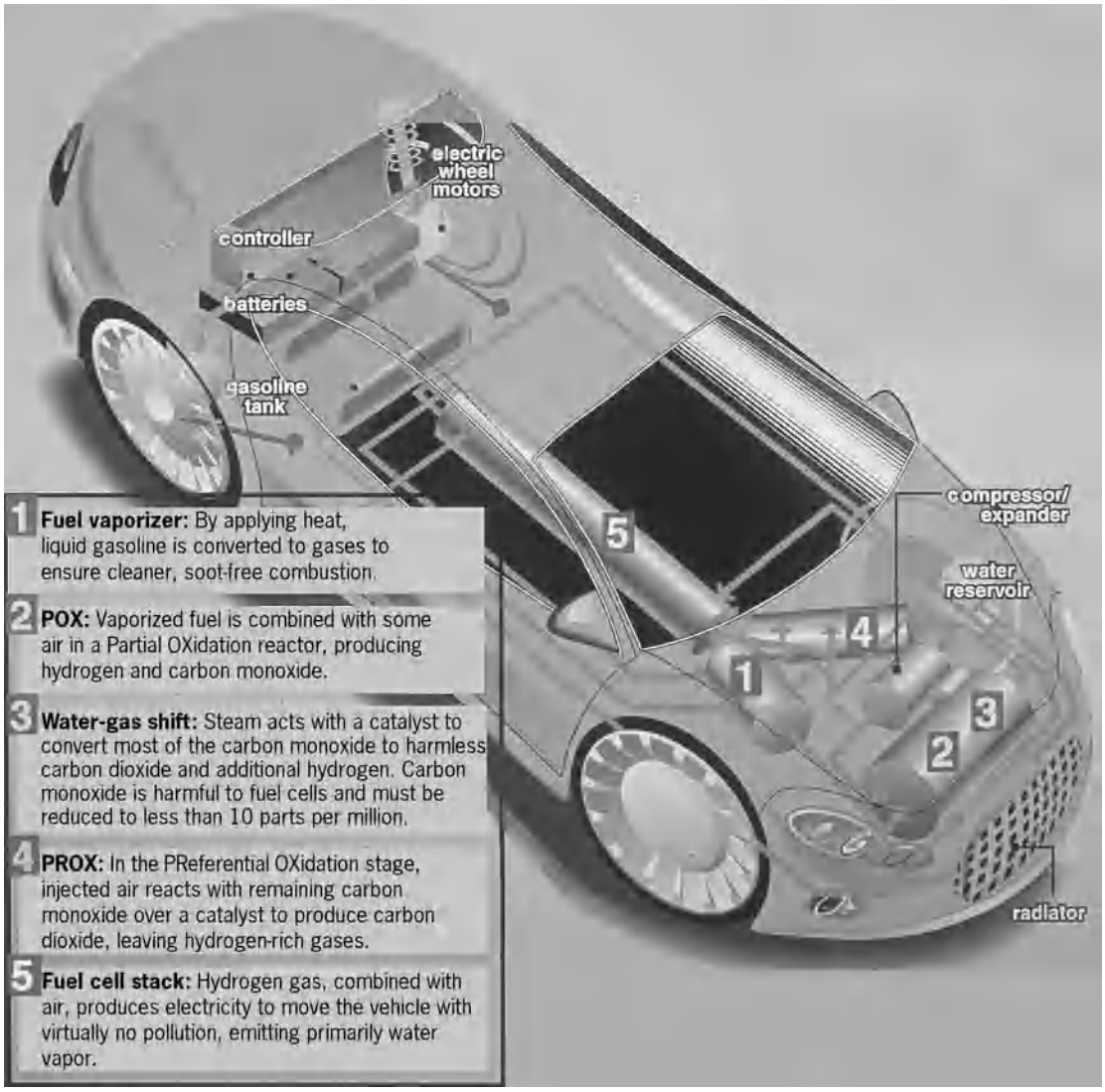


Fig. 3. Layout and five-step process that produces electricity from gasoline, which powers the fuel cells. (*Chrysler Corp.*)

hybrid vehicle, batteries provide secondary power. Fuel cells do not provide immediate output during a cold start. Until the fuel cells reach operating temperature, which may take about 5 min, a battery pack supplies the power for initial startup and vehicle movement. *See* FUEL CELL; HYDROGEN.

Three types of fuel cell under development for electric vehicles are the hydrogen-fueled, the methanol-fueled, and the gasoline-fueled. (1) A hydrogen-fueled cell runs on hydrogen gas stored in pressure tanks carried by the vehicle. Range of the vehicle is determined by the amount of compressed hydrogen that the tanks can hold. When hydrogen is used as the fuel, fuel cell operation produces no significant amounts of unwanted emissions. Water vapor and electricity are the only products. However, widespread use of hydrogen as a near-term vehicle fuel is unlikely because there exists no infrastructure of hydrogen refueling stations which are in place and accessible to the public. (2) To provide the fuel cell with hydrogen gas while avoiding the problems of hydrogen refueling, a methanol-to-hydrogen reformer onboard the vehicle can produce hydrogen gas from liquid methanol. Installed on production models, this would allow motorists to refuel vehicles in the conventional manner at existing service stations through any pump which dispensed methanol. However, use of a reformer to obtain hydrogen lowers vehicle efficiency and creates some emissions of carbon dioxide. (3) A gasoline-to-hydrogen reformer on the vehicle can extract hydrogen gas from gasoline. The hydrogen is then delivered to the fuel cell stack (**Fig. 3**). Use of gasoline could move fuel cell technology years closer to production in automotive vehicles, while reportedly improving fuel efficiency by 50% and emissions by 90%. *See* METHANOL.

Donald L. Anglin

Bibliography. Robert Bosch GmbH, *Automotive Handbook*, 6th ed., 2005; Society of Automotive Engineers, *SAE Handbook*, 3 vols., annually; M. H. Westbrook, *The Electric Car: Development and Future of Battery, Hybrid and Fuel-Cell Cars*, 2001.

Electrical breakdown

A large, usually abrupt rise in electric current in the presence of a small increase in electric voltage. Breakdown may be intentional and controlled or it may be accidental. Lightning is the most familiar example of breakdown.

Breakdown in gases and solids. In a gas, such as the atmosphere, the potential gradient may become high enough to accelerate the naturally present ions to velocities that cause further ionization upon collision with atoms. If the region of ionization does not extend between oppositely charged electrodes, the process is corona discharge. If the region of ionization bridges the gap between electrodes, thereby breaking down the insulation provided by the gas, the process is ionization discharge. When controlled by the ballast of a fluorescent lamp, for example, the process converts electric power to light. In a gas tube the process provides controlled rectification. *See* ARC DISCHARGE; BREAKDOWN POTENTIAL; CORONA DISCHARGE; ELECTRIC SPARK; ELECTRICAL CONDUCTION IN GASES; FLUORESCENT LAMP; GLOW DISCHARGE; IONIZATION.

In a solid, such as an insulator, when the electric field gradient exceeds 10^6 V/cm, valence bonds between atoms are ruptured and current flows. Such a disruptive current heats the solid abruptly; the rate of local temperature rise may fracture the insulator, the high temperature may carbonize or otherwise decompose the insulation, or as occasionally happens when lightning strikes a tree, the heat may ignite the insulator. *See* ELECTRIC INSULATOR; ELECTRICAL INSULATION.

In a semiconductor if the applied backward or reverse potential across a junction reaches a critical level, current increases rapidly with further rise in voltage. This avalanche characteristic is used for voltage regulation in the Zener diode. In a transistor the breakdown sets limits to the maximum instantaneous voltage that can safely be applied between collector and emitter. When the internal space charge extends from collector junction through the base to the emitter junction, even a voltage below the avalanche level can produce a short circuit, in which case the phenomenon is termed punch-through. *See* SEMICONDUCTOR; TRANSISTOR; ZENER DIODE.

Frank H. Rockett

Laser-induced breakdown in gases. For sufficiently high intensities, laser radiation can ionize (or breakdown) the gas atoms or molecules through which it is propagating. The laser radiation intensity required to produce ionization is dependent on the wavelength of the light and on the duration of the laser pulses as well as the gas species and pressure. In general, the breakdown by laser radiation can occur by two different mechanisms, a cascade- or collision-induced ionization, or a multiple-photon absorption process. The most often encountered mechanism is the cascade process, but multiple-photon ionization can be important under certain conditions.

Multiple-photon ionization. Multiple-photon ionization occurs when an atom absorbs directly from the laser radiation field a sufficient number of photons to ionize the atom. The number of photons required is simply the ionization potential of the atom divided by the energy of a photon. For example, the photon energy of the ruby laser at 0.69-micrometer wavelength is 1.78 eV. Therefore, the multiple-photon ionization of argon having an ionization potential of 15.8 eV requires the simultaneous absorption of nine photons. Because of the nature of this process, the ionization is strongly dependent on the laser intensity and wavelength. Experimental measurements of multiple-photon ionization have shown that the threshold for air at pressures below 100 torr (2 psi or 13 kilopascals) is 10^{18} W/m^2 (10^{17} W/ft^2) for a 50-picosecond-duration-pulse, 0.69-μm-wavelength laser. Multiple-photon ionization can be important for laser-induced breakdown at short wavelengths, short pulse durations, and low gas pressures. *See* IONIZATION POTENTIAL; PHOTON.

Cascade breakdown. For the cascade breakdown process, a free electron is accelerated in the radiation field of the laser, but only gains an increment of energy upon collision with atoms. This energy absorption continues until the electron has sufficient energy to ionize an atom by an inelastic collision, producing a second electron. The two electrons then gain additional energy from the field, producing four electrons and so forth. Under breakdown conditions this process continues until the gas is fully ionized. An electron gaining energy from the laser radiation field can also lose energy by processes such as electronic excitation of atoms, or electrons can be lost from the breakdown process by diffusion out of the laser beam or by recombination or by attachment to form negative ions. The rate of ionization necessary to overcome these loss processes then defines the breakdown threshold intensity.

Breakdown threshold. Since high-intensity lasers are usually pulsed devices, an important parameter affecting the breakdown threshold is the duration of the laser pulse. If the laser pulse is long compared to the characteristic electron-loss time, breakdown occurs when the rate of electron production just exceeds the losses; breakdown is determined by the laser intensity and is independent of the laser pulse duration. If the laser pulse is short relative to the characteristic loss times, breakdown occurs when there is sufficient fluence (energy per unit area) in the pulse to generate full ionization of the gas. For air at atmospheric pressure and large-diameter beams (such that electron diffusion is not important), the threshold is fluence-dependent for pulses less than approximately 10^{-7} s and is intensity-dependent for pulses having a longer duration. As an example, for air at atmospheric pressure and laser beams larger in diameter than 4×10^{-3} in. (10^{-2} cm), the threshold fluence for the 10.6-μm carbon dioxide (CO_2) laser wavelength is 10^5 joules/m^2 (10^4 joules/ft^2) for pulse durations less than 10^{-7} s, and for longer pulses the intensity threshold is 3×10^{13} W/m^2 (3×10^{12} W/ft^2). These theoretically predicted values have been experimentally verified.

Conditions for occurrence. Two important factors affecting the occurrence of breakdown are the source of the initiating electron and the presence of aerosol particles. The electron required for initiating the cascade process must be provided by some mechanism other than the cascade, such as photoionization from natural causes or a spark or discharge. For visible-wavelength laser-induced breakdown, the initial electron can be the result of multiple-photon absorption ionization of impurities, but for long-wavelength radiation, certainly 10.6 μm, an external source must provide it. Particulate matter, such as naturally occurring aerosols in the atmosphere, can have a dramatic effect on the breakdown threshold of gases. For particles of 0.1 μm and larger, the interaction of the laser radiation with the particle vaporizes the particle, ionizes the vapor, and subsequently ionizes the surrounding gas. The threshold for this process can be several orders of magnitude lower than the threshold for gas breakdown in the absence of aerosols. *See* ATMOSPHERIC CHEMISTRY.

Pressure dependence. The laser-induced gas breakdown has an interesting pressure dependence. The rate of laser energy absorption by free electrons increases with pressure because of the increase in the electron-atom collision frequency. This energy absorption increase continues as gas pressure increases until the collision frequency equals the oscillatory frequency of the laser radiation field. For pressures above this level, the rate of energy absorption decreases with increasing pressure, and the breakdown approaches that induced by a constant electric field in the high pressure limit. Therefore, there is a minimum in a plot of the threshold as a function of pressure that occurs when the collision frequency equals the oscillatory frequency. As an example, for atmospheric air and 10.6-μm-wavelength radiation this minimum occurs at approximately 44 atm (650 lb/in.^{2} or 4.5 megapascals).

For gas pressures lower than the minimum, the gas breakdown threshold varies as the inverse square of the wavelength; that is, the shorter the wavelength, the more difficult it is to break down the gas. This behavior has been verified for wavelengths from 0.69 to 10.6 μm. The threshold for gas breakdown at pressures above the minimum is independent of the laser wavelength.

Applications. Laser-induced gas breakdown is important as a limitation for propagating high-intensity laser radiation, and as a source for the production of a high-density, high-temperature plasma for a number of applications. Gas breakdown represents the first step in research leading to the study of laser-generated controlled fusion. The plasma produced by laser-induced breakdown can be used as a convenient source for soft x-rays and an intense source of ultraviolet or visible radiation for a fast-rise-time light source. *See* LASER; NUCLEAR FUSION; PLASMA (PHYSICS).

David C. Smith

Bibliography. G. Bekef, *Principles of Laser Plasmas*, 1976; L. A. Dossado and C. J. Fothergill (eds.), *Electrical Breakdown and Degradation in Polymers*, 1992; E. E. Kunhardt and L. H. Luessen (eds.), *Electrical Breakdown and Discharges in Gases*, 1983; Y. P. Raizer, *Laser-Induced Discharge Phenomena*, 1977; Y. P. Raizer and J. E. Allen, *Gas Discharge Physics*, 1997 reprint.

Electrical codes

Systematic bodies of rules governing the practical application, installation, and interconnection of electrically operated equipment, devices, and electrical wiring systems.

National Electrical Code. The basic code used throughout the United States, and a number of other countries throughout the world, for indoor and outdoor electrical installations to supply lighting, motors, appliances, and machines is the National Electrical Code, prepared under the direction of the National Fire Protection Association (NFPA). It is approved by the American National Standards Institute, and the 2005 edition of the code is known as NFPA 70-2005 (ANSI).

The National Electrical Code was originally drawn in 1897 as a result of the united efforts of various insurance, electrical, architectural, and allied interests. The original code was prepared by the National Conference on Standard Electrical Rules, composed of delegates from various interested national associations.

In 1911 the National Conference of Standard Electrical Rules was disbanded, and since then the NFPA has acted as the code's sponsor. Beginning with the 1920 edition, the National Electrical Code has been under the further auspices of the American National Standards Institute with the NFPA continuing in its role as administrative sponsor. Since that date, the NFPA committee that produces the code has been identified as ANSI Standards Committee C1 (formally USAS C1 or ASA C1).

The provisions of the National Electrical Code are under constant review by a number of panels whose members are selected to provide broad representation of electrical, industrial, and public interests. The code is amended in its periodic republication every 3 years or by tentative interim amendments (TIAs) which are announced by bulletins and through the technical press.

The National Electrical Code is purely advisory as far as the National Fire Protection Association is concerned, but it is very widely used for legal regulatory purposes. The code is administered by various local inspection agencies, whose decisions govern its application to individual installations. Local inspectors are largely members of the International Association of Electrical Inspectors. This organization, the National Electrical Manufacturers Association, the National Electrical Contractors Association, the Institute of Electrical and Electronic Engineers, the Edison Electric Institute, the Underwriters' Laboratories, Inc., the International Brotherhood of Electrical Workers, governmental groups, and independent experts all contribute to the

development and application of the National Electrical Code.

Compliance with the provisions of the code can effectively minimize fire and accident hazards in any electrical design. It sets forth requirements that constitute a minimum standard for the framework of electrical design. As stated in its introduction, the code is concerned with the "practical safeguarding of persons and property from hazards arising from the use of electricity." The National Electrical Code is recognized as a legal criterion of safe electrical design and installation in many jurisdictions throughout the country and the world. In some large municipalities within the United States, the National Electrical Code either is accepted with certain modifications or serves as the basis for a local electrical code. Where accepted by the local governing body, the National Electrical Code is used in court litigation and by insurance companies as a basis for insuring buildings.

Other standards. In addition to the National Electrical Code volume itself, other standards and recommended practices are made available in pamphlet form by the National Fire Protection Association, which oversees the promulgation and periodic revision of over 270 standards. These cover such special subjects as hospital operating rooms, municipal fire alarm systems, garages, aircraft hangars, and other equipment with great potential hazards due to improper design.

The National Electrical Safety Code (to be distinguished from the National Electrical Code) is published by the Institute of Electrical and Electronic Engineers, Inc. Designed as ANSI C2, this code presents basic provisions for safeguarding persons from hazards arising from the installation, operation, or maintenance of (1) conductors and equipment in electric supply stations, and (2) overhead and underground electric supply and communications lines. Basically, this code applies to the outdoor circuits of electric utility companies and to similar systems or equipment on commercial and industrial premises.

Municipal codes. The National Electrical Code is incorporated bodily or by reference in many municipal building ordinances, often with additional provisions or restrictions applicable in the particular locality. Some large cities have independent electrical codes; however, the actual provisions in most such codes tend to be basically similar to the National Electrical Code.

Testing of electrical products. Standards on the construction and assembly of many types of electrical equipment, materials, and appliances are set forth in literature issued by the National Electrical Manufacturers Association (NEMA) in conjunction with Underwriters Laboratories, Inc. (UL). The Underwriters Laboratories examines, tests, and determines the suitability of materials and equipment to be used according to code regulations. Each year, it publishes three volumes listing commercially available electrical products which have been found acceptable with reference to fire and accident hazards and which conform with the application and installation requirements of the code. The three volumes are titled *Electrical Construction Materials Directory, Electrical Appliance and Utilization Equipment Directory*, and *Hazardous Location Equipment Directory*. The Underwriters' Laboratories publishes other literature such as *Gas and Oil Equipment Directory* and *Fire Protection Equipment Directory*, dealing with special equipment involving hazard to life or property.

Testing of electrical equipment and products is permitted to be performed by any "nationally recognized testing laboratory," as stipulated by the Occupational Health and Safety Administration (OSHA). Underwriters Laboratories is one such company. Tests performed by other such companies are conducted in accordance with the applicable NEMA/UL standard.

Administration. Electrical codes are administered locally by inspectors. The inspectors, who may review plans and specifications, make periodic site visits during the construction process to visually examine electrical work during installation and after the work is completed to ensure compliance with applicable rules or ordinances.

Electrical inspection bureaus are maintained in many cities by the city itself. In other cities, a third-party inspection agency, such as the New York Board of Fire Underwriters, which provides electrical inspectors for any jurisdiction within New York State that wishes to avail itself of the Board's services, is used. Typically where third-party inspectors are used, the companies recognized to perform inspections within a given jurisdiction are identified by the local building department. In communities where codes are enforced by ordinance, inspections may be performed by municipal electrical inspectors. Utility inspectors examine the service entrance and metering installation for compliance with prevailing utility regulations.

Federal and state buildings are usually inspected by authorized government electrical inspectors. In these instances inspection includes both safety consideration and the requirements of the particular job specifications. Other specification (by underwriters or municipal inspectors) is often waived. *See* ELECTRONIC EQUIPMENT GROUNDING; GROUNDING; WIRING.

J. F. McPartland; Brian J. McPartland

Bibliography. J. F. McPartland, *Handbook of Practical Electrical Design*, 3d ed., 1999; J. F. McPartland, *National Electrical Code Handbook*, 25th ed., 2005; *National Electrical Safety Code*, NFPA 70-2005 (ANSI), 2005.

Electrical communications

That branch of electrical engineering dealing with the transmission and reception of information. Information can be transmitted over many different types of pathways, such as satellite channels, underwater acoustic channels, telephone cables, and

fiber-optic links. Characteristically, any communications link is noisy. The receiver never receives the information-bearing waveform as it was originally transmitted. Rather, what is received is, at best, the sum of what was transmitted and noise. In reality, what is more likely to be received is a distorted version of what was transmitted, with noise and perhaps interference. Consequently, the design and implementation of a communications link are dependent upon statistical signal-processing techniques in order to provide the most efficient extraction of the desired information from the received waveform. *See* COMMUNICATIONS CABLE; COMMUNICATIONS SATELLITE; DISTORTION (ELECTRONIC CIRCUITS); ELECTRICAL INTERFERENCE; ELECTRICAL NOISE; MICROWAVE; OPTICAL COMMUNICATIONS; RADIO-WAVE PROPAGATION; TELEPHONE SERVICE; TRANSMISSION LINES; UNDERWATER SOUND; WAVEGUIDE.

Broadly speaking, there are two basic classes of communication waveforms, those involving analog modulation and those involving digital modulation. The former type implies that the modulation process allows the actual information signal to modulate a high-frequency carrier for efficient transmission over a channel; this is achieved by using the continuum of amplitude values of an analog waveform. Examples of analog modulation systems include amplitude-modulation (AM) and frequency-modulation (FM) systems, as well as a variety of others such as single-sideband (SSB), double-sideband (DSB), and vestigial-sideband (VSB) systems. In digital modulation systems, the initial information-bearing waveform, assuming it is in analog form (such as voice or video), is first sampled, then quantized, and finally encoded in a digital format for carrier modulation and transmission over the channel. *See* MODULATION.

Analog modulation. Analog communication systems are often characterized as linear or nonlinear modulation systems. Linear systems include AM and single-, double-, and vestigal-sideband systems, whereas nonlinear systems include FM and phase-modulation (PM). Among the linear analog systems, AM is probably the most widely used, because its receiver structure can be implemented in the form of a simple envelope detector. However, AM is very inefficient from a power point of view, because its signal structure results in part of its total transmitted power being wasted in an unmodulated carrier instead of all its power going to the information-bearing portion of the waveform. *See* AMPLITUDE MODULATION; FREQUENCY MODULATION; PHASE MODULATION; SINGLE SIDEBAND.

This problem is remedied with either single- or double-sideband systems, but their receivers are more complex than the AM receiver. Furthermore, the single-sideband system, which has the most complex transmitter, requires only half the bandwidth of the other two, thereby giving it yet another advantage over AM. The vestigal-sideband system is essentially a compromise between the double sideband and the single sideband, for although implementation with it is somewhat easier than with a single sideband, its bandwidth is greater than that of the single sideband but less than that of the double sideband.

The disadvantage of any linear analog modulation system is that its ultimate performance, as measured by the ratio of the signal power of the final demodulated waveform to the noise power of the demodulated waveform, called the signal-to-noise ratio (SNR), is limited by this ratio at the input to the receiver. In nonlinear analog communications, this is not necessarily the case. There the output signal-to-noise ratio can be made significantly higher than that at the input, and consequently significantly higher than the output signal-to-noise ratio of a linear modulation scheme. However, the price for this improvement in performance is increased transmission bandwidth. *See* SIGNAL-TO-NOISE RATIO.

Digital modulation. In digital communications, it is necessary to convert the information into a digital format. At times, the information might originate in an appropriate digital format, as in the transfer of computer files from one computer to another (so-called computer communications). However, in many other instances, the information originates in analog form. This is the case for both voice communications and video transmission. Therefore, some means of digitizing the analog waveform is required.

The classical technique for accomplishing this is probably the use of pulse-code modulation (PCM). There the analog signal is first sampled as a minimum at what is known as the Nyquist frequency (that is, at a rate corresponding to twice the bandwidth of the waveform). The samples are quantized and encoded into a binary word. The digits of this word are used to modulate a carrier for transmission over the communications channel. *See* PULSE MODULATION.

Pulse-code modulation is one example of the general technique of source encoding. Other schemes attempt to remove as much redundancy as possible from the message in order to minimize the number of binary digits needed to reproduce it. Examples of practical source-coding schemes are delta modulation and differential pulse-code modulation. In both of these techniques, an estimate of the current sample is made, the error between the sample and its estimate is ascertained, and this error signal is encoded into binary digits to be modulated on the carrier.

Once the waveform is in digital format, many modulation techniques can be used. These consist of both binary schemes, whereby one of two possible signals is transmitted across the channel, and M-ary schemes with $M > 2$, whereby one of M signals is used to convey the information in any given signaling interval. Typically, $M = 2^n$, for some integer n, and if all M signals are equally likely (that is, the a priori probability of transmitting any one of them is $1/M$), then each transmitted symbol conveys n bits of information. Thus, a binary system conveys one bit of information each time a symbol is transmitted, a quarternary system conveys two bits of information each time a symbol is sent, and so on.

For binary systems, the most widely used modulations are phase-shift keying and frequency-shift

keying. In the former scheme, a radio-frequency carrier has its phase changed to 0 and π radians each time a binary 1 or a binary 0 is to be transmitted, respectively. In frequency-shift keying, the information is transmitted by sending one of two possible frequencies during each signaling interval.

Many modulations can be used with *M*-ary schemes. Perhaps conceptually the simplest are the*M*-ary versions of binary phase-shift or frequency-shift keying, for example, the transmission of a carrier which uses one of *M* possible phases during each signaling interval. However, other schemes make use of a combination of both amplitude and phase modulation; these techniques are termed quadrature amplitude modulation (QAM).

Communication theory. The theoretical foundations of digital communications stem from the results of statistical decision theory and information theory. By using concepts developed in the former discipline, it is possible, in principle, to design an optimal receiver for a given transmitted waveform and a given noise environment. For example, in the detection of one of *M* known signals, all of which have equal energy and equal probability to be transmitted a priori, in a noise environment that contains only radio-frequency receiver noise, one form of an optimum receiver consists of a parallel filter bank of *M* matched filters with one filter matched to each of the *M* possible transmitted waveforms. A matched filter has an impulse response which is the time inverse of the signal to which it is matched. *See* DECISION THEORY; ELECTRIC FILTER.

Information theory has as its key results the ultimate limitations and performance capabilities of any communication system. The most important of these is embodied in the so-called noisy channel-coding theorem, which states that, under a fairly broad set of conditions, it is possible to achieve perfectly reliable communication over a noisy (unreliable) channel. This is accomplished with the help of error-correction coding, a technique of adding a controlled amount of redundancy to the information sequence in order to correct the errors that are made when the data are transmitted over the noisy channel. This seemingly astounding result is not especially practical, since it leads to system characteristics which are not realistic, such as infinite delay before the message can be decoded. However, its tremendous significance is due to the fact that it shows the power of error-correction coding in the design of communication systems. *See* INFORMATION THEORY.

Multiple-access techniques. In many communication systems, multiple sources and multiple destinations are present, and the manner in which accessing the channel is achieved becomes important. Perhaps the most common examples of such multiple-access links are those used by commercial radio and television stations. These systems operate by assigning to each transmitted waveform a distinct frequency band which is adjacent to but nominally not overlapping with its neighboring bands. In this way, there is approximately no interference among the various users. This type of operation, known as frequency-division multiple accessing (FDMA), can be used with either analog or digital modulation formats. *See* RADIO; TELEVISION.

Because of certain problems that arise with frequency-division multiple accessing, such as the generation of intermodulation products when such a waveform is passed through a nonlinear amplifier, other types of multiple-accessing schemes have been devised. (An intermodulation product corresponds to a frequency generated at the output of a nonlinear device due to the nonlinear interaction of two or more frequencies at the input of the device.) In particular, when digital modulation is being employed, both time-division multiple access (TDMA) and code-division multiple access (CDMA) are possible alternatives.

In time-division multiple access, all users occupy the full bandwidth of the channel, but they transmit sequentially in time, and only one user has access to the channel at any given time. The system operates as follows. Each user is assigned a certain amount of channel time, and when the time slot of one user ends, the next user begins. After all the users have had an opportunity to transmit, the channel returns to the first user and the process is repeated. The time for the channel to return to a given user is referred to as the frame time. A predetermined pattern of bits, known as a frame synchronization pattern, is transmitted at the beginning of each frame. This pattern is known by each user, and it allows the users to synchronize to it so that they will know to transmit their data only during the allocated time interval, thus avoiding interference between the users. The advantage of the system over frequency-division multiple accessing is that the intermodulation problem disappears since, even if the channel is nonlinear, only one signal is on it at any time.

The technique of code-division multiple access operates by allowing multiple users to individually occupy the full bandwidth of the channel while simultaneously transmitting. This is accomplished by superimposing on the data of each user a wide-bandwidth digital waveform which identifies that user. Typically, a different identification sequence is assigned to every user. At the receiver this sequence for a given user allows the receiver to demodulate the data of that user while rejecting the waveforms of the other users.

Code-division multiple access is an application of a more general technique known as spread spectrum communication, in which the transmission bandwidth is purposely made much larger than the information bandwidth. This has the advantage that the resulting waveform is much less vulnerable to interference than a conventional waveform. Since the interference could be due to any one of a variety of sources, such as multipath or intentional jamming, this sytem has an inherent natural interference immunity which both frequency-division and time-division multiple access lack. *See* SPREAD SPECTRUM COMMUNICATION.

Other types of multiple-accessing schemes are designed to operate over a specific channel. One

example is space-division multiple access, which is used over a satellite communications link. In this scheme, spot beams on the satellite are used to illuminate different areas on the earth, so that the same frequency band can be used simultaneously by each of the spot beams without causing any interference. *See* MULTIPLEXING AND MULTIPLE ACCESS.

Random access. In many cases, the number of potential users is much greater than the number which can be simultaneously accommodated on the channel. However, if the percentage of time employed by each user is statistically very small, the users can compete for access to the channel. When a given user has a message ready for transmission, most of the time a vacant slot on the channel will be found, allowing data transmission. However, at times, all available slots on the channel are taken, and the user has to delay sending the message.

Systems that operate in this manner are referred to as random-access systems, and are typical of computer communication networks. Depending on the geographical size of such networks, they have come to have their own specialized names. For example, they are known as local-area networks (LANs) if they are concentrated over an area roughly the size of a few blocks or less. If the terminals which are furthest apart are within a few miles of one another, they are referred to as metropolitan-area networks (MANs). Networks that span still larger geographical distances are called wide-area networks (WANs). *See* DATA COMMUNICATIONS; LOCAL-AREA NETWORKS; WIDE-AREA NETWORKS.

Laurence B. Milstein

Bibliography. M. Schwartz, *Information Transmission, Modulation and Noise*, 4th ed., 1990; H. Taub, *Principles of Communication Systems*, 2d ed., 1986; R. E. Ziemer and W. H. Tranter, *Principles of Communications*, 3d ed., 1990.

Electrical conduction in gases

The process by means of which a net charge is transported through a gaseous medium. It encompasses a variety of effects and modes of conduction, ranging from the Townsend discharge at one extreme to the arc discharge at the other. The current in these two cases ranges from a fraction of 1 microampere in the first to thousands of amperes in the second. It covers a pressure range from less than 10^{-4} atm (10 pascals) to greater than 1 atm (10^5 pascals). *See* ARC DISCHARGE; TOWNSEND DISCHARGE.

In general, the feature which distinguishes gaseous conduction from conduction in a solid or liquid is the active part which the medium plays in the process. Not only does the gas permit the drift of free charges from one electrode to the other, but the gas itself may be ionized to produce other charges which can interact with the electrodes to liberate additional charges. Quite apparently, the current voltage characteristic may be nonlinear and multivalued. *See* ELECTRICAL CONDUCTIVITY OF METALS; ELECTROLYTIC CONDUCTANCE; SEMICONDUCTOR.

The applications of the effects encountered in this area are of significant commercial and scientific value. A few commercial applications are thyratrons, gaseous rectifiers, ignitrons, glow tubes, and gas-filled phototubes. These tubes are used in power supplies, control circuits, pulse production, voltage regulators, and heavy-duty applications such as welders. In addition, there are gaseous conduction devices widely used in research problems. Some of these are ion sources for mass spectrometers and nuclear accelerators, ionization vacuum gages, radiation detection and measurement instruments, and thermonuclear devices for the production of power.

The discussion of this complicated process will be divided into two parts. The first will deal with the basic effects involved, including production and loss of charges within the region and the motion of charges in the gas. The second part will deal with the mechanism of conduction.

Basic effects. To produce gaseous conduction, two conditions must obtain. First, there must be a source of free charges. Second, there must be an electric field to produce a directed motion of these charges. Considering the first of these, one finds that the free charge concentration is a result of a number of processes which produce and remove charges.

Sources of free charges. In many gaseous devices, a thermionic-emission cathode is included. The process of electron emission from a heated electrode is well known. Closely related to this as a source of electrons is field emission. Here, a strong positive field at a metallic surface lowers the barrier for electron emission. Thus, the electron current from a surface at a given temperature may be significantly increased. *See* THERMIONIC EMISSION.

Both of these effects result in electron production. Another effect, photon absorption, may result only in electrons if the absorber is a solid. However, if the photon interacts with a gas molecule or atom, ionization may result and both an electron and a positive ion can be obtained. The photon may come from some external source or it may be a secondary effect of the gaseous conduction. It may have a wavelength in the visible, ultraviolet, or x-ray region.

Conduction in flames is largely a result of the thermal production of ionization. This is a specialized field which has long been of interest in chemistry and combustion studies. To produce appreciable thermal ionization, the temperature must be high, as in a flame. If the effective temperature is known, the ionization concentration may be determined from statistical mechanics. Thermal ionization is also of tremendous importance in devices for production of power by thermonuclear processes, and in ion-propulsion equipment. A special form of this is surface ionization, in which a hot surface may cause ionization of a gas atom that comes in contact with it.

Another source of ionization is particle radiation due to cosmic rays, radioactive material in the walls or in the gas, or particles produced from an external source. These particles may then produce an ionized track in the gas. An example is the ionization produced by an α-particle from a polonium source in an ionization chamber.

In most of these methods of charge production, the sources are primary. That is, the presence of other free charges is not important in the production of ionization or electrons. Other processes are secondary in origin, although they may be of prime importance. It was pointed out that photons could originate either externally or as a secondary effect. Field emission could be a secondary effect, also. Other methods of ionization, however, are generally thought of as being secondary in origin. Ionization of the gas by electron impact is such a case. Here free electrons may gain enough energy in an electric field to interact with an atomic or molecular electron to produce an ion pair.

Cumulative ionization is an extension of ionization by impact. If the original electron and its offspring gain enough energy so each may produce another electron, and if this process is repeated over and over, the result is called an avalanche, and the ionization thus produced is referred to as cumulative ionization. This is the basis for particle detection in some ionization devices. *See* GEIGER-MÜLLER COUNTER; IONIZATION CHAMBER.

Another secondary source is electron emission from either electron or positive ion bombardment of a surface. This should not be confused with thermionic emission resulting from heating under bombardment. *See* SECONDARY EMISSION.

Other sources which may be important are atomic collisions, sputtering, and collisions of the second kind. In the first case, an atom or heavy ion may collide with an atom to produce ionization. This is quite unlikely until an energy of many times the ionization energy is obtained. The second is somewhat analogous to secondary electron emission. Here the positive ions strike a surface and knock out atoms or groups of atoms. Some of these come off as ions. In the third case, an excited atom may interact with an atom or molecule which is chemically different and has an ionization potential lower than the excitation potential of the excited atom. The result may be the decay of the excited state with ionization of the struck molecule or atom. Symbolically, this is shown by reaction (1), where A* is the excited atom, B the

$$A^* + B \rightarrow A + B^+ + e^- \qquad (1)$$

struck atom or molecule, and e^- an electron. *See* EXCITATION POTENTIAL; IONIZATION POTENTIAL; SPUTTERING.

Free-charge removal. The net free-charge concentration is a balance between charge-production and charge-removal processes. Recombination is one such process. Here, an electron or heavy negative ion and a positive ion may recombine. The energy transition may appear as electromagnetic radiation or may be carried off by a third body, if one is present. There are a wide variety of conditions which may lead to recombination. Where the temperature and electric field are high, the recombination will occur predominantly at the walls.

The method of charge removal is important from the aspect of conduction, however. If the charges move to the appropriate electrodes under the influence of the field and there recombine, then they contribute to the current. If they simply diffuse to the walls and recombine there with ions of the opposite sign, or if they recombine in the gas volume, they may not appear as part of the external current.

Motion of the charges. The motion of the charges within the gas will be largely influenced by the potential function. For the usual regular geometries, this could be calculated in principle if there were no charges present. However, in a gas with free charges distributed throughout, the problem is quite different. The charges modify the charge-free potential, but the potential itself determines how the charge will move. The motion of the charges further modifies the potential and so on. Although the situation can be described physically by Poisson's equation, it is generally impractical to carry out an analysis. As a practical result, the potential function must be determined by measurements which are made with probes. This requires careful procedures to obtain significant results.

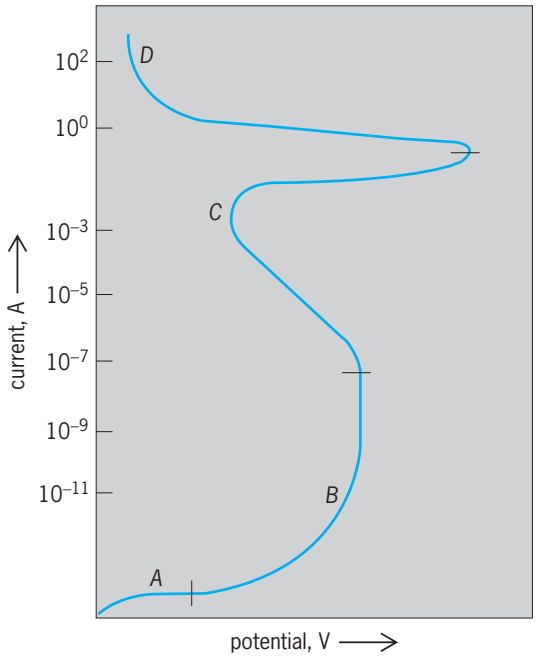
Diffusion of ions. Ion diffusion is a type of random motion which is always present and is the result of thermal or agitation energy. The randomness of the motion is brought about by the many collisions with molecules and other ions. A great difference exists in the motions of electrons and heavy ions. Because of low mass, electrons are easily deflected, so they move erratically. They diffuse badly, and follow field lines only generally. Again because the mass is small, an electron can give up appreciable energy only in an inelastic collision, in which excitation or ionization takes place. Hence, electron agitation and diffusion will be much greater in a pure inert gas than in a gas having many low-energy molecular states. Conversely, heavy ions exchange energy effectively at every collision. Diffusion is much less, so that they follow the electric field lines more closely than do electrons.

Mechanism of conduction. The ionic mobility μ relates drift velocity v to electric field X by Eq. (2). For

$$v = \mu X \qquad (2)$$

electrons, the mobility is high, and a drift velocity of 10^6 cm/s or greater may be obtained. The electronic mobility is not a true constant, but varies with field, pressure, temperature, and gas composition. For heavy ions, the mobility is much more nearly constant, but is still dependent on these quantities to some extent. Drift velocities are usually of the order of 10^3–10^4 cm/s. Thus, in a typical conduction device, an electron may move from one electrode to the other before a heavy ion is displaced appreciably.

It would appear from the foregoing that if accurate information about the important processes existed, one could predict the characteristics of the conduction process under given conditions. Unfortunately this is not the case. Generally, the situation is so complicated that the theory can yield only qualitative



Current-potential characteristics for a two-electrode device with constant pressure.

predictions. Accordingly, most of the information concerning the various forms of gaseous conduction is empirical. In the present description, it will be possible to mention the main features of a few of these modes.

The **illustration** shows a sample voltage-current characteristic for a two-electrode device with constant pressure. It is assumed that there is a constant source of ionization which could be any of the primary sources previously discussed. In region *A*, the current first rises and then over a limited range is relatively constant as the voltage across the electrodes is increased. The initial rise is the result of the collection of charges which were either recombining or diffusing to the walls. The nearly constant current region is the result of the collection of almost all of the charges.

In region *B*, further increase in voltage produces an increase in current. Here, ionization by electron impact is occurring. The situation is described by specifying that each free electron makes α additional ion pairs in traveling 1 cm in the direction of the field. The number of ion pairs produced per second in 1 cm at a distance x from the cathode (assuming parallel plate electrodes) is given by Eq. (3),

$$n = n_0 e^{\alpha x} \tag{3}$$

where n_0 is a constant depending on the initial number of electrons. This is a form of the Townsend equation, and α is the first Townsend coefficient. In the region *B* in the illustration, the increase in current represents an increase in α. Near the end of this region, the current increases more rapidly with applied field. Here, additional effects are taking place, such as the photoelectric process and secondary emission. This situation is described by Eq. (4), where β is the second Townsend coefficient, i_0 is the initial electron current at the cathode, and i is the anode current as a function of plate separation x; β is also a function of electric field.

$$i = i_0 \frac{(\alpha - \beta)\, e^{(\alpha-\beta)x}}{\alpha - \beta\, e^{(\alpha-\beta)x}} \tag{4}$$

At the end of the region, the slope becomes infinite, and if the external resistance is not too large, the current will jump in a discontinuous fashion. The transition is referred to as a spark, and the potential at which it occurs is the breakdown or sparking potential. The region *B* is called a Townsend discharge and is not self-sustained. Thus, if the source of primary ionization were removed, the discharge would cease. *See* BREAKDOWN POTENTIAL; ELECTRIC SPARK.

As the potential reaches the sparking potential, a transition occurs to the region *C*. This is the self-sustained glow discharge region. Over an extensive current range, the voltage drop remains substantially constant. During the current increase, a glow occurs at the cathode, and at the upper end of the range, the cathode is completely covered. At this point, a further current increase can be achieved only if the potential drop across the discharge is increased. This portion of the characteristic is known as the abnormal glow. Throughout this portion of the discharge characteristic curve, secondary effects are quite important. Particularly vital are the effects of cumulative ionization and secondary emission at the cathode. *See* GLOW DISCHARGE.

Further increase in current leads to another mode of discharge, the arc. This is shown as region *D* in the illustration. Characteristic of this mode is the low cathode potential fall and the very high current density. It is generally felt that the predominant effect in the production of the large number of electrons at the cathode necessary for the arc is thermionic emission. This is consistent with the very high temperatures known to exist either generally or locally on the cathode. Although the arc type of discharge has very great commercial value, the mechanism of its operation is not very well understood.

In addition to these general types of conduction, there are very special cases of considerable interest. Some of these are the corona discharge, radiofrequency or electrodeless discharge, hot-cathode discharge, and discharges in the presence of a magnetic field.

Glenn H. Miller

Bibliography. M. N. Hirsh and H. J. Oskan (eds.), *Gaseous Electronics*, vol. 1: *Electrical Discharges*, 1978; E. E. Kunhardt and L. H. Luessen (eds.), *Electrical Breakdown by Discharges in Gases*, 2 vols., 1983; Y. P. Raizer and J. E. Allen, *Gas Discharge Physics*, 1997.

Electrical conductivity of metals

The property of a metal that measures its ability to conduct electricity, following Ohm's law. Electrical conductivity σ (measured in $\Omega^{-1} \cdot m^{-1}$) is the

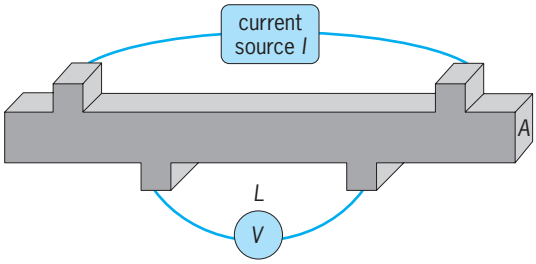


Fig. 1. Schematic of a four-probe measurement of electrical conductivity. Current *I* is fed through the outer leads, and voltage drop *V* is measured on the inner leads. In this way the contact potential drop experienced by the applied current is localized at the junctions with the outer leads. Measuring *V* with minimal current through the voltmeter minimizes the contact potential contribution to the measured conductance $G = I/V$.

reciprocal of the resistivity ρ [in $\Omega \cdot$ m, where resistance R in Ohms (Ω) is voltage drop V in volts (V) divided by current I in amperes (A)]. Ohm's law, $V = IR$, may also be written in the form of Eq. (1), where j is current density (current per

$$j = \sigma E \qquad \text{or} \qquad E = \rho j \tag{1}$$

unit area, I/A, measured in A/m^2) and $E = V/L$ is electric field or electrical potential gradient, measured in V/m, where V is the voltage drop measured across a length L of material (**Fig. 1**). The conductivity σ is an intrinsic property of a pure material, related to the measured conductance $G = I/V$ via $G = \sigma A/L$, just as resistance $R = V/I$ relates to resistivity ρ via $R = \rho L/A$. Thus a large-gauge conductor [large area (A)] of short length has a high conductance, but the physical dimensions do not affect the conductivity σ or the resistivity ρ. Positive current flows from higher to lower voltage, and σ is never negative. *See* ELECTRICAL RESISTANCE; ELECTRICAL RESISTIVITY; OHM'S LAW.

Metals versus insulators. Metals have large conductivity (typically greater than $10^5\ \Omega^{-1} \cdot \text{m}^{-1}$ but never greater at room temperature than that of silver, $0.66 \times 10^8\ \Omega^{-1} \cdot \text{m}^{-1}$.) As temperature T is lowered, conductivity almost always increases in metals (**Fig. 2**). These properties contrast with semiconductors and insulators, which have smaller conductivity and different temperature dependence. The reason concerns the spectrum of quantum excited states in matter (**Fig. 3**). Outer electrons (occupying frontier orbitals in chemical language) have a continuous range of energy levels in solids, called a band of quantum states. In metals the highest-energy occupied state and the least-energy empty state differ only infinitesimally in energy, whereas in semiconductors or insulators they are separated by a significant energy gap. In both semiconductors and metals the states that correspond to the frontier orbitals are mobile, meaning that they permit electrons to travel from atom to atom. In semiconductors the energy gap is moderate (typically 0.5–3.0 eV). If electrons are introduced into empty states above the gap, a semiconductor behaves much like a metal. In an insulator the energy gap is larger, 5 eV or more, and electrons typically get trapped on atoms or defects, rendering them comparatively immobile. *See* BAND THEORY OF SOLIDS; ELECTRIC INSULATOR; HOLE STATES IN SOLIDS; SEMICONDUCTOR.

Propagation of electrical current. To outside observers, electrical current flow in a wire resembles water flow in a pipe, but the analogy does not explain what is going on internally. In water flow, molecules have a net drift velocity, which is preserved by intermolecular collisions. Locally there is Galilean invariance, which means that an internal observer moving with the hydrodynamic current and observing individual molecules would not be able to detect the current. In a metal the electrons disengage from their parent positive ions. The ions vibrate around fixed positions, not participating in the current. When electrical current j flows, the electrons have a net drift velocity which can be detected by a local observer watching ions as well as electrons. The system does not have Galilean invariance. Collisions of electrons do not conserve j, and act to restore electrons to the $j = 0$ state of thermal equilibrium, not the moving local equilibrium of hydrodynamics. This is the source of resistance in metals. Electrical current is less "collective" than hydrodynamic current. However, at very low temperature many metals become superconductors, seen in the data of Fig. 2. This is a different state of electronic matter in which electrons behave more collectively than they do in the "normal" or higher-temperature resistive state of metallic electronic matter. The collective (and quantum-mechanically phase-coherent) superconducting state allows $j \neq 0$ with $E = 0$, thus making $\sigma = \infty$. *See* SUPERCONDUCTIVITY.

Electrons in quantum theory are waves as well as particles, and can diffract around the crystalline

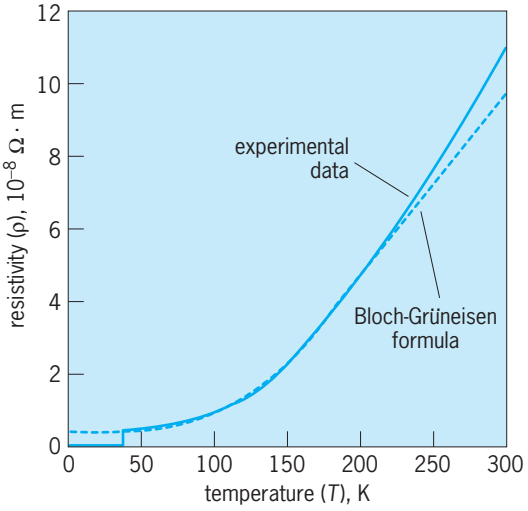


Fig. 2. Electrical conductivity of magnesium diboride (MgB_2), plotted, as is customary for metals, as resistivity (ρ) versus temperature (T). The unit $10^{-8}\ \Omega \cdot$ m is often written as 1 micro-ohm cm or 1 $\mu\Omega$-cm. Below $T = 40$ K $= -233^\circ$C, MgB_2 is a superconductor with $\rho = 0$. A fit using Bloch-Grünisen theory, explained in the text, is also shown. (*Data from Z. X. Ye et al., Electron scattering dependence of dendritic magnetic instability in superconducting MgB_2 films, Appl. Phys. Lett., 86:5284–5286, 2004*)

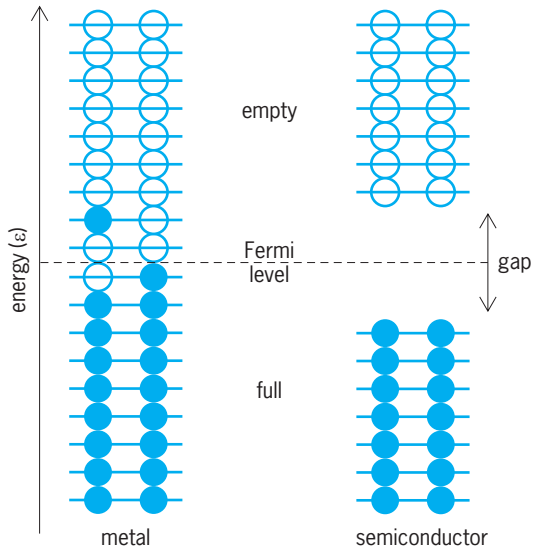


Fig. 3. Schematic energy spectrum in a metal versus a semiconductor. Filled and empty circles represent full and empty quantum states, two per level corresponding to the two spin orientations, just as in an atom. Adjacent levels are infinitesimally close except at the Fermi level of a semiconductor, which lies in the energy gap where there are no quantum states. The metal states are shown with a filled state above the Fermi level and a vacant state or hole below the Fermi level, as would happen by ordinary thermal excitation at room temperature. Semiconductors are likely to have few thermal excitations. In such a state the conductivity σ would be close to zero.

lattice of ions. In perfect crystals, if ions were not vibrating, the electron waves would remain coherent over the whole sample size, and conductivity would be infinite. The actual finite value occurs because electrons scatter from defects in the lattice, from atomic vibrations, and from each other. The last source of scattering is small, and the other sources are minimized in pure crystals at low temperature. Unlike molecules in water, which collide frequently, electrons in good metals collide infrequently and have long "mean free paths" or distances of travel between collisions. The collision frequency for water molecules at room temperature (300 K), denoted $1/\tau$, is about 6×10^{12}/s. This comes from dividing the thermal velocity, given by Eq. (2), by the distance $\sim 10^{-10}$ m between collisions. For electrons in metals the frequency of collisions with atomic vibrations at $T = 300$ K is given by Eq. (3). In these formulas, k_B is Boltzmann's constant, M is the mass of a water molecule, and $\hbar$ is Planck's constant divided by 2π. Compared to atoms, electrons have much higher mean velocities, $v_{el} \sim (2\text{–}10) \times 10^5$ m/s, and v_{el} is independent of temperature. This is a quantum effect caused by the Pauli principle. In a free-electron model, where all diffraction from atoms is ignored, the value is called the Fermi velocity and is given Eq. (4), where n is the number of electrons per unit volume, of the order of 10^{29} m^{-3}, and m is the electron mass. As a consequence of the larger velocity, the mean free path for electrons, estimated as velocity divided by collision rate, is typically 10 to 20 interatomic spacings, 50 times longer than for water molecules. *See* CRYSTAL DEFECTS; EXCLUSION PRINCIPLE; FREE-ELECTRON THEORY OF METALS; LATTICE VIBRATIONS; QUANTUM THEORY OF MATTER.

$$v_{th} = \sqrt{3k_BT/M} \approx 600 \text{ m/s} \qquad (2)$$

$$1/\tau \approx (1 - 10)k_BT/\hbar \sim 10^{14}/\text{s} \qquad (3)$$

$$v_F = \hbar(3\pi^2 n)^{1/3}/m \qquad (4)$$

Bloch-Boltzmann theory. The long mean free path of electrons in metals is the key to formulating a theory. In a perfect crystal the quantum description of electrons as particle-wave entities is similar to the idea of electromagnetic waves confined in a resonant cavity. The wavelengths have to "fit" the dimensions of the cavity, which causes quantization of wavelength λ. The individual quantum states or orbitals are labeled by their wavevectors $k = 2\pi/\lambda$. The orbital labeled by k has an energy $\varepsilon(k)$, and is pictured as a wave oscillating with a space and time dependence given approximately by $\cos [kx - \varepsilon(k)t/\hbar]$, like a traveling electromagnetic wave with frequency $\omega(k) = \varepsilon(k)/\hbar$. A wave packet built from such orbitals moves with velocity given by Eq. (5). There is no simple formula for the function $\varepsilon(k)$. It has to be measured by photoemission or computed using band theory, and varies from metal to metal. The wave packets in real metals do not remain phase-coherent over the whole sample, only over the distance between collisions. Provided this distance is longer than a wavelength, the wavelength and wavevector remain meaningful quantities, which can be used to build a theory, as was done by Felix Bloch in 1928.

$$v(k) = [d\varepsilon(k)/dk]/\hbar \qquad (5)$$

Bloch's theory starts from the recognition of Bloch orbitals with wavevectors k. It then adopts from Boltzmann the notion of the distribution $F(k)$ which gives the probability that the state k is occupied. Because of the Pauli principle, this probability cannot exceed 2 (one for spin "up" and one for spin "down"). For states of low energy [$\varepsilon(k)$ far below the Fermi level ε_F], the occupancy is very close to 2, and for states of high energy [$\varepsilon(k)$ far above the Fermi level ε_F], the occupancy is very close to 0. In thermal equilibrium the occupancy $F(k)$ is equal to the Fermi-Dirac function, given by Eq. (6). The conductivity is calculated by summing the velocities of the occupied states, yielding Eq. (7).

$$f(k) = \frac{2}{\exp\left[\dfrac{\varepsilon(k) - \varepsilon_F}{k_BT}\right] + 1} \qquad (6)$$

$$\sigma = j/E = \frac{-e}{E \cdot Vol} \sum_k v(k)F(k) \qquad (7)$$

When the field E is zero, the distribution $F(k)$ is $f(k)$ in Eq. (6), which has $j = 0$ because states with

positive and negative velocities $v(k)$ are equally populated. The effect of the field E is to cause the k-vectors of the occupied states to change according to Bloch's law, Eq. (8). This is the quantum version

$$\hbar dk/dt = -eE \tag{8}$$

of Newtonian acceleration by the electric force $-eE$, using de Broglie's momentum $\hbar k$ of a quantum wave. The acceleration is continuously resisted by scattering. Therefore the total shift of the typical k-vector is given by Eq. (9), where τ is the time between collisions. Then, using $F(k) \approx f(k - \Delta k)$, we can write the answer for the conductivity as Eq. (10), where $(n/m)_{\text{eff}}$ is given by Eq. (11).

$$\Delta k \approx -eE\tau/\hbar \tag{9}$$

$$\sigma = \left(\frac{n}{m}\right)_{\text{eff}} e^2\tau \tag{10}$$

$$\left(\frac{n}{m}\right)_{\text{eff}} = \frac{-1}{\hbar Vol}\sum_k v(k)\frac{df(k)}{dk} = \frac{1}{Vol}\sum_k \left[\frac{d^2\varepsilon}{\hbar^2 dk^2}\right] f(k) \tag{11}$$

See NONRELATIVISTIC QUANTUM THEORY.

The last part of Eq. (11) is derived by integration by parts after substituting the expression given by Eq. (5) for $v(k)$, and can be interpreted as electron density, given by Eq. (12), times reciprocal effective mass, given by Eq. (13). The Bloch theory recovers a

$$n = \left(\sum f(k)\right) \Big/ Vol \tag{12}$$

$$1/m^* = d^2\varepsilon/d(\hbar k)^2 \tag{13}$$

result proposed on classical grounds by Paul Drude shortly after the discovery of the electron, namely $\sigma = ne^2\tau/m$. However, the interpretation is quite different. A proper quantum theory for the scattering rate τ is needed, and was provided by Bloch by making a quantum generalization of the Boltzmann scattering operator. *See* BOLTZMANN TRANSPORT EQUATION.

An approximate solution of Bloch's theory, known as the Bloch-Grüneisen formula, is often very successful in fitting data, as shown in Fig. 2 for magnesium diboride (MgB_2). In this example, theory deviates from experiment at higher temperatures because high-frequency lattice vibrations, involving boron atoms, are not well represented by the form of the approximate theory. The extrapolated $T = 0$ resistivity value, $\rho_0 = 0.4 \times 10^{-8}\ \Omega \cdot \text{m}$, is called the residual resistivity, and would be zero in a perfect crystal.

Bloch's theory has failures, but they are outnumbered by the successes. One failure is that collective behavior can sometimes be seen at low temperature, the outstanding example being superconductivity. Other examples are more controversial, and include so-called Luttinger liquid behavior, which definitely happens in ideal one-dimensional theoretical models and probably is seen in careful experiments on

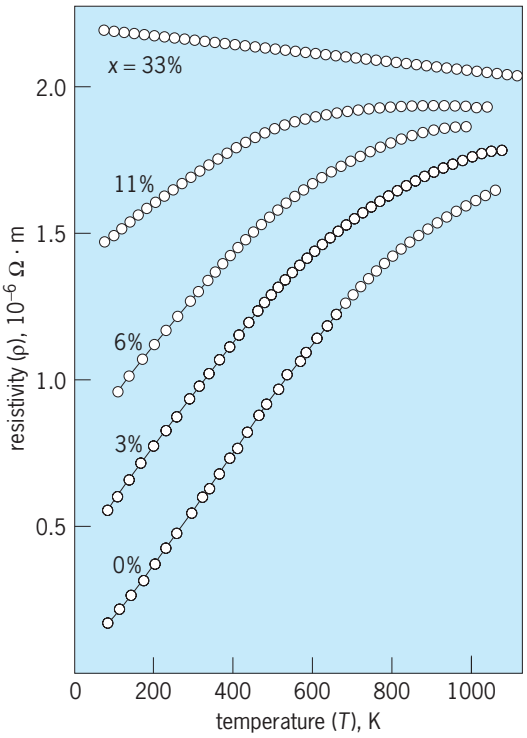


Fig. 4. Resistivity ρ of $Ti_{1-x}Al_x$ alloys as function of temperature (T) for various aluminum concentrations x. (*From J. H. Mooij, Electrical conduction in concentrated disordered transition-metal alloys, Phys. Stat. Sol. (a), 17: 521–530, 1973*)

such systems as carbon nanotubes. Another case is transport by so-called sliding charge-density waves, perhaps seen in quasi-one-dimensional metals like $NbSe_3$. A more common failure is that in metals which are very impure or which have very strong scattering, the mean free path between scattering events may become so short that the wavelength and wavevector of electron quantum states is no longer definable. This undermines the basis of Bloch's theory. *See* CHARGE-DENSITY WAVE.

Such a failure of the theory (technically, a failure of the quasiparticle picture) is illustrated in **Fig. 4**, showing the resistivity of $Ti_{1-x}Al_x$ alloys. At the lower aluminum concentrations, $x = 3\%$ and 6%, the $\rho(T)$ curves shift upward, obeying Mattheissen's rule, given by Eq. (14), where the residual resistivity

$$\rho = \rho_0 + \rho_{\text{pure}}(T) \tag{14}$$

ρ_0 is proportional to impurity concentration x. For the larger concentrations, a too short mean free path causes the quasiparticle picture underlying Bloch-Boltzmann theory to fail. This in turn causes failure of Mattheissen's rule. The quasiparticle picture also fails at high temperature for all concentrations x. *See* MATTHIESSEN'S RULE.

A more abstract formulation of nonequilibrium effects like conductivity exists under the name of Kubo formulas. These provide a route to improved theories.

Philip B. Allen

Bibliography. N. W. Ashcroft and N. D. Mermin, *Solid State Physics*, Sauders, Philadelphia, 1976; J.

Bass, Electrical resistivity of pure metals and dilute alloys in *Landolt-Börnstein Tables 15a*, Springer, Berlin, 1980; O. Gunnarsson, M. Calandra, and J. E. Han, Colloquium: Saturation of electrical resistivity, *Rev. Mod. Phys.*, 75:1085–1099 2003; C. Kittel, *Introduction to Solid State Physics*, 8th ed., Wiley, New York, 2005; J. M. Ziman, *Electrons and Phonons*, Oxford, 1960, reprint 2001.

Electrical connector

A device that joins electrical conductors mechanically and electrically to other conductors and to the terminals of apparatus and equipment. The term covers a wide range of devices designed, for example, to connect small conductors employed in communication circuits, or at the other extreme, large cables and bus-bars.

Electrical connectors are applied to conductors in a variety of ways. Soldered connectors have a tube or hole of approximately the same diameter as the conductor. The conductor and connector are heated, the conductor inserted, and solder flowed into the joint until it is filled. Solderless connectors are applied by clamping the conductor or conductors in a bolted assembly or by staking or crimping under great mechanical force, usually by means of special tools designed for the purpose.

Industrial and power types. Connectors for electric power systems are commonly cast and assembled with bolts. The material most commonly used is copper alloy, which may vary in composition and properties depending upon the intended use of the connector. Where heavy currents flow through the body of the connector, an alloy of relatively high conductivity is used. Where the connector serves principally to clamp conductors together, an alloy of higher strength and lower conductivity may be used. An important consideration governing the choice of alloy is that it must have a coefficient of thermal expansion very close to that of the conductor material itself (copper or aluminum), so that the connection will remain secure through wide temperature changes.

Connectors designed for severe outdoor service and heavy current and mechanical loading are called power connectors and are widely used in power substations for cable-to-stud connections at equipment, cable-to-bus, bus-to-bus, and bus-to-line connections.

For industrial application, more compact constructions and greater versatility may be required, and connectors with such features are called industrial connectors. In practice, both classes are employed in power systems and industrial plants, the choice depending upon the actual service conditions.

Typical connector types are in-line splice couplers, T-tap connectors, terminal lugs, and stud connectors. Couplers join conductors end to end. T-tap connectors join a through conductor to another conductor at right angles to it (**Fig. 1***a*). Terminal lugs join the conductor to a drilled tongue for bolting to the terminals of equipment (Fig. 1*b*). Stud connectors join the conductor to equipment studs; the stud clamp is threaded or smooth to match the stud. Many variations of these types are made.

Split-bolt connectors are a compact construction widely used for splices and taps in building wiring. The bolt-shape casting has a wide and deep slot lengthwise. The conductors are inserted in the slot and the nut is drawn up, clamping the conductors together inside the bolt (Fig. 1*c*). Newer types of connectors are made to serve this same function.

Expansion connectors or flexible connectors allow some limited motion between the connected conductors. The clamp portions of the connector are joined by short lengths of flexible copper braid and may also be held in alignment by a telescoping guide.

Heavily tinned copper-alloy bolted connectors or aluminum-alloy-body connectors may be used to connect aluminum conductors. They are applied with an oxide-inhibiting compound. Where copper and aluminum conductors are to be clamped together, they are separated by bimetal or tinned washers to prevent galvanic corrosion.

In electronic equipment assembly, involving hookup of wires to terminals and of other components onto printed circuit boards, connections are made by soldering the wire or component to the terminal or printed circuit board. Although soldered terminals are by far the most common type of wiring connections in electronic chassis applications and in manufacture of electronic and electrical equipment, the pressure-wrap connection is a tool-applied method used in the field that speeds up connections where very many have to be made, as in wiring telephone and signal and control equipment. The tool that applies the wire wrap to the terminal stud assures an extremely tight, low-resistance, high-reliability connection by slightly elongating the wire as it is wrapped, putting a tight constrictive force on the terminal.

Separable types. These consist of matched plugs and receptacles, which may be readily separated to disconnect a conductor or group of conductors from the circuit or system. Separable connectors are commonly used for the connection of portable appliances and equipment to an electric wiring system. In electronics and communication they are used for connecting various components so that a single component can be readily removed for servicing without disturbing the circuitry.

For connections to an electric power supply, separable connectors consist of receptacles, usually fixed and permanently connected to the wiring system,

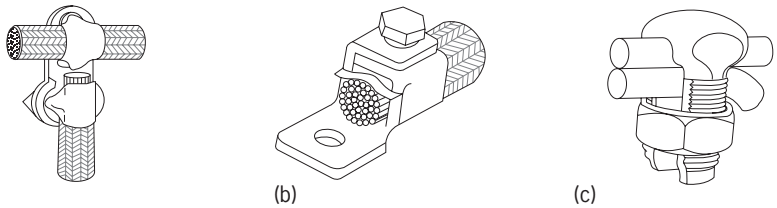


Fig. 1. Types of connectors. (*a*) T-tap connector. (*b*) Terminal lug. (*c*) Split-bolt connector.

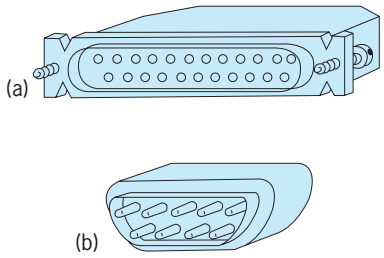


Fig. 2. Typical subminiature computer communications connectors. (*a*) 25-pin RS-232 plug. (*b*) 9-pin RS-422 jack.

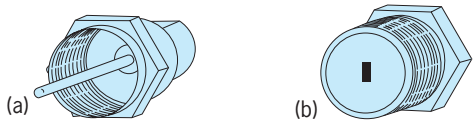


Fig. 3. Typical coaxial cable connection: 5/8 in., 24 threads per inch. (*a*) Jack. (*b*) Plug.

and attachment plugs, usually connected to a cord or cable extending to the appliance or equipment served. Other constructions provide cable-to-cable and cable-to-appliance connections.

Connectors for power systems are typically two-wire, three-wire, or four-wire, depending upon the circuit requirements. They are made in a wide range of types and sizes for applications involving various voltage, current, mechanical, and environmental requirements. They are so constructed that a particular receptacle will accept only a matching plug of identical characteristics, rating, and circuit configuration. Two-wire connectors that can be coupled in only one position are said to be polarized. Multiconductor connectors are designed, by a nonuniform arrangement of contacts or by a key in the assembly housing, so that they may be coupled in only one position.

Locking types. These are designed so that, when coupled, they may not be separated by a strain or pull on the cord or cable. In a typical construction the plug is inserted and twisted through a small arc, locking it securely in place. For hazardous areas, where the spark caused by separating a connector under load may be dangerous, connectors are designed so that the circuit is made or broken within the receptacle only in the fully engaged and locked position. Receptacles may also be interlocked mechanically with a switch, allowing insertion or removal of the plug only when the switch is in the open position.

Multiwire connectors used in electronic, communication, and control circuits contain more connections of much smaller current capacity than power connectors (**Fig. 2**). Connector constructions are designed primarily for mechanical and electrical contact reliability.

Coaxial connectors provide for connecting and disconnecting a coaxial cable from an equipment circuit (**Fig. 3**). They are so constructed as to maintain the conductor configuration, through the separable connection, and the characteristic impedance of the coaxial cable. Coaxial connectors are made of copper or aluminum with contact and current-carrying surfaces plated with silver or gold. They are made in a variety of sizes adaptable to the cables with which they are used. They are designed for cable-to-chassis and cable-to-cable connections in plug, jack, receptacle, feed-through, right-angle, and T-adapter designs. *See* COAXIAL CABLE.

Common receptacles. Plug receptacles, sometimes called convenience outlets, are a type of wiring device distributed throughout buildings for the attachment of cord plug caps from portable lamps and appliances. In residences at least one such outlet must be provided for every 12 linear feet (3.66 m) or major fraction of wall perimeter. The receptacle is conventionally two-wire, 120-V, 15-A, and supplied by a 15- or 20-A branch circuit. The slots which accept the attachment plug are parallel. The device may consist of a single receptacle, a duplex receptacle, or a triplex receptacle (for one, two, or three pairs of blade slots for plug attachment, respectively) in a single outlet box. More attachment plugs can be provided by ganging two or more devices in a common outlet box under a common wall plate. Receptacles may also be combined with other devices (as switches and pilot lights).

In split-circuit duplex receptacles, each half of the device is supplied independently through a separate terminal, with typically one part connected directly to the branch circuit conductors and the other connected through a switch. Grounding receptacles have an additional contact that accepts the third round or U-shaped prong of a grounding attachment plug (**Fig. 4**). Grounding receptacles will accept either grounding or nongrounding attachment plugs,

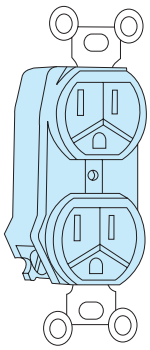


Fig. 4. Grounding-type 15-A receptacle.

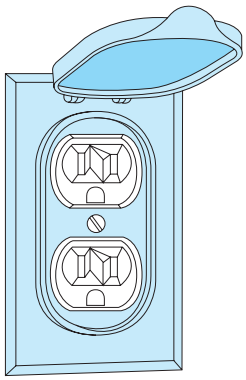


Fig. 5. Weatherproof receptacle with hinged cover.

but grounding attachment plugs can be used only in grounding receptacles. The grounding contact of the receptacle is connected internally to a supporting strap providing a ground both through the outlet box and the grounding conductor, armor, or raceway of the wiring system. *See* GROUNDING.

Multioutlet assemblies are a type of surface raceway designed to hold conductors and plug receptacles, which are usually spaced uniformly along the length of assembly.

Weatherproof receptacles (**Fig. 5**) are equipped with a weatherproof plate, gasketed around the edges, with a hinged or threaded cover over the device. *See* WIRING.

J. F. McPartland; Brian J. McPartland

Bibliography. T. Croft and W. I. Summers (eds.), *American Electricians Handbook*, 13th ed., 1996; IEEE, *National Electrical Safety Code 1997*, 1996; J. F. McPartland, *The Handbook of Practical Electrical Design*, 3d ed., 1999; J. F. McPartland, *National Electrical Code Handbook*, 23d ed., 1999.

Electrical degree

A time interval equal to 1/360 of the time required for one complete cycle of alternating current. Mechanical rotation is often measured in degrees, 360° constituting one complete revolution. In describing alternating voltages and currents, the time for one complete cycle is considered to be equivalent to 360 electrical degrees (360°) or 2π electrical radians. For example, if the frequency f is 60 cycles per second (60 Hz), 360° corresponds to 1/60 second and 1 electrical degree to 1/12,600 second.

There is a definite relationship between electrical and mechanical degrees in rotating electric generators and motors. **Figure 1** shows typical coil and angular relationships in a two-pole alternator. As the magnetic field in the machine moves relative to the coils in the armature winding, the coils are linked sequentially by the fluxes of north and south magnetic poles; two flux reversals induce one cycle of voltage in a given coil. Thus, in a two-pole machine such as in Fig. 1, 360° of electrical cycle corresponds to 360° of mechanical rotation, and an angle measured in mechanical degrees has the same value in electrical degrees. However, in a machine with more than two poles, one electrical cycle is generated per pair of poles per revolution. For example, a six-pole machine generates three cycles of voltage in each armature coil per revolution. In this case, each mechanical degree is equivalent to 3 electrical degrees. The four-pole case is illustrated in **Fig. 2**. In general, the relationship below is valid, where p is the number

$$\text{Number of electrical degrees in a given angle} = \frac{p}{2} \cdot \begin{pmatrix}\text{number of mechanical}\\ \text{degrees in that angle}\end{pmatrix}$$

of magnetic poles of either the rotor or the stator. It follows that the electrical angle between the centers of succeeding poles of opposite polarity is always 180 electrical degrees.

The concept of electrical degrees simplifies the analysis of multipolar machines by allowing them to be analyzed on a two-pole basis. Furthermore, it permits trigonometry to be used in solving alternating-current problems. *See* ALTERNATING CURRENT; ELECTRIC ROTATING MACHINERY; GENERATOR; MOTOR; WINDINGS IN ELECTRIC MACHINERY.

George McPherson, Jr.

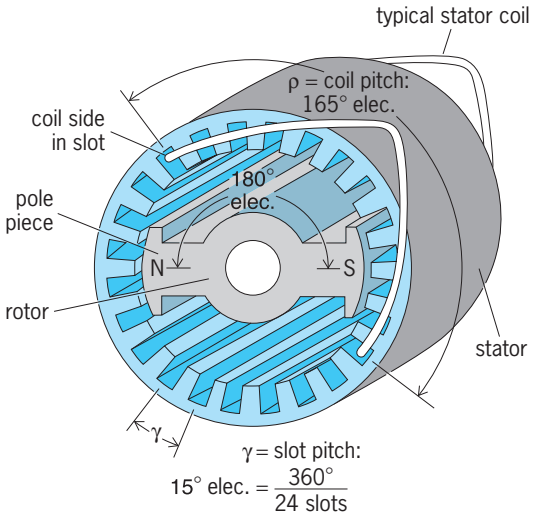


Fig. 1. Coil and angular relationships in a two-pole alternator.

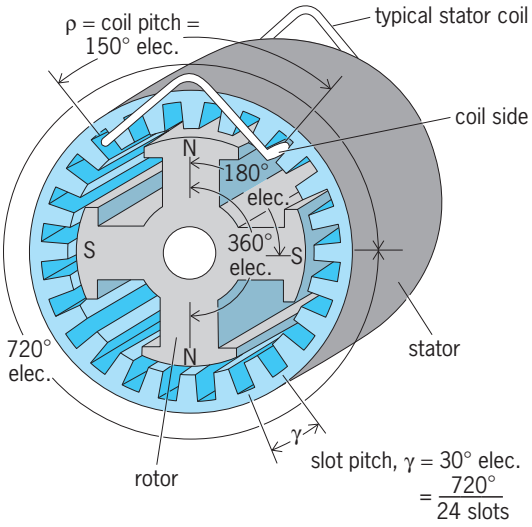


Fig. 2. Coil and angular relationships in a four-pole alternator.

Electrical energy measurement

The measurement of the integral, with respect to time, of the power in an electric circuit. The absolute unit of measurement of electrical energy is the joule, or the charge in coulombs times the potential difference in volts. The joule, however, is too small (1 watt-second) for commercial use, and the more commonly used unit is the watt-hour (3.6×10^3 joules). The most common measurement

application is in the utility field. *See* ELECTRICAL UNITS AND STANDARDS.

Electrical energy is one of the most accurately measured commodities sold to the general public. Many methods of measurement, with different degrees of accuracy, are possible. The choice depends on the requirements and complexities of the measurement problems. Basically, measurements of electrical energy may be classified into two categories, direct-current power and alternating-current power. The fundamental concepts of measurement are, however, the same for both.

Methods of measurement. There are two types of methods of measuring electrical energy: electric instruments and timing means, and electricity meters.

Electric instruments and timing means. These make use of conventional procedures for measuring electric power and time. The required accuracy of measurement dictates the type and quality of the measuring devices used (for example, portable instruments, laboratory instruments, potentiometers, stopwatches, chronographs, and electronic timers). Typical methods are listed below. *See* ELECTRIC POWER MEASUREMENT.

1. Measurement of energy on a direct-current circuit by reading the line voltage and load current at regular intervals over a measured period of time. The frequency of reading selected (such as one per second, one per 10 s) depends upon the steadiness of the load being measured, the time duration of the test, and the accuracy of measurement desired. The first dc electric energy meter was an electroplating cell in series with the load. It deposited a mass of metal on an electrode exactly proportional to the total charge transported to the load. The electrode was weighed periodically to determine the energy used. Errors were introduced if line voltage was not constant. It was replaced by more convenient instruments. *See* CURRENT MEASUREMENT; VOLTAGE MEASUREMENT.

In electrical energy measurements, the losses in the instruments must be considered. Unless negligible from a practical standpoint, they should be deducted from the total energy measured. If the voltmeter losses are included in the total energy measured, then watt-hours $= (VI - V^2R)t/3600$, where V is the average line voltage (volts), I is the average line current (amperes), R is the voltmeter resistance (ohms), and t is the time (seconds).

2. Measurement of energy on a direct-current circuit by controlling the voltage and current at constant predetermined values for a predetermined time interval. This method is common for controlling the energy used for a scientific experiment or for determining the accuracy of a watt-hour meter. For best accuracy, potentiometers and electronic timers are desirable.

3. Measurement of energy on an alternating-current circuit by reading the watts input to the load at regular intervals over a measured period of time. This method is similar to the first, except that the power input is measured by a wattmeter.

4. Measurement of energy on an alternating-current circuit by controlling the voltage, current, and watts input to the load at constant predetermined values. This method is similar to the second, except that the power input is measured by a wattmeter. A common application of this method is to determine the standard of measurement of electrical energy, the watt-hour.

5. Measurement of energy by recording the watts input to the load on a linear chart progressed uniformly with time. This method makes use of a conventional power record produced by a recording wattmeter. The area under the load record over a period of time is the energy measurement.

Electricity meters. These are the most common devices for measuring the vast quantities of electrical energy used by industry and the general public. The same fundamentals of measurement apply as for electric power measurement, but in addition the electric meter provides the time-integrating means necessary to electric energy measurement.

A single meter is sometimes used to measure the energy consumed in two or more circuits. However, multistator meters are generally required for this purpose. Totalization is also accomplished with fair accuracy, if the power is from the same voltage source, by paralleling secondaries of instrument current transformers of the same ratio at the meter. Errors can result through unbalanced loading or use of transformers with dissimilar characteristics.

Watt-hour meters are generally connected to measure the losses of their respective current circuits. These losses are extremely small compared to the total energy being measured and are present only under load conditions.

Other errors result from the ratio and phase angle errors in instrument transformers. With modern transformers these errors can generally be neglected for commercial metering. If considered of sufficient importance, they can usually be compensated for in adjusting the calibration of the watt-hour meter. For particularly accurate measurements of energy over short periods of time, portable standard watt-hour meters may be used. Errors may also arise due to integral-cycle control. This is a well-established method of power control in which a silicon-controlled rectifier circuit acts to turn the power on for several cycles and off for a different number of cycles. The main source of error in the induction meter is the difference between the mechanical time-constants for two on-off intervals, making the meter read high.

Watt-hour meters used for the billing of residential, commercial, and industrial loads are highly developed devices. Over the last years many significant improvements have been made, including improvements in bearings, insulating materials, mechanical construction, and sealing techniques which exclude dust and other foreign material. *See* WATT-HOUR METER.

Automatic remote reading. Various aspects of the energy crisis spurred active development of automatic meter-reading systems with the functional

capability of providing meter data for proposed new rate structures, for example, time-of-day pricing, and initiating control of residential loads, such as electric hot-water heaters.

Automatic meter-reading systems under development generally consist of a utility-operated, minicomputer-controlled reading center, which initiates and transmits commands over a communication system to a terminal at each residential meter. The terminal carries out the command, sending the meter reading back to the reading center, or activating the control of a residential load.

Several communication media have been proposed and tested by system developers, including radio, CATV, use of the existing subscriber phone lines, and communication over the electric distribution system itself.

Quantities other than watt-hours. Included in the field of electrical energy measurement are demand, var hours, and volt-ampere hours.

Demand. The American National Standards Institute defines the demand for an installation or system as "the load which is drawn from the source of supply at the receiving terminals, averaged over a suitable and specified interval of time. Demand is expressed in kilowatts, kilovolt-amperes, amperes, kilovars and other suitable units" (ANSI C12.1-1988).

This measurement provides the user with information as to the loading pattern or the maximum loading of equipments rather than the average loading recorded by the watt-hour meter. It is used by the utilities as a rate structure tool.

Var hour. ANSI defines the var hour (reactive volt-ampere hour) as the "unit for expressing the integral, of reactive power in vars over an interval of time expressed in hours" (ANSI C12.1-1988).

This measurement is generally made by using reactors or phase-shifting transformers to supply to conventional meters a voltage equal to, but in quadrature with, the line voltage.

Volt-ampere hour. This is the unit for expressing the integral of apparent power in volt-amperes over an interval of time expressed in hours. Measurement of this unit is more complicated than for active or reactive energy and requires greater compromises in power-factor range, accuracy, or both. Typical methods include: (1) Conventional watt-hour meters with reactors or phase-shifting transformers tapped to provide an in-phase line voltage and current relationship applied to the meter at the mean of the expected range of power-factor variation. (2) A combination of a watt-hour and a var-hour meter mechanically acting on a rotatable sphere to add vectorially watt-hours and var-hours to obtain volt-ampere hours, volt-ampere demand, or both.

Measurement of volt-ampere hours is sometimes preferred over var-hours because it is a more direct measurement and possibly gives a more accurate picture of the average system power factor. This would not necessarily be true, however, where simultaneous active and reactive demand are measured and recorded. *See* ELECTRICAL MEASUREMENTS.

William H. Hartwig

Bibliography. D. A. Bell, *Electronic Instrumentation and Measurement*, 2d ed., 1994; *Code for Electrical Metering*, ANSI C12.11988, 8th ed., 1988; D. Fink and H. W. Beaty (eds.), *Standard Handbook for Electrical Engineers*, 14th ed., 1999; Instrument Society of America, *ISA Standards and Practices for Instrumentation*, 12th ed., 1995; L. Schnell (ed.), *Technology of Electrical Measurements*, 1993; L. M. Thompson, *Electrical Measurements and Calibration*: *Fundamentals and Applications*, 2d ed., 1994.

Electrical engineering

The branch of engineering that deals with electric and magnetic forces and their effects. *See* ELECTRICITY; ENGINEERING; MAGNETISM.

Scope. Electrical engineers design computers and incorporate them into devices and systems. They design two-way communications systems such as telephones and fiber-optic systems, and one-way communications systems such as radio and television, including satellite systems. They design control systems, such as aircraft collision-avoidance systems, and a variety of systems used in medical electronics. Electrical engineers are involved with generation, control, and delivery of electric power to homes, offices, and industry. Electric power lights, heats, and cools working and living space and operates the many devices used in homes and offices. Electrical engineers analyze and interpret computer-aided tomography data (CAT scans), seismic data from earthquakes and well drilling, and data from space probes, voice synthesizers, and handwriting recognition. They design systems that educate and entertain, such as computers and computer networks, compact-disk players, and multimedia systems. *See* CHARACTER RECOGNITION; COMMUNICATIONS SATELLITE; COMPACT DISK; COMPUTER; COMPUTERIZED TOMOGRAPHY; CONTROL SYSTEMS; DIGITAL COMPUTER; ELECTRIC HEATING; ELECTRIC POWER GENERATION; ELECTRIC POWER SYSTEMS; ELECTRIC POWER TRANSMISSION; ELECTRICAL COMMUNICATIONS; ILLUMINATION; MULTIMEDIA TECHNOLOGY; OPTICAL COMMUNICATIONS; OPTICAL FIBERS; RADIO; TELEPHONE SERVICE; TELEVISION; VOICE RESPONSE.

Early development. Early experimenters developed batteries, electromagnets, and electric motors. In 1832, M. Faraday showed that electric and magnetic effects are not separate, and combined the results of various observations into a single theory of electromagnetic induction. The first major commercial application of electromagnetism was the telegraph, demonstrated in 1844. *See* BATTERY; ELECTROMAGNET; ELECTROMAGNETIC INDUCTION; MOTOR; TELEGRAPHY.

The last quarter of the nineteenth century saw rapid developments in lighting, stemming from the invention of the incandescent lamp, and in electric power generation and delivery. During this period, lighting moved to homes, offices, and schools, and motors powered new forms of transportation and began to transform factories. A major technical

debate in this period was whether to generate and distribute electric energy as direct current or alternating current. After the invention of the transformer in 1883, alternating-current generation and delivery quickly became the dominant system, because of its greater efficiency and versatility. *See* ALTERNATING CURRENT; DIRECT CURRENT; ELECTRIC ROTATING MACHINERY; INCANDESCENT LAMP; TRANSFORMER.

The latter nineteenth and early twentieth centuries saw the development of the telephone and wireless telegraphy. Telegraph wires provided communications for railroads and nearly instantaneous message capability for business, industry, and personal data, while the telephone industry grew rapidly. Wireless telegraphy linked continents with high-powered spark transmitters.

Electronics. The audion, invented in 1906 and later called the triode, was a vacuum tube (or valve) with three electrical elements. The triode was used to design and build amplifiers, a basic element in modern electronics. Amplifiers enable a weak information signal to control a local power source, yielding a more powerful output. Amplification and other applications of tubes enabled the rapid development of the electronics industry. The junction transistor, whose invention in 1948 followed that of the point-contact transistor in the previous year, became the dominant electronic device by the 1960s, replacing vacuum tubes. *See* AMPLIFIER; TRANSISTOR; VACUUM TUBE.

The integrated circuit, invented in 1959, quickly revolutionized electronics. This device or process enables designers to put many transistors and associated components into a single package. The number of electronic devices in a single integrated-circuit chip now routinely exceeds 1 million. These units have made the personal computer a reality, and are installed in home appliances, cars, offices, and factories. *See* ELECTRONICS; EMBEDDED SYSTEMS; INTEGRATED CIRCUITS; MICROCOMPUTER; MICROPROCESSOR.

Feedback. Electronic devices can produce, along with amplification, distorted output signals. An invention of the 1920s, called negative feedback, reduces this problem. A fraction of the amplifier output signal is subtracted from the input signal. While this subtraction reduces gain and signal level, it also reduces distortion. The idea illustrates the trade-offs that must be made throughout the design process, as competing requirements are balanced. While positive feedback can sometimes find application, it is often a serious problem in electronics. Positive feedback is responsible for the screeching sound in a public address system when a person inadvertently points a microphone at a loudspeaker. *See* DISTORTION (ELECTRONIC CIRCUITS); FEEDBACK CIRCUIT.

Control systems. Control systems use feedback. Common examples are thermostats that control furnaces or air conditioners. Automobile cruise controls and automatic pilots for aircraft are applications. Other examples are found in chemical processes and steel-rolling mills. In control systems, electronic circuits enable desired results to be compared with actual results, and the feedback enables the reduction of the error to nearly zero. Electrical engineers work with other engineers in the design and development of control systems. *See* AUTOPILOT; PROCESS CONTROL; THERMOSTAT.

Computers. Electrical engineers have a major role in the continuing development and application of computers, both large and small. They work with computer engineers, computer scientists, and other professionals. Electrical engineers design the data-storage systems, commonly called disks (floppy or hard), the central processing unit (CPU), and the display. The disks are magnetic storage devices, and proper operation of the disks requires a sophisticated control system. The CPU, which is the heart of the computer, includes very large scale integrated (VLSI) circuits for computations, for temporary storage during computation, and for control of printers, disk drives, and displays. The display may be a large vacuum tube with a phosphorescent screen, whose design requires knowledge of electromagnetic theory and materials, or it may be a liquid-crystal display, a technology that has developed rapidly. Computer capability continues to double every 18–24 months. *See* COMPUTER STORAGE TECHNOLOGY; ELECTRONIC DISPLAY.

The computer-communication system known as the internet or worldwide web is a linkage of computers around the world. The internet uses telephone lines, and in turn the telephone system uses satellites, microwave communications, wire communications, and radio. This system allows users to exchange data and ideas rapidly, reliably and at very low cost. *See* DATA COMMUNICATIONS.

Optics. In optical communications, another major contribution of electrical engineering, light pulses and waves carry data over very small strands of glass. The glass replaces copper wire, and the new systems are less expensive, more reliable, and in many respects safer to the user than those they replace. Fiber-optic networks provide two-way video and audio links for universities, colleges, public schools, libraries, hospitals, and court houses.

Challenges. The integration of communications equipment, control systems, computers, and other devices and processes into reliable, easily understood, and practical systems is a major challenge, which has given rise to the discipline of systems engineering. Electrical engineering must respond to numerous demands, including those for more efficient and effective lights and motors; better communications; faster and more reliable transfer of funds, orders, and inventory information in the business world; and the need of medical professionals for access to medical data and advice from all parts of the world. *See* INFORMATION SYSTEMS ENGINEERING; MEDICAL INFORMATION SYSTEMS; SYSTEMS ENGINEERING.

Edwin C. Jones, Jr.

Bibliography. J. E. Brittain (ed.), *Turning Points in American Electrical History*, 1977; D. G. Fink and H. W. Beaty (eds.), *Standard Handbook for Electrical Engineers*, 14th ed., 1999; T. P. Hughes, *Networks of Power*, 1983, reprint 1993; J. D. Irwin and

D. V. Kerns, Jr., *Introduction to Electrical Engineering*, 1995; C. R. Paul, S. A. Nasar, and L. E. Unnewehr, *Introduction to Electrical Engineering*, 2d ed., 1992; J. D. Ryder, *Engineers and Electrons*, 1984.

Electrical impedance

The measure of the opposition that an electrical circuit presents to the passage of a current when a voltage is applied. In quantitative terms, it is the complex ratio of the voltage to the current in an alternating-current (ac) circuit.

The **illustration** shows a generalized ac circuit, which may be composed of the interconnection of various types of circuit elements. If the voltage is $v = V_0 \sin(\omega t + \theta)$ and the current is $i = I_0 \sin(\omega t + \beta)$, then the circuit can be analyzed by considering a complex voltage $V = V_0 e^{j\theta}$ and a complex current $I = I_0 e^{j\beta}$. The impedance of the circuit is then given by Eq. (1), and Z is a complex number given by Eq. (2). R, the real part of the impedance,

$$Z = \frac{V}{I} \tag{1}$$

$$Z = R + jX \tag{2}$$

is the resistance of the circuit, and X, the imaginary part of the impedance, is the reactance of the circuit. The units of impedance are ohms. *See* ELECTRICAL RESISTANCE; REACTANCE.

The modulus of the impedance is given by Eq. (3),

$$|Z| = \sqrt{R^2 + X^2} \tag{3}$$

and this is the ratio of the maximum voltage to the maximum current. The phase angle of the impedance is given by Eq. (4), and this is the angle by

$$\psi = \tan^{-1} \frac{X}{R} \tag{4}$$

which the current waveform lags the voltage waveform. The power factor of the circuit is $\cos \psi$ and is given by Eq. (5). The power dissipated in the circuit is given by Eq. (6).

$$\cos \psi = \frac{R}{\sqrt{R^2 + X^2}} \tag{5}$$

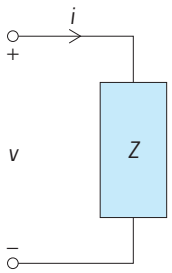
$$P = {}^1/_2 |V| |I| \cos \psi \tag{6}$$

For a circuit composed of resistors, inductors, and capacitors, R is always positive, but X may be positive or negative. For example, in a circuit consisting of a resistor R and inductor L in series, the impedance at frequency ω is given by Eq. (7), so R is greater than

$$Z = R + jL\omega \tag{7}$$

0 and X is greater than 0. For a circuit consisting of a resistor R and capacitor C in series, the impedance at frequency ω is given by Eq. (8) so R is greater than

$$Z = R + \frac{1}{jC\omega} = R - j\frac{1}{C\omega} \tag{8}$$



Generalized alternating-current circuit.

0 and X is less than 0. *See* ALTERNATING-CURRENT CIRCUIT THEORY.

J. O. Scanlan

Electrical instability

A condition that is generic to all physical systems that use the amplification of signals, power, or energy in various forms. Electrical circuits are perhaps the best-known example because the use of amplification is very common there. In order for instability to occur, a closed (circular) path must exist, and the net amplification around it must be sufficiently large that any random signal (for example, noise) that travels around the closed path grows uncontrollably. The condition is eventually limited by some physical aspect of the system such as nonlinearity, saturation, or power available. The energy to create the uncontrollable signal growth is obtained by conversion from some other form of energy. For example, a high-frequency electrical oscillation can draw its energy from a direct- current (dc) power supply. *See* ENERGY; LINEARITY; SATURATION.

There can be various degrees of stability and instability in an electrical circuit or system, and instability may be desirable or undesirable depending on the context. Desirable instability is displayed by a cross-coupled positive-feedback latch in a computer memory cell, or an oscillator that is used to create the high-frequency carrier wave of a radio transmitter. An undesirable instability might arise in the loss of control over a power transmission system due to surges on the generators, leading to a necessary decoupling of the system. Another example is the high-pitched whine heard when a radio-station telephone caller leaves the home radio receiver on and thus allows feedback from the radio receiver to the telephone, that is, a closed-loop signal path. *See* FEEDBACK CIRCUIT; OSCILLATOR.

Yet another example is a computer simulation program that fails because of lack of convergence when trying to solve for the signal levels in a circuit. This could be due to a numerical instability in the computer algorithm. *See* ALGORITHM; SIMULATION.

Electrical instability can occur at zero frequency (dc), for example, in the cross-coupled memory latch, but its accidental occurrences are generally at high frequencies. A configuration may be designed to be stable at low frequencies, but inadvertent phase shifts or time delays at higher frequencies may

create a positive-feedback situation. For this reason, the design of amplifiers, feedback control systems, and so forth necessarily involves much attention to the higher-frequency aspects even though the desired use may be primarily at lower frequencies. *See* AMPLIFIER; CONTROL SYSTEMS.

The intellectual basis for the design of electrical circuits and systems to ensure their stability is quite well developed, based on linear mathematics. Computer tools are extensively used in complex application areas. However, the situation is less satisfactory in the design of deliberately unstable systems, such as oscillators, because their behavior is finally explained by less tangible nonlinear limiting mechanisms. *See* CONTROL SYSTEM STABILITY; LINEAR SYSTEM ANALYSIS.

Miles A. Copeland

Electrical insulation

A nonconducting material that provides electric isolation of two parts at different voltages. To accomplish this, an insulator must meet two primary requirements: it must have an electrical resistivity and a dielectric strength sufficiently high for the given application. The secondary requirements relate to thermal and mechanical properties. Occasionally, tertiary requirements relating to dielectric loss and dielectric constant must also be observed. A complementary requirement is that the required properties not deteriorate in a given environment and desired lifetime. *See* CONDUCTOR (ELECTRICITY).

Electric insulation is generally a vital factor in both the technical and economic feasibility of complex power and electronic systems. The generation and transmission of electric power depend critically upon the performance of electric insulation, and insulation now plays an even more crucial role because of the energy shortage.

Insulation Requirements

The important requirements for good insulation will now be discussed.

Dielectric strength. The basic difference between a conductor and a dielectric is the availability of charge carriers with high mobility in a conductor, whereas free charge has little mobility in a dielectric. Dielectric strength is a measure of the electric field required to cause sudden and substantial motion of charge across the surface or through the bulk of a dielectric. Dielectric strength usually deteriorates with increased moisture content and with elevated temperature. For high-voltage (on the order of kilovolts) applications, dielectric strength is the most important single property of the insulation.

Steady-state strength. The voltage at which a sudden high-current discharge punctures an insulator is called the breakdown voltage. In a uniform field this voltage divided by the electrode gap is called the dielectric strength, the breakdown field strength, the electric strength, or the breakdown stress. In a nonuniform field it is important to differentiate between the maximum field and the average field at breakdown. *See* BREAKDOWN POTENTIAL; ELECTRICAL BREAKDOWN.

An insulator with an operational dielectric strength of 4–12 kV/mm (100–300 V/mil) is considered good. Many of the insulators listed with a dielectric strength more than an order of magnitude higher than this are ordinarily stressed at these lower values in practical operation. The reason is that imperfections in the form of voids, cracks, filamentary defects, and so forth occur whenever long lengths and thick sections of insulating material are manufactured.

Whenever two dielectrics are in series in an alternating electric field, the electric field E is higher in the medium of lower dielectric constant κ by the ratio of dielectric constants, $E_2/E_1 = \kappa_1/\kappa_2$. Therefore the electric field is higher in a cavity, which generally has a dielectric constant of 1, by a factor of 2–3, than it is in the insulator. Furthermore, the dielectric strength is much less in the defect than in the insulator (**Table 1**; **Fig. 1**). Thus, in addition to operating insulators at much lower than intrinsic dielectric strength, schemes such as oil impregnation are used

TABLE 1. Properties of various good insulators

Insulator	Resistivity, 10^{14} ohm-cm	Dielectric strength, 10^3 V/mil*	Power factor $\times 10^{-4}$	Dielectric constant	Tensile strength, 10^3 lb/in.^{2†}	Chemical resistance	Flammability
Lexan	10–1000	8–16	10–30	3	8	Good	Self-extinguishing
Kapton H	1000–10,000	3–8	20–50	3	22	High	Self-extinguishing
Kel-F	10–1000	2–6	20–40	2–3	5	High	No
Mylar	0.1–1	4–16	30–120	3	20	Good	Yes
Parylene	10–10,000	6–10	2	2–3	10	Very good	Yes
Polyethylene N	100–500	1–17	3–30	2	4–6	Good	Yes
Polytetrafluoroethylene	0.1–1000	1–7	1–4	2	4–5	High	Very low
Air (1 atm, 1 mm gap)	—	0.1	10^{-3}	1	0	Stable	No
Sulfur hexafluoride (1 atm, 1 mm gap)	—	0.2–0.3	10^{-3}	1	0	Stable	No
Vacuum ($<10^{-5}$ torr or 1.3×10^{-3} Pa, 1 mm gap)	—	2–3	0	1	0	Stable	No

*10^3 V/mil = 4×10^5 V/cm.
†10^3 lb/in.^{2} = 6.9 (MPa).

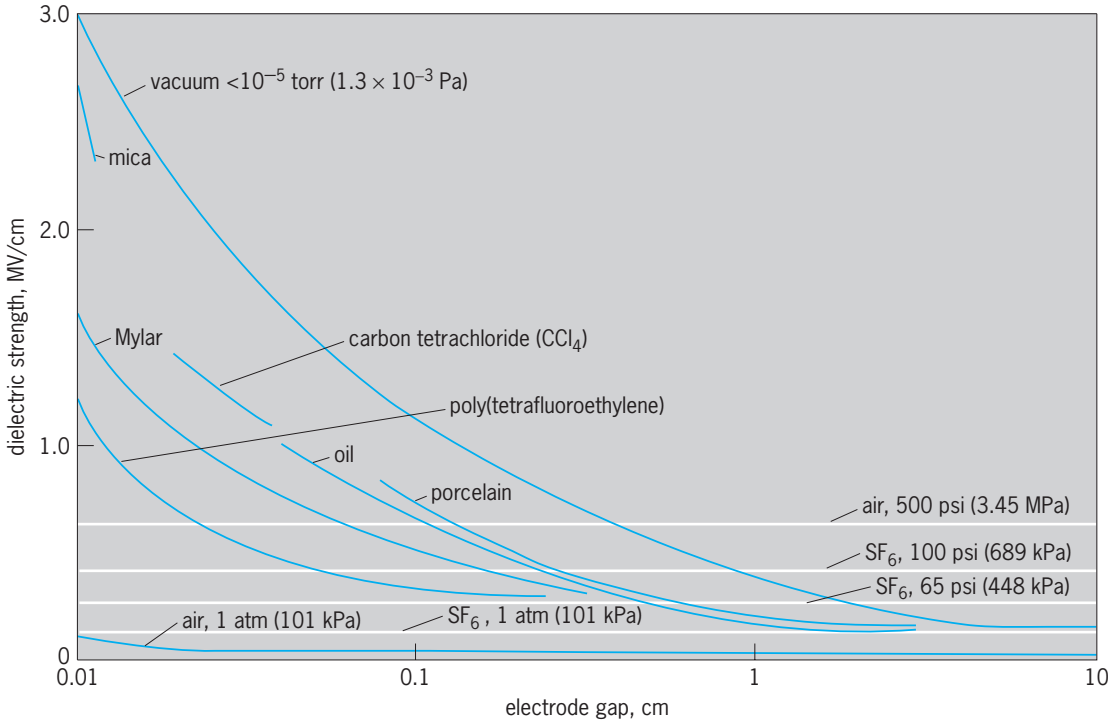


Fig. 1. Uniform field dielectric strength for solid, liquid, gas, and vacuum insulation.

when possible to fill the defects to increase both their dielectric constant and their dielectric strength.

Although solid extruded dielectric cable is now employed up to at least 420 kV system voltage and operates at average fields in the range of 15 kV/mm, oil-impregnated-paper or paper-polypropylene laminate (PPLP) wound-tape insulation can be used at even higher voltages. The probability that defects in the tape will line up in proximity is much less than in a solid piece of the same insulation material. The tape thickness also limits the size of the defect in the direction of the electric field. *See* DIELECTRIC MATERIALS.

Laboratory testing of insulation over a wide range of conditions is conducted to determine the effects of aging, high temperature, heat cycling, ultraviolet radiation, moisture, and so forth. A statistical analysis of the data based upon the weak-link Weibull statistics is then used to determine safe operating voltages and electric fields for the required lifetimes and environments. Operational withstand stress is sometimes arbitrarily defined as being at least two standard deviations below the average dielectric strength in the laboratory.

Impulse strength. Impulse strength is a measure of an insulator's ability to withstand a rapid rise and rapid fall (measured in microseconds) of electric stress. In addition to steady-state dielectric strength, an insulator must be able to meet a basic impulse level (BIL) and a basic switching level (BSL) surge (not as high a rate of voltage increase as for BIL) requirement. The impulse breakdown electric field strength is greatest for solids, followed by liquids, gases, and vacuum. For example, a development in tape insulation consisting of a combination of cellulose paper bound to polypropylene film or fibers [known as polypropylene laminated paper (PPLP)] and impregnated with oil has an impulse strength of 2.7 MV/cm, whereas the impulse strength of sulfur hexafluoride from 100 to 400 kilopascals (15 to 60 lb/in.^{2}) goes only from 0.2 to 0.45 MV/cm. Thus, in designing transmission cables, sufur hexafluoride cables are limited by their BIL, whereas most solid insulation cables are limited only by their steady-state dielectric strength. This is why solid insulation cables are so much more compact than sulfur hexafluoride cables for the same ratings, in addition to the fact that they have higher steady-state dielectric strength. However, when a solid dielectric breaks down electrically, the punctured and surrounding regions must be replaced. After sulfur hexafluoride breaks down, the gas is self-healing and recovers dielectric strength. If no damage has been incurred by the conductors, operation may resume.

The impulse strength for sulfur hexafluoride increases with increasing pressure but tends to level off above 400 kPa (60 lb/in.^{2}). The impulse strength of liquids increases less rapidly with increasing pressure, leveling off above 2 megapascals (300 lb/in.^{2}). For example, carbon tetrachloride's impulse strength goes from 2.2 to 2.6 MV/cm in going from 100 kPa to 2 MPa (15 to 300 lb/in.^{2}). This pressure dependence for liquids tends to disappear for rates of field increase greater than 4 MV/cm-μs. This is experimental evidence that a "bubble" mechanism (see below) cannot be operative in such short times.

Resistivity. Resistivity is a measure of how much current will be drained away from the conductor through the bulk (bulk resistivity) or along the surface (surface resistivity) of the dielectric.

An insulator with resistivity equal to or greater than 10^{13} ohm-cm may be considered good. While resistivity measurements of conductors are straightforward, measurements on insulators are much more difficult and are subject to a much wider range of values for a given material. A dc measurement of current is made after application of a uniform electric field for a specified period of time. A time variation from 1 s to 1 min can lead to orders-of-magnitude variation in the measured resistivity, and the "true" resistivity can be nearly impossible to determine.

Since a resistivity measurement is nondestructive, it can be used as an on-line quality and uniformity check during manufacturing processes. During actual operation in the field, a sharp decrease in resistivity usually signals an ensuing service failure. *See* ELECTRICAL RESISTIVITY; INSULATION RESISTANCE TESTING; NONDESTRUCTIVE EVALUATION.

Dielectric loss. When a dielectric is subjected to an alternating field, a time-varying polarization of the atoms and molecules in the dielectric is produced. The response time of polarization phenomena varies from nearly instantaneous for electronic polarization to seconds or even hours for some forms of orientational polarization. If the time constant for polarization results in a lag between the applied field and polarization, a loss results, of which the power factor is a measure. The importance of the power factor varies with application and tends to become more important at high temperatures, where the loss tends to increase and can cause thermal runaway, and in cryogenic systems, where the the energy required to remove the heat that results from dielectric loss can be great.

Dielectric loss is a function of both temperature and frequency and generally decreases as these two variables decrease. However, it is not a monotonic function of these parameters, and at various discrete values of temperature and frequency quasiresonances occur where the loss can be large. A sharp rise in dielectric loss, usually accompanied by a rapid temperature increase, is an indication of impending breakdown.

Neglecting these high-loss sites, **Table 2** indicates the power factor in the three temperature regions of general interest, 300 K or 80°F (conventional), 77 K or −321°F (cryoresistive), and 4 K or −452°F (superconducting). Values of less than 10^{-3} are good at 300 K, values less than 10^{-4} are good at 77 K, and values less than 10^{-5} are good at 4 K. With the increasing need to deliver large blocks of power (greater than 3000 MW) for long distances (greater than 80 km or 48 mi) from remote energy sources, advanced power transmission must have insulation capable of meeting cryogenic requirements. A power factor of less than 10^{-5} at the operating cryogenic temperature is generally desirable, since the overall power loss is amplified by the refrigeration inefficiency, which may require as much as 500 W per watt of dissipation. *See* SUPERCONDUCTING DEVICES.

TABLE 2. Power factor of insulators at three temperature

Insulator	Power factor × 10^{-4} 300 K (80°F)	77 K (−321°F)	4 K (−452°F)
Kapton H	20–50	5	0.5
Polyethylene (low density)	3–30	0.1	0.02
Polytetrafluoroethylene	1–4	0.9	0.02
Kraft paper	50	20	7
Vacuum (<10^{-5} torr or 1.3×10^{-3} Pa)	0	0	0

Ideally the power factor should be independent of electric field stress. However, owing to processes unrelated to polarization, such as electron emission and ionization, the dielectric loss tends to increase with increasing electric field, and this is reflected in an increasing power factor.

The power factor is the cosine of the phase angle ϕ between the voltage and current, or equivalently the ratio of resistance to impedance for the medium. Occasionally confusion arises regarding terminology when the dielectric loss is related to the complementary angle δ, which satisfies Eq. (1) for small δ. Thus

$$\text{Power factor} = \cos\phi = \sin\delta \doteq \tan\delta \doteq \delta \qquad (1)$$

the power factor is often used interchangeably with the terms loss tangent and loss angle.

Dielectric constant. The dielectric constant, also known as the relative permittivity or specific inductive capacity (SIC), is a measure of the ability of the dielectric to become polarized, taken as the ratio of the charge required to bring the system to the same voltage level relative to the charge required if the dielectric were vacuum. It is thus a pure number, but is in fact not a constant, and may vary with temperature, frequency, and electric-field intensity. For highly polar materials such as water, the variation with temperature and frequency is dramatic, going from 80 at 300 K (80°F) to 3 at 77 K (−321°F), or from 80 at 60 Hz to 20 at 10 kHz.

In addition to the problem of intensification of the electric field in regions of relatively lower dielectric constant, a low-dielectric-constant insulation is desirable for two more reasons. In ac transmission cables, the lower the dielectric constant, the more the current and the voltage will be in phase. Thus, more usable power will be delivered, without the need for reactive compensation. Furthermore, in reducing the charging current (which is proportional to the dielectric constant), concomitant power losses (related to the square of the charging current) are also reduced. A high dielectric constant is desirable in capacitors, since the capacitance is proportional to it. *See* CAPACITANCE; DIELECTRIC MEASUREMENTS; PERMITTIVITY.

Dielectric absorption. When high voltage is applied to a solid dielectric, charge tends to be absorbed into the dielectric. When the voltage is removed, some of this change comes back out of the dielectric over time. This property is known as dielectric absorption. For example, if high voltage is applied to a polymer film capacitor for several minutes, the

capacitor is shorted for a few seconds, and then the capacitor is left open-circuit, a voltage will appear across the capacitor terminals as a result of the absorption charge which evolves from the capacitor film. A "perfect" capacitor, such as a vacuum capacitor, would have no dielectric absorption. Many capacitors used in electrical or electronic applications, including most high-voltage capacitors, are based on polymer films and do exhibit dielectric absorption. Poly(tetraflyoroethylene) [PTFE] and polystyrene capacitor films exhibit among the lowest dielectric absorption.

Insulator Properties

Table 1 summarizes the properties of various good insulators at or near 60 Hz and ambient temperature (about 300 K or 80°F). The range of values indicates not only differences between measurements of different investigators, but differences resulting from small variations in a parameter. For example, the dielectric strength depends on the thickness and area of the sample being tested. It generally decreases with increasing thickness and area. The voltage at which electrical breakdown occurs increases less than linearly with thickness. Thus the corresponding value of electric field decreases.

All insulators may be classified as either solid or fluid. Solid insulation is further divided into flexible and rigid types.

Flexible insulation. Flexible hydrocarbon insulation is generally either thermoplastic or thermosetting. Thermosets are initially soft, and can be extruded by using only pressure. Following heat treatment, when they return to ambient temperature, they are tougher and harder. After thermosetting, nonrubber thermosets are harder, stronger, and have more dimensional stability than the thermoplastics. Thermoplastics are softened by heating, and when cool become hard again. They are heat-extruded.

Thermoplastics. The first seven insulators listed in Table 1 are thermoplastics, as are nylon, polystyrene, and poly(vinyl chloride). Those listed in Table 1 are solid polymers that are used at ambient temperature and have good potential for cryogenic applications, and their monomeric structures are shown in **Fig. 2**. Polymers come in amorphous and ordered states which have different dielectric properties. The ordered state is frequently called "crystalline," but is not the same as the crystalline state of solids. *See* POLYMER.

Lexan is a trade name for a group of polyesters formed from carbonic acid, and generically called polycarbonate (PC). Polycarbonate has good electrical and mechanical properties, good dimensional stability, good resistance to creep, high distortion temperature, and ease of molding and extrusion. *See* POLYESTER RESINS.

Kapton H is a trade name for polypyromellitimide (PPMI). PPMI has exceptional resistance to thermal degradation, because it is one of the polyimides that incorporates multiple bonds along the backbone of the chain. It has good mechanical strength and good

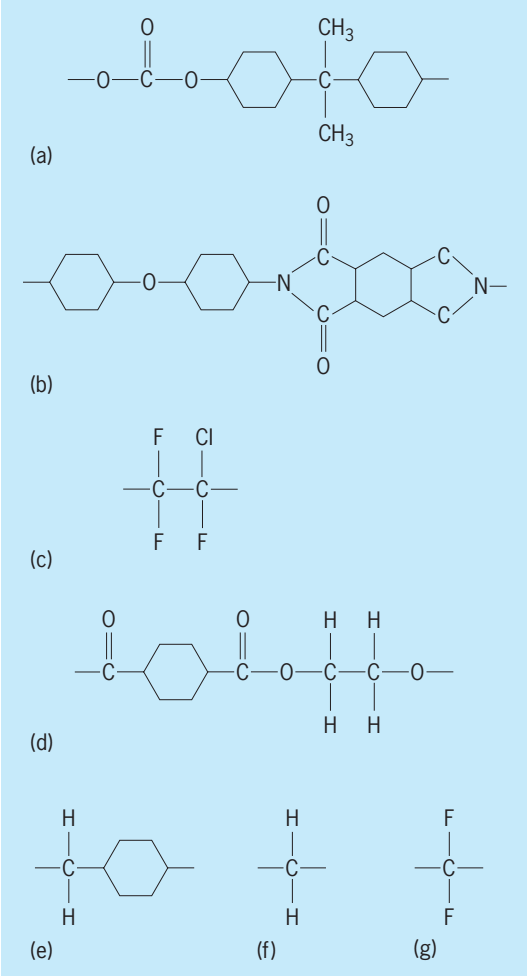


Fig. 2. Monomeric structures of polymers used as electrical insulators. (*a*) Lexan. (*b*) Kapton H. (*c*) Kel-F. (*d*) Mylar. (*e*) Parylene. (*f*) Polyethylene. (*g*) Polytetrafluoroethylene.

dimensional stability at both high and low temperatures, retaining ductility down to cryogenic temperatures. *See* HETEROCYCLIC POLYMER.

Kel-F is a trade name for poly(chlorotrifluoroethylene) [CTFE]. It is one of the best insulators in its ability to exclude moisture, in high-temperature strength, in chemical inertness, and in low coefficient of friction. *See* POLYFLUOROOLEFIN RESINS.

Mylar is a trade name for poly(ethylene terephthalate) [PET]. PET is a high-molecular-weight polyester with a stiff polymer chain and resilient interchain bonds that retain good mechanical properties beyond 150°C (300°F). Because of their toughness, PET films are used not only for electrical insulation, but also for magnetic tapes. *See* MAGNETIC RECORDING.

Parylene is a trade name for poly(*p*-xylene) [PPX]. PPX is linear and highly regular, and was developed to be highly radiation-resistant. It has good dimensional stability down to cryogenic temperatures. It has not had widespread use owing to its cost. *See* POLY-P-XYLYLENE RESINS.

Polyethylene (PE) is a widely used inexpensive electrical insulator with good mechanical and

excellent dielectric properties. Polyethylene can be cross-linked chemically or by irradiation to increase its mechanical strength and improve its high-temperature properties. Chemical cross linking leaves impurities which increase loss. Electron-beam cross linking results in lower loss. Cross-linked polyethylene (XLPE) power cables are manufactured for up to 420 kV system voltage using chemical cross linking, while XLPE high-frequency signal coaxial cables are made using electron-beam cross linking to reduce high-frequency loss. Although polyethylene has a higher dielectric strength than chemically cross-linked XLPE, its low melting point of 118°C (244°F) makes it less desirable for extruded dielectric cables than XLPE. Even after cross linking, melting of crystallites in XLPE results in a precipitous drop in yield stress above 90°C (194°F) and very large thermal expansion in the same temperature range. These characteristics make implementation of splices and terminations problematic for XLPE cables rated to 105°C (221°F). Ethylene propylene rubber (EPR) with inorganic filler is also be used in extruded solid dielectric power cables. Since the EPR polymer is nearly amorphous and the compound gains its physical properties from the inorganic filler, the mechanical and thermal properties of EPR cable dielectric are much more stable with temperature, both low and high, than XLPE, to the point that EPR cable can be rated to operate at 140°C (284°F). *See* POLYOLEFIN RESINS.

Polytetrafluoroethylene and its copolymer with hexafluoropropylene are highly ordered and orientable polymers made of chains of CF_2 with its strong C-F bond. This is why their structure involves little or no branching or cross linking. Thus this polymer has a low coefficient of friction, is stiff, inert, and insoluble, and has a high melting point of 327°C (621°F). Because of its high symmetry and tight bonding, it has one of the lowest known power factors and dielectric absorption for a solid, as well as other excellent dielectric properties. It may be processed by powder and sinter techniques as well as by extrusion or injection molding.

Polystyrene is a good insulator that is used in thin films or where it will not be stressed or bent, because it is stiff and brittle when it is thick. Fillers in the form of silica, alumina, mica, and so forth reduce the brittleness (probably by inhibiting crack propagation), improve the thermal conductivity, and reduce thermal expansion. Hydrated fillers, such as aluminum trihydrate, inhibit tracking, since the waters of hydration tend to be arc-quenching. Polystyrene capacitor film has among the lowest loss ($\sim 10^{-5}$) and dielectric absorption of all capacitor films. *See* POLYSTYRENE RESIN.

Thermosets. The various rubber compounds, neoprene, and epoxies are thermosets. Natural rubber, butyl rubber, buna S rubber, and silicone rubber are good insulators that are tough and flexible. Although chemically and mechanically tough, neoprene does not have good dielectric properties. The epoxies are not very flexible but are used at joints of flexible insulation because of their good bonding, electrical, and mechanical properties. *See* POLYETHER RESINS; RUBBER.

Cellulose paper. Cellulose paper insulation is neither thermoplastic nor thermosetting. It is widely used in cables and rotating machinery in multilayers and impregnated with oil. It has a relatively high dielectric loss that hardly decreases with decreasing temperature (Table 2), which rules it out for cryogenic applications. Oil-impregnated cellulose paper has very high dielectric strength perpendicular to the paper (through the layers of oil-impregnated paper) but relatively low dielectric strength parallel to the paper surface. Because of its high dielectric strength, the high loss has not been a deterrent to its use in conventional ambient-temperature applications. However, the high dielectric strength deteriorates quickly if moisture permeates the paper. *See* CELLULOSE; PAPER.

Rigid insulation. This second major group of solid insulators includes glass, mica, epoxies, ceramoplastics, porcelain, alumina, other ceramics, and fiber-reinforced plastic (FRP). Rather than being used to insulate wires and cables except for mica, these materials are used in equipment terminations (potheads) and as support insulators (in tension or compression) for overhead lines whose primary dielectric is air. These rigid structures must be shock-resistant, be relatively water-impervious, and be able to endure corona discharges over their surfaces. *See* CORONA DISCHARGE.

For use as termination insulators, such as for transformers, underground transmission lines, reactors, and generators, both high bulk dielectric strength and high surface dielectric strength are important. For use in connection with equipment such as overhead transmission lines, the requirement for bulk dielectric strength is less severe, and the emphasis is on the surface. Surface dielectric strength of several kilovolts per inch is adequate, and the creepage paths are usually designed for approximately 400 V/cm (1 kV/in.).

In addition to normal electromagnetic operating forces and windage stress, these insulators must be able to withstand switching surges and shocks from short-curcuit currents. These fault currents may be 10 times higher than the normal current, resulting in forces that are 100 times higher, since the force is proportional to the square of the current.

Porcelain is widely used for such applications. It is a hard, brittle material made by firing feldspar, quartz, clay, and other minerals at high temperature. Glass is a completely amorphous material containing 50–90% silica. Glass has a dielectric strength of about 2 MV/cm at 300 K (80°F). Mica is a mineral that occurs naturally in a laminated form. For thin sections of 0.002 cm thickness, mica has a dielectric strength as high as 10 MV/cm perpendicular to the direction of lamination. Fiber-reinforced plastics, often composites of glass fiber and epoxy, covered with silicone rubber are replacing glass and porcelain in many applications, including transmission-line insulators and high-voltage terminations. This structure has the advantages of much lower weight, improved

performance under polluted conditions as a result of the silicone rubber exterior insulation, and being explosion-proof when operated under pressure. *See* GLASS; MICA; PORCELAIN.

Fluid insulation. Liquids, gases, and vacuum fall in the category of fluid insulation. For all of these, the electrical structure must be such as to contain the fluid in the regions of high electric stress.

Liquids. The main types of insulating liquids are the mineral oils, silicones, chlorinated hydrocarbons, and the fluorocarbons with dielectric strengths on the order of megavolts per centimeter. Many other liquids also have good dielectric strength, such as carbon tetrachloride, toluene, hexane, benzene, chlorobenzene, alcohol, and even deionized water. However, special problems with liquids, such as chemical activity, high thermal expansion, thermal instability, low boiling point, and tracking (conducting carbon residue) after arcing, rule out many liquids. Even transformer oil in combination with the insulating paper in activated transformers produces water which must be filtered out.

Gases. Most gases have a dielectric constant of about 1, and low dielectric loss. **Table 3** compares the relative dielectric strengths of various gases. Those containing fluorine or chlorine are strongly electronegative. Of course, the temperature at which the gas condenses out of the gaseous state is an important consideration, as well as chemical stability, tracking, reactivity, and cost.

Air is used as a dielectric in a wide variety of applications, ranging from electronics to high-voltage (765-kV) and high-power (2000-MW) electric transmission lines. Dry air is a reasonably good insulator (Table 1). However, its dielectric strength decreases with increasing gap (Fig. 1).

Sulfur hexafluoride (SF_6; Table 1) has, at 1 atm (101 kPa), about three times the dielectric strength of air, carbon dioxide, or nitrogen, with this difference increasing with pressure and gap in uniform electric fields (Fig. 1). Sulfur hexafluoride is an electronegative gas which impedes electrical breakdown by capturing free electrons and forming negative ions. Not only does sulfur hexafluoride possess a high dielectric strength, but also, as a result of very rapid recombination deionization after being decomposed in an arc, it recovers its dielectric strength after arcing much more rapidly than air. This property makes it 100 times as effective as air and other gases for use in power circuit breakers. It also is used frequently in electrical utility substations as an insulating medium.

TABLE 3. Comparison of various gaseous dielectrics

Gas	Condensation temperature (at 1 atm or 101 kPa), °C (°F)	Dielectric strength relative to nitrogen
N_2	−195.8 (−320.4)	1.0
CO_2	−78 (−108)	0.9
SF_6	−63.8 (−82.8)	2.5–3
CCl_4	76 (169)	6.3
$CFCl_3$	23.8 (74.8)	3–4.5
CF_2Cl_2	−29.8 (−21.6)	2.4–2.5
CF_2ClCF_2Cl	3.6 (38.5)	2.8
CF_3CN	63 (−81)	4
C_2F_5CN	−30 (−22)	5
C_3F_7CN	1 (34)	6
C_3F_8	−36.7 (−34.1)	2–2.9
C_4F_6	−5 (23)	4
C_4F_{10}	−2 (28)	2.5
C_5F_8	25 (77)	6
C_5F_{10}	22 (72)	>4
He	−269 (−452)	<0.2

Vacuum. Vacuum (that is, pressures of less than 10^{-3} Pa or 10^{-5} torr) has one of the highest dielectric strengths in gaps ranging from 0.1 to 1 mm (0.004 to 0.04 in.; Fig. 1; Table 1). However, as the gap increases, its dielectric strength decreases rapidly. A perfect vacuum should be a perfect insulator, since there would be no charge carriers present to contribute to electrical conductance. The limitation of vacuum as a dielectric results from the effects of a high electric field at the surface of electrodes in vacuum, which can cause field emission of electrons, rather than because a perfect vacuum is far from being realized in the laboratory.

The dielectric properties of vacuum can degenerate rapidly because vacuum offers no resistance to the motion of charge carriers once they are introduced into the vacuum region. The collision mean free path is of the order of meters at pressures less than 10^{-2} Pa (10^{-4} torr), and this is below the pressure range for a gas discharge. However, avalanche processes occur at the electrodes, and vapor from the electrodes becomes the arcing medium.

Vacuum has found increasing applications as an insulator in electron tubes, photocells, high-frequency capacitors, electron microscopes, particle accelerators, distribution voltage (∼15 kV) circuit breakers, and so forth. It is used as an arcing medium with exceptional rate of recovery of dielectric strength. The realization of its full potential is strongly linked to the constitution of the electrodes.

Nature of Electrical Breakdown

Electrical breakdown in gases first began to be understood in 1889, and breakdown in solids in 1935. Breakdown in liquids is less well understood. Breakdown in vacuum (in terms of an encompassing predictive theory) is least well understood, despite experimental investigations going back to 1897. *See* ELECTRICAL BREAKDOWN.

Gases. The Townsend avalanche criterion satisfactorily accounts for the threshold for electrical breakdown in nonelectronegative gases. This theory may be modified to include the process of electron capture by electronegative gases. When the space-charge field becomes large, other mechanisms enter in, such as streamers. At very high pressures, the electric field strength for breakdown is so high that field emission from the electrodes enters into the process.

The Townsend criterion for avalanche is given by Eq. (2). Here α, the first Townsend coefficient, is

$$\gamma(e^{\alpha\delta} - 1) = 1 \tag{2}$$

the number of ionizing collisions per unit length in

the direction of the field made by an electron; and γ, the second Townsend coefficient, is the number of secondary electrons produced at the cathode per electron produced by impact ionization in the gap, δ. Thus the dielectric strength of a gas is a function of both the gas species and the cathode material.

Although impact ionization in the gas produces equal numbers of positive ions and electrons, a space charge would develop even with equal displacement of the negative and positive charges. Since the electrons have greater mobility, this further separates the charges. In those situations when enough positive-ion space charge is produced at the original avalanche head to cause the space-charge field to be almost as large as the applied field, streamer breakdown occurs.

In highly nonuniform fields, where the electric field is very high at one electrode and low at the other, localized breakdown (called corona) which does not bridge the gap can occur. Corona can reduce the electric field around the pointed electrode, in effect rounding it off. This results in a higher breakdown voltage than otherwise, a phenomenon known as corona stabilization.

Most systems in the field become contaminated with conducting particles which greatly reduce the gaseous dielectric strength. This problem is important enough to warrant incorporating particle traps in such equipment, particularly those using electronegative gases. *See* ELECTRICAL CONDUCTION IN GASES.

Solids. For ambient temperature and pressure, the densities of solids are 10^3–10^4 times those of gases. This results in dielectric strengths 10–10^2 times higher than gases. At these higher fields, quantum-mechanical processes such as electron tunneling enter in. Additionally, through excitation of lattice vibration a new mechanism for reducing the energy of the accelerated electrons is provided. *See* LATTICE VIBRATIONS; PHONON; QUANTUM MECHANICS.

For the purpose of considering electrical breakdown, solids may be put in the categories of polar crystals, nonpolar crystals, quasicrystalline (highly ordered), and amorphous. On the basis of only density, the dielectric strength of solids and liquids may be expected to be comparable, and this is so. The alkali halides such as potassium bromide, KBr, and sodium chloride, NaCl, are examples of polar crystals. Mica is an example of a nonpolar crystal. The polymeric solids such as polyethylene can vary between highly ordered and amorphous. Glass is totally amorphous. In their absence of structure, the amorphous materials more nearly resemble and behave like highly viscous liquids. *See* AMORPHOUS SOLID; POLAR MOLECULE.

Above 0 K (-459.67°F) some electrons in the high-energy tail of the electron energy distribution have enough energy to find themselves in the conduction band of an insulator. This is why insulators have a slight conductivity. A. R. Von Hippel assumed that above a critical electric field, these thermally excited electrons would gain energy faster in the conduction band than they would lose it. Thus they would gain enough energy to ionize the solid and cause breakdown. The values of critical field, calculated from this theory, agreed reasonably well with the measured values of the breakdown fields for the alkali halides. *See* BAND THEORY OF SOLIDS.

A simplifying assumption in Von Hippel's treatment is that the electrons in the conduction band all have the same energy. A consequence of this assumption is that since there is no ionization below the critical field, the prebreakdown current can be associated only with these free electrons, implying that (1) the prebreakdown current should be extremely small, (2) the prebreakdown current should be independent of the electric field below the critical field, and (3) the ionization coefficient should be discontinuous—going from zero to a large value at the critical field. Since these predictions are at variance with experiment, new theories were developed. H. Fröhlich assumed an energy distribution for the electrons in the conduction band. C. M. Zener assumed tunneling of electrons into the conduction band because of the high electric field.

Both the Von Hippel and Fröhlich theories neglected interaction of the conduction electrons with the lattice vibrations (phonons). Introducing a further refinement, F. Seitz demonstrated that the phonon interaction is dominant in nonpolar crystals, and is nonnegligible in polar crystals.

Despite the success of the above approaches, there are solids in which other mechanisms also occur, such as thermal, electromechanical, gas-discharge, and electrochemical breakdown. In thermal breakdown, part of the solid reaches a critical temperature at the breakdown field, causing chemical deterioration or melting. Even when thermal breakdown is avoided by using thin specimens, massive electrodes, and pulsed voltages, a gas-discharge mechanism of breakdown may initiate in voids in the specimen, or between the electrodes and the solid. Such discharges may cause treeing (Lichtenberg figures) in the solid, and may chemically and mechanically degrade the solid to the point of breakdown.

Liquids. Strictly speaking, liquids do not have an intrinsic dielectric strength. It is necessary to specify the electrode composition, geometry, and the time duration of the applied field. For very long stress durations, the time is important even for solids. Thus, except for electrode composition, this distinction between liquids and solids may be only a matter of degree. In liquids, microprojections on the cathode and migration of suspended particles (due to the gradient in the field) to the high-field regions contribute to a decrease in dielectric strength. Dissociation of the liquid by field-emitted electrons, electrochemical processes, and thermal bubble formation further complicate electrical breakdown in liquids.

A conditioning effect occurs, also common to vacuum, in which the breakdown strength increases progressively with the first few breakdowns, provided that the breakdown current is limited. Other commonalities with vacuum are curvature and area effects. Contrary to expectation, up to a point the dielectric strength is higher for electrodes of smaller

radius of curvature, and higher for electrodes of smaller area.

The fact that the breakdown strength increases with pressure is indicative that a bubble is formed during the breakdown process. However, if the voltage is applied in the nanosecond range, the pressure dependence vanishes and another breakdown mechanism becomes dominant. A piece of evidence in support of the bubble mechanism of breakdown for longer stress duration is the direct relationship between the boiling point and the breakdown strength for the *n*-alkanes.

Vacuum. As pointed out above, breakdown in vacuum has a number of features in common with liquid breakdown. L. Cranberg suggested that breakdown is initiated when a charged clump of loosely adhering material is removed from one electrode surface by the electric field, strikes the opposite electrode, and thus causes a high enough temperature to produce local evaporation. A discharge then ensues in the metal vapor. (The clump here is analogous to particle-initiated gas or liquid breakdown.)

Cranberg's model predicts that $VE > K$, where V is the breakdown voltage, E is the macroscopic electric field at the electrode where the clump originates, and K is a constant characteristic of a given pair of electrodes. For a uniform field and gap d, this implies that V is proportional to $d^{1/2}$, which agrees reasonably well with experiment. However, it does not predict the area effect, nor the polarity effect where the breakdown voltage is much higher when the pointed electrode is the anode. Moreover, it incorrectly predicts the curvature effect.

L. B. Snoddy was the first to suggest anode vaporization as a result of electron bombardment to account for breakdown. A. J. Ahearn was the first to suggest local heating of the cathode as initiating breakdown. L. C. Van Atta, R. J. Van de Graaff, and H. A. Barton were the first to hypothesize a particle interchange multiplication process. The basic assumption here is that, at a critical voltage, a free charged particle upon striking an electrode produces an avalanche of charged particles by secondary emission, with photoemission also playing a role.

There have since been many variations on these models. However, none has yet correctly predicted (even qualitatively) the area and curvature effects, as well as the gap dependence of approximately $d^{1/2}$. M. Rabinowitz introduced a hypothesis that predicts at least qualitatively the known experimental results, without reference to any particular model in terms of processes that are assumed to occur. He observed that the initiation of vacuum breakdown and gap conduction processes occurs so fast (on the order of 10^{-9} s), compared with the time constants of most breakdown circuits, that only the capacitively stored energy of the electrodes and electrode supports discharges within this time. (Light travels less than 30 cm or 1 ft in 10^{-9} s). His assumption was simply that breakdown cannot occur until there is sufficient stored energy in the electric field in the gap, because only this energy (or a fraction of it) is available to break down the gap.

The energy in the field is the capacitively stored energy, $^1/_2CV^2$. The hypothesis leads to Eq. (3), where

$$W = {}^1/_2 f C V^2 \tag{3}$$

W is the critical energy needed to initiate breakdown (characteristic of the electrodes), C is the capacitance of the electrode system, and f is the fraction of the capacitively stored energy available to produce breakdown, $0 < f < 1$. For a uniform field, this predicts that the breakdown voltage V is proportional to $d^{1/2}$. The curvature and area effects are also qualitatively predicted. Mario Rabinowitz; Steven A. Boggs

Bibliography. R. Bartnikas et al. (eds.), *Engineering Dielectrics*, vol. 1, 1979, vol. 2A, 1983, vol. 2B, 1987, vol. 3, 1993; A. Bradwell et al. (eds.), *Electrical Insulation*, 1983; L. A. Dissado and J. C. Fothergill, *Electrical Degradation and Breakdown in Polymers*, 1992; H. G. Erdman (ed.), *Electrical Insulating Oils*, 1988; D. Kind and H. Karner, *High Voltage Insulation Technology*, 1985; G. G. Raju, *Dielectrics in Electric Field*, 2003.

Electrical interference

Disturbance to the normal or expected operation of electrical or electronic devices, equipment, and systems. Electrical interference is sometimes called radio-frequency interference (RFI) or electromagnetic interference (EMI). Such disturbances may range from nuisances (for example, an electric shaver jamming a nearby broadcast radio receiver) to catastrophes (for example, midair aircraft collision due to EMI navigation errors in holding patterns during bad weather in terminal areas). Electrical noise is a broader term that includes those phenomena that are generally termed electrical interference, but also includes naturally occurring currents or voltages that are more or less continuous and cannot be completely eliminated.

Electrical interference originates from one or more of the following: transmitters such as those used for broadcast, communication, radar, and navigation; artificial incidental emission sources such as from sparking of motor brushes, automotive ignition, and fluorescent lamps; and natural phenomena such as lightning and electrostatic discharges. The interference emissions may be radiated through space or conducted along the paths of electrical wires in power lines and signal cables. *See* COMMUNICATIONS CABLE; ELECTROMAGNETIC RADIATION; TRANSMISSION LINES.

Electrical interference is usually controlled by national regulations, such as those recommended by the International Special Committee on Radio Interference (CISPR), of which there are 25 member nations, including the United States. In the United States, the Federal Communications Commission, through Part 15 (low-level radiators such as computers, wireless microphones, and radio receivers) and Part 18 (industrial, scientific, and medical devices) of its regulations, attempts to mitigate

interference through radiated and conducted emission control. The military sector uses its own standard, MIL-STD-461B, to designate required control of both emission and susceptibility on all purchased equipment and subsystems having electrical or electronic elements. The intelligence agencies regulate allowable emission levels from equipment carrying classified information (known as TEMPEST) through their specifications, NACSIM-5100A.

Excessive electrical disturbance to power mains and through radiation is a form of environmental pollution. Unlike other forms of pollution, however, electrical noise cannot usually be detected with the human senses. Instrumented antennas and conduction probes are used with receivers or spectrum analyzers to detect the presence of emissions and to determine if they might create (or to diagnose) an EMI problem. They are also used for specifications and regulation certification. *See* SPECTRUM ANALYZER.

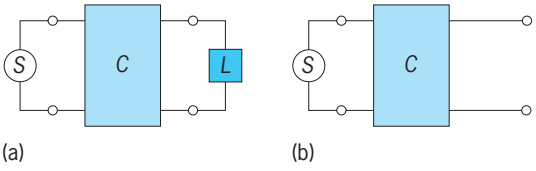
There are several means of mitigating electrical interference, including grounding, bonding, shielding, filtering, and transient control. There are two major approaches. One is to reduce radiation from boxes, cabinets, racks, and consoles by the design of multilayer printed circuit boards and backplanes, or to use metal or metallized plastic cases with EMI-protected apertures. The other involves reduction of radiation to or from the interconnecting cable by using filter pin connectors or foil-wrap shields terminated with complete coverage by the connector backshell at the respective box bulkhead. *See* BONDING; ELECTRIC FILTER; ELECTRIC TRANSIENT; ELECTRICAL NOISE; ELECTRICAL SHIELDING; ELECTROMAGNETIC COMPATIBILITY.

Donald R. J. White

Bibliography. K. Javor, *Introduction to the Control of Electromagnetic Interference*, 1993; M. Mardiguian, *Electromagnetic Control in Components and Devices*, 1971; M. Mardiguian, *EMI Control Methodology and Procedures*, 1989; D. R. J. White, *EMI Control in the Design of Printed Circuit Boards*, vol. 1, 1996.

Electrical loading

The connection of a specific circuit (called the load) to another circuit. **Illustration** *a* shows a source of electrical energy (*S*), a coupling circuit (*C*), and a load (*L*). In power transmission, *S* is the generator of the power plant, *C* represents the connecting transmission lines and transformers, and *L* is, for instance, a household with its electrical appliances. The entire system operates to supply the load demand of *L*, that is, the electrical energy required by the load. In a high-fidelity sound system, *S* is the output of the phonograph cartridge or the tape playback head, *C* is the amplifier, and *L* is the coil of the loudspeaker. *See* ELECTRIC POWER SYSTEMS; SOUND-REPRODUCING SYSTEMS.



Systems (*a*) under load and (*b*) under no load.

When the load is disconnected, the system operates under no-load conditions (illus. *b*).

Loading is used in certain wire-line transmissions in order to improve their characteristics. This is often done by connecting loading inductors along telephone lines, or along certain data transmission circuits. These inductors reduce the attenuation of the line, thus improving the transmission of signals. Waveguides, special antennas, and coaxial cables are often loaded with specific circuits to achieve specific purposes. *See* ANTENNA (ELECTROMAGNETISM); COAXIAL CABLE; INDUCTOR; TRANSMISSION LINES; WAVEGUIDE.

Shlomo Karni

Bibliography. D. J. Fink and H. W. Beaty (eds.), *Standard Handbook for Electrical Engineers*, 14th ed., 1999; M. Kaufman and A. H. Seidman, *Handbook for Electronics Engineering Technicians*, 2d ed., 1985; J. Markus and C. Weston, *Essential Circuits Reference Guide*, 1988.

Electrical measurements

Measurements of the many quantities by which the behavior of electricity is characterized. Such measurements are of great economic importance because of their impact upon industrial, commercial, and scientific activities. Measurements are required of all electrical quantities, possibly extending over a wide dynamic range and at frequencies ranging from 0 to 10^{12} Hz. The International System of Units (SI) is in universal use for all electrical measurements. Electrical measurements are ultimately based on comparisons with realizations, that is, reference standards, of the various SI units. These reference standards are maintained by the National Institute of Standards and Technology in the United States, and by the national standards laboratories of many other countries. Many electrical measurements are made to measure nonelectrical quantities. For example, platinum-resistance-thermometer measurements facilitate accurate temperature measurements, and eddy-current conductivity measurements are made to assess heat treatment of aluminum alloys. *See* ELECTRICAL UNITS AND STANDARDS.

Direct-current measurements. Direct-current (dc) measurements include measurements of resistance, voltage, and current in circuits in which a steady current is maintained.

Resistance measurement. Resistance is defined as the ratio of voltage to current. For many conductors this ratio is nearly constant, but depends to a varying extent on temperature, voltage, and other environmental conditions. The best standard resistors are made from wires of special alloys chosen for low dependence on temperature and for stability. They may have drift rates as low as 0.01 part per million (ppm) per year. *See* ELECTRICAL RESISTANCE; RESISTOR.

The SI unit of resistance, the ohm, is realized by means of a quantized Hall resistance standard. This is based upon the value of the ratio of fundamental constants h/e^2, where h is Planck's constant and e is the charge of the electron, and does not vary with time. Wire-wound standard resistors can be compared with this standard with an uncertainty of about 1 part in 10^8. *See* HALL EFFECT.

For resistance values below about 1 kilohm, the greatest accuracy is achieved by the use of four-terminal measurements, although this introduces considerable complication into resistive bridges.

Resistors in series produce a value equal to their sum, but resistors in parallel produce a resistance whose reciprocal, called its conductance, is equal to the sum of the conductances of the components. For this reason it is sometimes more convenient to use conductance in siemens, rather than the resistance (in ohms); for example, a resistance of 0.1 Ω is equally well expressed as a conductance of 10 S. *See* CONDUCTANCE.

The principal instruments for accurate resistance measurement are bridges derived from the basic four-arm Wheatstone bridge, and resistance boxes. In the Wheatstone bridge the same current is passed through the unknown and reference resistors, and the resistance ratio is effectively deduced from the ratio of potential differences across them. Alternatively, by using a dc current comparator, the currents through the resistors are adjusted so as to produce equal potential differences across them, when the resistance ratio is given by the ratio of currents. The achievement of constant resistance ratios is much easier than that of constant resistance, so that if a bridge is used only to measure the ratio between a standard resistor and an unknown, it need not be constructed of resistors of the high and expensive quality of the standard. When used in this way, many commercial bridges can provide considerably better accuracy than is available by using their built-in reference resistors. *See* WHEATSTONE BRIDGE.

Many multirange digital electronic instruments measure resistance potentiometrically, that is, by measuring the voltage drop across the terminals to which the resistor is connected when a known current is passed through them. The current is then defined by the voltage drop across an internal reference resistor. For high values of resistance, above a megohm, an alternative technique is to measure the integrated current into a capacitor (over a suitably defined time interval) by measuring the final capacitor voltage. Both methods are capable of considerable refinement and extension. *See* OHMMETER; RESISTANCE MEASUREMENT.

Voltage measurement. The SI unit of voltage, the volt, is realized by using arrays of Josephson junctions. This standard is based on frequency and the ratio of fundamental constants e/h, so the accuracy is limited by the measurement of frequency. The secondary standards, which may be either saturated Weston mercury-cadmium standard cells or semiconductor Zener-diode-based electronic voltage standards, can be compared with the Josephson array with uncertainties of about 1 part in 10^8. Josephson arrays can produce voltages between 200 μV and 10 V.

Weston standard cells are sensitive to temperature, vibration, and the passage of current, and have to be kept in temperature-controlled enclosures stabilized to about 1 millikelvin. Electronic standards do not have these problems, but they may be sensitive to humidity and pressure, and their short-term noise and long-term drift in value are both slightly worse than those of standard cells. Voltage standards are usually compared back to back so that the resulting small voltage difference can be measured by a sensitive digital voltmeter which need not be calibrated to such a high accuracy as if the full voltage of the standard had to be measured.

At the highest levels of accuracy, higher voltages are measured potentiometrically, by using a null detector to compare the measured voltage against the voltage drop across a tapping of a resistive divider, which is standardized (in principle) against a standard cell. Dividers intended for this use up to 2 kV are usually called volt ratio boxes, and sometimes contain built-in self-calibration facilities. Resistive dividers for the accurate measurement of still higher voltages, up to about 500 kV, need increasing care to minimize the effects of the high electric fields, which may give errors due to current leakage through corona discharges. These can be avoided by the use of guarding techniques. *See* CORONA DISCHARGE; POTENTIOMETER.

The Zener diode reference standard is the basis for most commercial voltage measuring instruments, voltage standards, and voltage calibrators. The relative insensitivity to vibration and other environmental and transportation effects makes the diodes particularly useful as transfer standards. Under favorable conditions these devices are stable to a few parts per million per year. *See* ZENER DIODE.

Most dc digital voltmeters, which are the instruments in widest use for voltage measurement, are essentially analog-to-digital converters which are standardized by reference to their built-in reference diodes. Some of these instruments may be microprocessor-controlled, in which case they may be remotely or automatically operated, and may include built-in error correction for linearity over the indicated range. The basic range in most digital voltmeters is between 1 and 10 V, near the reference voltage. Other ranges are provided by means of resistive dividers, or amplifiers in which gain is stabilized by feedback resistance ratios. In this way these instruments provide measurements over the approximate range from 10 nanovolts to 10 kV. *See* DIGITAL-TO-ANALOG CONVERTER; MICROPROCESSOR; VOLTAGE MEASUREMENT; VOLTMETER.

Current measurement. The most accurate measurements of direct currents less than about 1 A are made by measuring the voltage across the potential terminals of a resistor when the current is passed through it. It is important that the resistor has been calibrated under the conditions of use, or that allowances are made for the differences. Higher currents, up to about 50 kA, are best measured by means

of a dc current comparator, which accurately provides the ratio of the high current to a much lower one which is measured as above. At lower accuracies, resistive shunts may be used up to about 5000 A, but the effective calibration of such shunts is a difficult process. *See* CURRENT COMPARATOR; CURRENT MEASUREMENT.

Power frequency measurements. The principal needs for accurate measurement at power frequencies (usually 60 or 50 Hz, but possibly up to 35 kHz for lighting systems) arise from the sale of electrical energy. There are also widespread needs for less accurate measurement for the monitoring and control of generators, distribution systems, and plants.

Alternating-current voltage and current measurements. Alternating-current (ac) voltages are established with reference to the dc voltage standards by the use of thermal converters. These are small devices, usually in an evacuated glass envelope, in which the temperature rise of a small heater is compared by means of a thermocouple when the heater is operated sequentially by an alternating voltage and by a reference (dc) voltage. Suitable series resistors, which have been independently established to be free from variation with frequency, permit direct measurement of power frequency voltages up to about 1 kV. Greater accuracy is provided by multijunction (thermocouple) thermal converters, although these are much more difficult and expensive to make. Thermal converters essentially measure the root-mean-square voltage; for a sinusoidal wave the peak value is $(2)^{1/2}$ times the root-mean-square value. Practical wave-forms are usually not sinusoidal, and the relationship between the mean and peak values will depend on the harmonic content or form factor of the waveform. Improvements in digital electronics have led to alternative approaches to ac measurement. For example, a line frequency waveform may be analyzed by using fast sample-and-hold circuits and, in principle, be calibrated relative to a dc reference standard. Also, electronic root-mean-square detectors may now be used instead of thermal converters as the basis of measuring instruments. *See* NONSINUSOIDAL WAVEFORM; THERMAL CONVERTERS.

Voltages above a few hundred volts are usually measured by means of a voltage transformer, which is an accurately wound transformer operating under lightly loaded conditions. These transformers are usually calibrated by comparison with a high-voltage capacitative divider, which can be independently evaluated. This permits the precise measurement of the voltage ratio and phase error of the transformer under its specified conditions of use.

The principal instrument for the comparison and generation of variable alternating voltages below about 1 kV is the inductive voltage divider. In principle, this resembles a series of tapped autotransformers connected to give a multidecade arrangement. Inductive voltage dividers are very accurate and stable devices, and at best provide an uncertainty of division of less than 1 in 10^8. They are widely used as the variable elements in bridges or measurement systems. *See* INDUCTIVE VOLTAGE DIVIDER.

Alternating currents of less than a few amperes are measured by the voltage drop across a resistor, whose phase angle has been established as adequately small by bridge methods. Higher currents are usually measured through the use of current transformers, which are carefully constructed (often toroidal) transformers operating under near-short-circuited conditions. The performance of a current transformer is established by calibration against an ac current comparator, which establishes precise current ratios by the injection of compensating currents to give an exact flux balance. *See* INSTRUMENT TRANSFORMER.

Commercial instruments for measurement of ac quantities are usually dc measuring instruments, giving a reading of the voltage obtained from some form of ac-dc transducer. This may be a thermal converter, or a series of diodes arranged to have a square-law response, in which case the indication is substantially the root-mean-square value. Some lower-grade instruments measure the value of the rectified signal, which is usually more nearly related to the peak value.

Power and energy measurements. The principal complication in power measurement is that the power is the vector product of the voltage V and current I, so a wattmeter is essentially a vector multiplier of these quantities. Although electromechanical instruments such as dynamometers still exist, electronic instruments are now cheaper and more accurate, and are much more widely used, especially where integration into an automated system is required. A long-established type of energy meter consists of a revolving conducting disk driven by the interaction between induced currents generated by voltage and current coils. Such meters are remarkable for their accuracy (particularly at low energy levels), stability, and low cost. Nevertheless, modern electronic energy meters that are competitive with the rotating disk types are available; they offer the advantages of digital indication and the possibility of remote interrogation via power line communications. *See* ELECTRICAL ENERGY MEASUREMENT; WATT-HOUR METER.

The great majority of wattmeters work at levels of current and voltage less than about 10 A and 250 V, single phase. High-power levels such as the output from power stations at several hundred megawatts per phase are measured through the use of current and voltage transformers.

Several wattmeters have appeared in which the multiplication is carried out electronically. In one method a pulse height is determined by the instantaneous value of one variable and the pulse width by the instantaneous value of the other. The integrated energy of the pulse string then gives the power. A second approach is direct simultaneous sampling of the voltage and current, which are then reduced to digital values. The product of each sample pair is stored, and the accumulated total of the series is a

measure of the power. Another approach to power measurement is the use of the quarter-squares principle, using the identity below. In this method the

$$(V + I\cos\theta)^2 - (V - I\cos\theta)^2 = 4\,VI\,\cos\theta$$

outputs of two square-law devices are summed to give a direct power reading. All these approaches have been shown to be capable of producing accuracy comparable to that of the best wattmeters of more conventional design. *See* ELECTRIC POWER MEASUREMENT; WATTMETER.

Audio- and radio-frequency measurements. At the highest level of accuracy, audio-frequency (af) voltages are measured in the same way as at power frequencies, through the use of thermal converters, and employing inductive voltage dividers to provide variable ratios. Such methods provide uncertainties of about 5 ppm at 1 kHz, increasing to about 30 ppm at 1 MHz.

Both pointer and digital instruments for voltage measurement are similar to those in use at power frequencies. Current measurement is needed less often at these frequencies, and generally only small currents (less than about 1 A) are involved. These can readily be measured using a resistive transducer.

Since a typical af electrical component has both reactance and resistance, the ratio of af voltage to current depends upon the impedance of the component. A large number of bridges are available for the measurement of impedance at these frequencies, and most of these are especially suited to the measurement of particular parameters or ranges of values. The most accurate bridges are those based on transformers and inductive dividers, generally at frequencies below 100 kHz, and using coaxial systems throughout. Automatic network analyzers permit measurements to be made with great convenience within the range 10 kHz–40 GHz. Both vector and scalar instruments are available, with the latter being less expensive but not providing phase information. *See* CAPACITANCE MEASUREMENT; INDUCTANCE MEASUREMENT; RADIO-FREQUENCY IMPEDANCE MEASUREMENTS.

For power measurement, the use of dynamometer wattmeters is usually restricted to frequencies below 10 kHz. The range of electronic wattmeters, that is, those performing electronic vector multiplication, may often extend to 1 MHz, but in this frequency range the measurement of power by vector multiplication of voltage and current is superseded by the use of resistive bolometers or calorimeters. These devices provide a response when they are heated by a known fraction of the power to be measured; this approach may be used up to and beyond microwave frequencies. *See* BOLOMETER.

Microwave frequencies. At frequencies above about 300 MHz, where the circuit dimensions become an appreciable fraction of the wavelength, it becomes necessary to use totally enclosed circuits (waveguides), and the measurements of voltage, current, and lumped impedance are replaced by those of wave impedance, reflection and transmission coefficients, voltage standing-wave ratio, attenuation, and power. The measurement techniques, other than for power measurement, are drastically changed. *See* ATTENUATION (ELECTRICITY); MICROWAVE IMPEDANCE MEASUREMENT; MICROWAVE NOISE STANDARDS; MICROWAVE POWER MEASUREMENT.

Fast pulse measurements. The measurement of electrical pulses with rise times of a few picoseconds is of growing importance in the fields of communications, computing, and digital electronics. The most commonly used instrument for measuring repetitive pulse trains is the sampling oscilloscope, and this is calibrated through electrooptical techniques in which short laser pulses are used to study electrical pulses as they travel along transmission lines on electrooptic crystals. In a variant of this, a small electrooptic crystal probe carrying fast laser pulses is brought close to the transmission line carrying the electric pulse, so that the electric field of the pulse affects the optical transmission of the probe and hence modifies the laser pulses. *See* ELECTROOPTICS; OPTICAL PULSES.

Automated measurement systems. There has been a noticeable trend toward the use of automated measurement systems for electrical measurements, facilitated by the readiness with which modern digital electronic instruments may be interfaced with computers. Many of these instruments have built-in microprocessors, which improve their convenience in use, accuracy, and reliability. A built-in reference standard may be used for self-calibration, and the results of many measurements may be stored, processed, and displayed in a variety of ways. A complex measurement system may comprise several such instruments operating under computer control and automatically producing measurement reports. The overall effect is that the performance of accurate measurements becomes less labor-intensive and more cost-effective. R. Gareth Jones; Owen C. Jones

Bibliography. D. A. Bell, *Electronic Instrumentation and Measurements*, 2d ed., 1994; *Code for Electrical Metering*, ANSI C12.1-1995, 1995; C. H. Dix and A. E. Bailey, Electrical standards of measurement, *Proceedings of the IEE*, vol. 122, no. 10R, October 1975; D. Fink and H. W. Beaty (eds.), *Standard Handbook for Electrical Engineers*, 14th ed., 2000; A. Hartland, Quantum standards for electrical units, *Contemp. Phys.*, 29:477–498, 1988; Institute of Measurement and Control, London, *A Guide to Measuring Direct and Alternating Current and Voltage below 1 MHz*, 2003; Institute of Measurement and Control, London, *A Guide to Measuring Resistance and Impedance below 1 MHz*, 1999; Instrument Society of America, *ISA Standards Library for Measurement and Control*, 1999; L. D. Jones and A. F. Chin, *Electronic Instruments and Measurements*, 2d ed., 1991; B. P. Kibble and G. H. Rayner, *Coaxial AC Bridges*, 1984; L. Schnell (ed.), *Technology of Electrical Measurements*, 1993; L. M. Thompson, *Electrical Measurements and Calibration*: *Fundamentals and Applications*, 2d ed., 1994.

Electrical model

A mathematical description or electrical equivalent circuit that represents the behavior of a device or system. Models for complex systems are often represented by networks of models for simpler electrical devices such as resistors, capacitors, transistors, and transformers. By using analogies between current or voltage and other physical parameters, equivalent circuits can also be used to analyze thermal, mechanical, magnetic, and acoustic systems. A recurring issue in development and in application of models is the use of simplifying assumptions to allow a compromise between accuracy and complexity. A hierarchy of models often exists, ranging from highly accurate but complex physics-based nonlinear time-domain computer models to linear equivalent-circuit models suitable for hand calculation. The best model for a particular application is the simplest one that predicts the relevant behavior with acceptable accuracy. *See* EQUIVALENT CIRCUIT; NETWORK THEORY.

Electrical models can be divided into nonlinear, large-signal and linear, small-signal models. Each of these can be further divided into time-invariant and time- or frequency-dependent categories.

Large-signal models. These are usually derived by applying physical laws to generalized devices. Since electromagnetic effects propagate across space at the speed of light, the most detailed and accurate version of these models needs to contain a description of device behavior across space and time. Such "distributed-circuit" models take the form of partial differential equations whose independent variables consist of three-dimensional spatial coordinates and time. Electromagnetic field models of this degree of complexity are required where the dimensions of the circuit are large compared with the wavelength at which the circuit operates, or when electromagnetic wave radiation is of interest. Unfortunately, these equations are difficult to solve, even numerically by computer simulation, and usually need to be simplified. *See* MAXWELL'S EQUATIONS.

At lower operating frequencies, where the physical dimensions of a circuit are a small fraction of a wavelength, spatial variations of electrical quantities can be neglected and a system of ordinary differential equations, with time as the sole independent variable, can model the circuit. In most electrical systems modeled at this "lumped-circuit" level, the variables of interest are the voltages at certain locations (nodes) in the circuit and the currents passing through these nodes. Other variables such as electric charge or magnetic flux also may be of interest. Physical models at this level of detail, without additional approximation, are often still highly complex and require lengthy computer solutions. *See* DIFFERENTIAL EQUATION.

An alternative approach is empirical modeling, in which a simple mathematical form is assumed for the model, with key physical parameters being determined by measurements on representative devices. Such models further reduce predictive power for increased simplicity.

Successive simplifications are required to reduce circuit-model complexity to a point where pencil-and-paper calculations can produce an initial new system design. Once this preliminary design is completed, computer simulations using more detailed models can further refine physical circuit-performance predictions. Depending on the degree of simplification employed in the design process, it may be necessary to construct circuit prototypes to verify adequate circuit performance. If the agreement between circuit-model predicted performance and actual physical performance is inadequate, the simplified circuit models needs to be modified to account for these parasitic (unmodeled) effects and the entire design sequence needs to be repeated. In an attempt to avoid this expensive and time-consuming process, much effort has been expended in the continual development and enhancement of computer tools for circuit modeling. *See* COMPUTER-AIDED CIRCUIT DESIGN.

Many devices with lumped-circuit models can be represented with three controlling nodes or terminals. One of these can be selected as a reference to which the other voltages are compared. The result is a two-port network (**Fig. 1***a*), with four variables: two port voltages and two terminal currents. Such a system can be characterized by using two independent equations. Ideally these equations should be expressible in closed-form: two of the variables are expressed as functions of the remaining two. For example, with voltages v_1 and v_2 in Fig. 1*a* as the independent variables, the closed-form model equations can be expressed as Eqs. (1) and (2). For devices

$$i_1(t) = f_1[v_1(t), v_2(t)] \quad (1)$$

$$i_2(t) = f_2[v_1(t), v_2(t)] \quad (2)$$

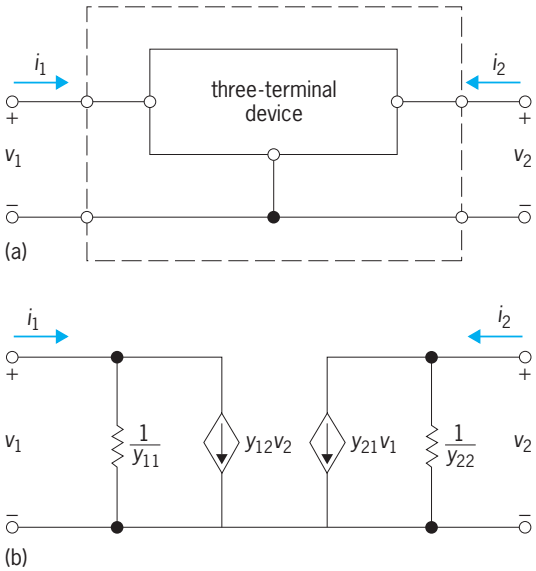


Fig. 1. Three-terminal device. (*a*) Two-port representation. (*b*) Y-parameter equivalent circuit. The diamond-shaped symbols represent voltage-controlled current sources.

controlled by more than three terminals, more port variables and more equations must be added to the model.

Once a suitable model is chosen, it can be used to determine the terminal currents and voltages that result when the device is connected in an electrical circuit. There are two main application areas for large-signal models. One is to find the response of a system to time-varying inputs. The other is to determine the voltages and currents of a circuit with zero input variations, as a preliminary to small-signal analysis.

Small-signal modeling. In many applications, current and voltage variations are small enough that the device's characteristics are approximately independent of signal size. This allows the device to be represented by a simplified linear equivalent circuit. To find the values of the equivalent circuit's elements, the input signal is replaced with its average value, and all currents and voltages are found by using the large-signal model. The results represent the quiescent operating point or Q point for the circuit. All currents and voltages can be approximated by series expansions around the Q point. For small enough signal variations, terms above first order can be neglected. For example (putting aside time derivatives for the moment), the currents in Eqs. (1) and (2) can be written as Eqs. (3) and (4), where the Δ's denote

$$i_1 = I_1 + \Delta i_1 \cong I_1 + \frac{\partial i_1}{\partial v_1} \cdot \Delta v_1 + \frac{\partial i_1}{\partial v_2} \cdot \Delta v_2 \quad (3)$$

$$i_2 = I_2 + \Delta i_2 \cong I_2 + \frac{\partial i_2}{\partial v_1} \cdot \Delta v_1 + \frac{\partial i_2}{\partial v_2} \cdot \Delta v_2 \quad (4)$$

signal-dependent variations, I_1 and I_2 are the Q-point currents, and the derivatives are evaluated at the Q point.

An equivalent circuit can then be drawn in which only the variations of the signal are modeled; all signal-independent voltage and current sources are set to zero. The nonlinear device is replaced in the circuit by its linearized equivalent, whose element values are given by the derivatives in the series expansion. Equations (3) and (4) lead to the equivalent circuit shown in Fig. 1*b*, where the parameters of Eqs. (5) are the derivatives of the nonlinear equations evaluated with v_1 and v_2 set equal to their Q-point values. These so-called *y*-parameters have units of conductance. Other small-signal representations result from different choices of dependent and independent variables.

$$y_1 = \frac{\partial i_1}{\partial v_1} y_{12} = \frac{\partial i_1}{\partial v_2}$$
$$y_{21} = \frac{\partial i_2}{\partial v_1} y_{22} = \frac{\partial i_2}{\partial v_2} \quad (5)$$

Linearity has several benefits. One is the applicability of systematic solution methods for linear circuits that simplify both manual and computer solutions. Linearity also allows superposition: a linear system's response to the sum of two inputs equals the sum of its responses to the inputs applied separately. An important consequence of superposition is that the system is completely specified by its response to sinusoids of arbitrary frequency. By using Fourier analysis, any input can be written as a unique sum of sinusoids whose magnitude and phase as functions of frequency make up the spectrum of the input. The linear system's response to each sinusoid is a sinusoid of the same frequency. The gain (amplitude increase) and phase shift vary with frequency but not with input amplitude. This frequency response determines the output spectrum for arbitrary inputs. Linear circuits can be analyzed entirely in the frequency domain. *See* FOURIER SERIES AND TRANSFORMS; GAIN; LINEARITY; RESPONSE; SUPERPOSITION THEOREM (ELECTRIC NETWORKS).

Neglecting time derivatives in the small-signal circuit derivation restricts the model's use to the midband region, where the gain is independent of frequency; in the midband equivalent circuit, all inductors and capacitors are treated either as short or open circuits.

Frequency-domain equivalent circuit. At high frequencies, parasitic effects inevitably reduce the gain. This frequency dependence is usually modeled by addition of capacitors or inductors to the model. The values of these elements depend (usually nonlinearly) on the Q point. The external circuit, of course, may also contain frequency-dependent elements. In the time domain, the current and voltage in an inductor or capacitor are related through time integrals or derivatives. These elements can be represented in the frequency domain with an imaginary impedance, directly or inversely proportional to frequency. The circuit can then be analyzed by using complex algebra, which is much easier than solving the time-domain differential equations. As an example of a frequency-domain model, **Fig. 2** shows a high-frequency small-signal equivalent circuit for a bipolar transistor. *See* ALTERNATING-CURRENT CIRCUIT THEORY; TRANSISTOR.

Distortion. If the signal amplitudes are too large, the small-signal approximation is invalid, causing distortion: the generation of frequency components in the output that are not present in the input. Distortion is analyzed in the time domain by using large-signal models. *See* DISTORTION (ELECTRONIC CIRCUITS).

Noise models. Many electrical devices produce random current variations due to the discrete charge of the electron. This electrical noise limits the

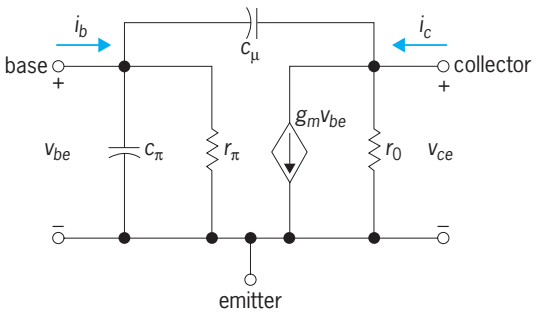


Fig. 2. High-frequency hybrid-π model for a bipolar transistor.

minimum usable signal amplitude. In the time domain, noise signals can be described only statistically. The statistical distributions can be used to represent noise in the frequency domain. The noise power at any frequency can be determined from the autocorrelation function, a measure of how closely two values of the signal measured at different times correspond as the time between the measurements is varied. Physical or empirical models are used to predict the spectrum's dependence on the Q point. *See* CIRCUIT (ELECTRONICS); ELECTRICAL NOISE.

Robert M. Fox; Philip V. Lopresti

Bibliography. G. Massobrio and P. Antognetti, *Semiconductor Device Modeling with SPICE*, 2d ed., 1993; A. S. Sedra and K. C. Smith, *Microelectronic Circuits*, 5th ed., 2004; Y. P. Tsividis, *Operation and Modeling of the MOS Transistor*, 3d ed., 2006; M. E. Van Valkenburg, *Network Analysis*, 3d ed., 1974.

Electrical noise

Interfering and unwanted currents or voltages in an electrical device or system. Electrical noise, or simply noise, has a significant effect on the design and operation of almost all electrical and optical systems which are used to communicate or process information. Noise is responsible for the familiar static observed on home radio receivers, the clicking sounds on frequency-modulation (FM) radios operating in fringe (near-threshold) areas, and the "snow"-type granularity on the viewing screen of a television receiver displaying a weak signal. In general, noise provides the fundamental limitation to the range over which radio or optical signals can be transmitted and received with integrity. Noise is, therefore, of great importance to the engineers who design and operate such systems.

It is convenient to differentiate between noise which results from human activity and that which is naturally occurring. Noise which results from human activity, such as that generated by an electrical appliance or an automotive ignition, can usually be eliminated or minimized by good design practice (shielding, filtering, equipment location, and so forth). Naturally occurring noise can be further subdivided into that which is irregular or erratic in nature and that which is more or less continuous. An example of noise which is irregular or erratic is that associated with an electrical storm. This type of noise is sometimes dealt with in the system design, but since it is only occasionally present, it does not ordinarily constitute a design limitation. On the other hand, naturally occurring noise which is essentially continuous in time is responsible for the fundamental limitation cited above. The remainder of the article therefore concentrates on this type of noise. *See* ATMOSPHERIC ELECTRICITY; CROSSTALK; ELECTRIC FILTER; ELECTRICAL INTERFERENCE; ELECTRICAL SHIELDING; ELECTROMAGNETIC COMPATIBILITY; ELECTRONIC EQUIPMENT GROUNDING; GROUNDING; STATIC.

Sources. Most noise generation is a consequence of the spontaneous fluctuations which occur within matter at the microscopic level. In electrical circuits these fluctuations give rise to what are commonly referred to as thermal noise and shot noise. Thermal noise is generated by the random motion of free electrons in a resistor or any conductor with resistance. The random motion, and thus the noise generated, is proportional to the temperature of the medium. At absolute zero temperature on the Kelvin scale (−459.67°F), all motion ceases and no noise is generated. Shot noise is most commonly identified with the fluctuations in the current of a vacuum tube caused by the random emission of electrons from its heated cathode. Shot noise is also observed in semiconductor devices as random fluctuations in carrier density when an electric field is applied. There are other types of noise associated with electrical circuits, but shot noise and thermal noise are by far the most important. *See* FREE-ELECTRON THEORY OF METALS; KINETIC THEORY OF MATTER; SEMICONDUCTOR; VACUUM TUBE.

In a system in which signals are transmitted through the atmosphere [for example, amplitude-modulation (AM) or FM radio broadcast, or satellite communications], the receiving system will always receive noise as well as the desired signals. This noise is a result of thermal radiation from the Earth, planets, Sun, Moon, the galaxy (galactic noise), radio-emitting stars, and atmospheric gases. In addition, there is a small background level of thermal radiation, uniformly distributed, which is believed associated with the big bang origin of the universe. All of these noise sources, weighted by the directional characteristics of the receiving antenna, will contribute to the overall system noise. *See* COSMIC BACKGROUND RADIATION; HEAT RADIATION; TERRESTRIAL RADIATION.

In an optical communications system, a signal level is represented by a number of energy packets called photons. The mean arrival rate of the photons at the detector is proportional to the optical intensity or signal strength. At the detector (a photodiode), the photons are absorbed, each creating a hole-electron pair and thus a current in which the electrons are randomly positioned in time and in which the mean number of electrons is proportional to the optical intensity. The statistical nature of this process gives rise to fluctuations in the number of photons representative of a given level and, subsequently, the number of electrons generated to represent that level. If the detector has internal gain as in an avalanche photodiode, each hole-electron pair can create additional hole-electron pairs. This process, however, is statistical in nature, resulting in a mean value of gain but giving rise to additional fluctuations in the generated current. *See* MICROWAVE SOLID-STATE DEVICES; OPTICAL COMMUNICATIONS; OPTICAL DETECTORS; PHOTODIODE; PHOTON.

Mathematical analysis. Because of the statistical or random nature of noise, noise voltages and currents must be dealt with by using the branch of mathematics that deals with random variables, that is, probability theory. A random variable can be characterized by a probability density function $p(x)$. The definition of

$p(x)$ is such that the probability that the random variable x lies between x_1 and x_2 is given by Eq. (1). Ob-

$$\text{Prob}\,(x_1 < x < x_2) = \int_{x_1}^{x_2} p(x)\,dx \tag{1}$$

viously, $p(x)$ can be normalized as shown in Eq. (2).

$$\int_{-\infty}^{\infty} p(x)\,dx = 1 \tag{2}$$

Investigations have shown that fluctuation noise, in general, has a probability density function which is gaussian or normal, provided that the noise can be represented as a large number of independent overlapping samples. The gaussian distribution has the form given in Eq. (3), where σ is the standard de-

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2} \tag{3}$$

viation, x is the random variable (noise voltage or current), and the average value is zero. *See* DISTRIBUTION (PROBABILITY); PROBABILITY.

If the average value is nonzero, the distribution will have the form illustrated by Eq. (4), where V is

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-V)/2\sigma^2} \tag{4}$$

the average value. V might be a signal voltage, for example, with the total instantaneous voltage fluctuating about that level.

In the foregoing, it has been assumed that the statistical properties of the random variable are invariant with time. The statistical process associated with such a system is said to be a stationary process. In experimentally collecting data to determine the probability distribution of a random variable (for example, a noise voltage), two methods are available. First, a large number of samples can be collected from a system at a sequence of times. Second, and alternately, the outputs of a large number of identical systems can be sampled simultaneously. The latter method is called the ensemble method. If the statistics collected by either of these methods are invariant with time, the process is stationary. Furthermore, if both methods yield identical statistics, the process is ergodic. Most noise processes are both stationary and ergodic. For such processes the mean value [V in Eq. (4)] is just the time average of the noise, and the variance σ^2 is the average of the square of the fluctuations about the mean value. *See* PROBABILITY (PHYSICS); STATISTICAL MECHANICS.

Two additional functions that are useful in characterizing random noise are the autocorrelation function $R(\tau)$ and the power spectral density $G(w)$. $R(\tau)$ provides a measure of the dependence between any two values of the noise variable $x(t)$ and $x(t + \tau)$, as specified in Eq. (5). $G(w)$ gives the distribution of

$$R(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t)\,x(t+\tau)\,dt \tag{5}$$

average power in noise as a function of frequency w. $G(w)$ and $R(\tau)$ are Fourier transform pairs; that is, they satisfy Eqs. (6), where $j = \sqrt{-1}$. This is a use-

$$G(w) = \int_{-\infty}^{\infty} e^{-jw\tau} R(\tau)\,d\tau$$
$$R(\tau) = \int_{-\infty}^{\infty} e^{-jw\tau} G(w)\,dw \tag{6}$$

ful relationship because it relates the autocorrelation function of a noise voltage at the output of a linear system to a function $G(w)$ which can be expressed in terms of the transfer function of the system. For example, if white noise (uniform power spectrum) is passed through a system with transfer function $H(w)$, the autocorrelation function at the output of the system is simply the Fourier transform of $H(w)^2$. *See* FOURIER SERIES AND TRANSFORMS.

Many network calculations involving shot noise and thermal noise can be handled simply by considering the noise sources as ordinary signal sources but with the distinction that mean-squared values are employed. Total voltages (or currents) can be determined by adding mean-squared values.

Noise figure and noise temperature. A quantity of considerable interest in describing the characteristics of an electrical or optical signal is the ratio of signal power to noise power (S/N) measured at various points in the system. In particular, the signal power–to–noise power ratio at the output of a network, $(S/N)_{\text{out}}$, can be related to the signal power–to–noise power ratio at the input of the network, $(S/N)_{\text{in}}$, by a term called the noise figure F, as expressed by Eq. (7). F is a quality factor of the net-

$$F = \frac{(S/N)_{\text{in}}}{(S/N)_{\text{out}}} \tag{7}$$

work because it provides a measure of the degradation introduced by the network to the signal-to-noise ratio. An ideal, lossless, noise-free network is one for which $F = 1$. Frequently F is expressed in decibels as shown by Eq. (8). Typical signal-to-noise ratio min-

$$F_{\text{dB}} = 10\,\log F \tag{8}$$

imum requirements may range from 3 to 15 dB for many detection systems or may be as large as 50 dB for a high-quality voice communications circuit. *See* SIGNAL-TO-NOISE RATIO.

In radio astronomy and space communications and, in general, systems which employ very low-noise first-stage amplifiers in the receiving system (for example, masers and parametric amplifiers), it has become common practice to characterize the amplifier and, indeed, the system in terms of an equivalent noise temperature. The noise figure defined previously can be expressed in terms of this equivalent noise temperature, as shown in Eq. (9),

$$F = \frac{T_e}{T_o} + 1 \tag{9}$$

where T_e is the equivalent noise temperature and T_o is the ambient temperature (usually taken as 290 K or 62°F). Then equivalent noise temperature is

expressed in terms of noise figure by Eq. (10). Tem-

$$T_e = (F - 1)T_o \quad \text{K} \tag{10}$$

peratures are always expressed in kelvins. *See* AMPLIFIER; MASER; PARAMETRIC AMPLIFIER; SPACE COMMUNICATIONS.

If the system consists of a series of cascaded networks, each with an equivalent noise temperature T_i and gain G_i, it can be shown that the overall equivalent noise temperature referred to the input terminals of the first stage is given by Eq. (11). For

$$T_e = T_1 + \frac{T_2}{G_1} + \frac{T_3}{G_1 G_2} + \cdots + \frac{T_n}{G_1 G_2 \cdots G_{n-1}} \tag{11}$$

this system or any system in which the overall equivalent noise temperature is known, the noise power spectral density is simply given by kT_e watts/hertz, where k is Boltzmann's constant (1.38×10^{-23} joule/K) and T_e is in kelvins. *See* BOLTZMANN CONSTANT.

Thermal noise. Spontaneous fluctuations in a resistor at temperature T will give rise to a noise voltage at its terminals with mean-squared value given by Eq. (12), where k is Boltzmann's constant, R is the

$$\overline{V_N^2} = 4kTRB \tag{12}$$

resistance in ohms, and B is the bandwidth in hertz. Equation (12) is often referred to as the Nyquist formula and the thermal noise as Nyquist noise or Johnson noise in recognition of the original investigators.

Equation (12) is actually an approximate formula, but it is valid with useful accuracy over all radio and microwave frequencies of general interest. It is not, however, valid at optical frequencies. The Nyquist formula shows that the spectral density $4kTR$ is constant. Noise with a constant or uniform distribution with frequency is referred to as white noise.

In a communications system, thermal noise is evidenced throughout the receiving system: the antenna, the transmission lines, amplifiers, and so forth. In systems where very low noise is vital to operation (such as those used in space communications), cryogenic cooling of the first stage of amplification is frequently employed. Liquid helium and liquid nitrogen have commonly been used for this purpose. Receiver noise temperatures approaching 4 K (-452°F) have been demonstrated. In these systems, antenna noise temperature becomes very important.

Antenna noise temperature is defined formally in Eq. (13), where A_e is the effective receiving area of

$$T_A = \frac{A_e}{\lambda^2} \int\int T(\theta, \phi)\, P_N(\theta, \phi)\, d\Omega \tag{13}$$

the antenna in m^2, λ is the wavelength in m, $P_N(\theta,\phi)$ is the normalized power pattern of the antenna, and $T(\theta,\phi)$ is the temperature distribution over all space in kelvins. $T(\theta,\phi)$ includes contributions from the atmosphere, the galaxy, radio-emitting stars, and so forth, as well as thermal radiation from the Earth itself. In Eq. (13), Ω is a solid angle and the integration is carried out over 4π steradians. $P_N(\theta, \phi)$ will depend on the pointing direction of the antenna; consequently, the antenna temperature will depend on the pointing direction of the antenna. At microwave frequencies, the antenna temperature is usually a minimum when the antenna is pointed to zenith. *See* ANTENNA (ELECTROMAGNETISM).

Shot noise. Shot noise, like thermal noise, is thermal in origin. However, it differs from thermal noise in two significant ways: (1) for shot noise to be present in a circuit, a voltage (for example, cathode to anode) or electric field must be impressed; (2) shot noise is not characterized by a uniform, wide-band (white) spectral density, but rather has a high-frequency cutoff given approximately by the inverse transit time of an electron from cathode to anode.

The probability that N electrons will be emitted from a heated cathode in a time interval τ is given by the Poisson distribution, Eq. (14), where $\overline{n}$ is the

$$P(N) = \frac{(\overline{n}\tau)^N e^{-\overline{n}\tau}}{N!} \tag{14}$$

average number of electrons emitted per second. If the electric field is sufficiently strong, all the emitted electrons are attracted to the anode, and the fluctuàtions in current are given by the fluctuations in emission. This is called the temperature-limited case, and the mean-squared noise current is given by the shot-effect formula of Eq. (15), where e is the

$$\overline{i_s^2} = 2e\overline{I}B \tag{15}$$

charge on an electron, $\overline{I}$ is the average current, and B is the bandwidth in hertz.

If the applied electric field is not sufficiently strong to pull all the electrons to the anode, a negative space charge develops in the vicinity of the cathode which tends to inhibit shot-effect fluctuations. This is the space-charge-limited case, and the shot noise is then approximated by Eq. (16), where Γ^2 is a space-charge

$$\overline{i_s^2} = 2e\overline{I}\Gamma^2 B \tag{16}$$

reduction factor varying commonly between 0.01 and 1. The value of Γ^2 depends on physical geometry, cathode temperature, and applied voltage. *See* SPACE CHARGE.

Noise in optical detectors. The reception of an optical signal involves the use of a photodetector in which the optical signal is converted to an electric current. Thus, both optical noise and electronic noise can be expected to influence the signal-to-noise ratio at the receiver output.

Optical noise relates to the random arrival of photons at the detector and subsequent generation of hole-electron pairs. The number of hole-electron pairs generated during the period t to $t + \tau$ is a random variable having the Poisson distribution given

by Eq. (17), where Λ is given by Eq. (18), where η is

$$P(n) = \frac{\Lambda^n e^{-\Lambda}}{n!} \tag{17}$$

$$\Lambda = \int_t^{t+\tau} \frac{\eta}{hf} p(t)\, dt \tag{18}$$

the quantum efficiency (fraction of light absorbed), h is Planck's constant, f is the optical frequency, and $p(t)$ is the optical power.

For the case where $p(t)$ has a constant value p_0, an expression for the signal-to-noise ratio is given as Eq. (19), where e is charge on the electron, P_N is the

$$S/N = \frac{\left(\frac{\eta e p_0}{hf}\right)^2}{\frac{2\eta e^2 p_0}{hf} B + P_N} \tag{19}$$

mean-square thermal (electronic) noise power associated with amplification of the generated current, and B is the bandwidth in hertz.

Many modern optical communications systems employ avalanche photodiodes. In the avalanche photodiode, a received photon creates a hole-electron pair, which in turn, by means of ionization collisions, creates additional hole-electron pairs. Thus the detector has internal gain. However, the process which gives rise to the gain is itself a random process, and therefore, additonal fluctuation noise is generated. The signal-to-noise ratio for the optical receiver employing an avalanche photodetector is given by Eq. (20), where G represents the interval

$$S/N = \frac{\left(\frac{neGp_0}{hf}\right)^2}{\frac{2ne^2G^2Fp_0B}{hf} + P_N} \tag{20}$$

gain and F is the excess noise factor.

If, in Eq. (19) or (20), the thermal noise term is sufficiently small compared to the optically generated noise so as to be neglected, the resulting S/N is referred to as the quantum limit. In most present-day systems, however, thermal noise is found to be the dominant term.

Noise measurement. The most common apparatus used to measure noise is a total power receiver consisting of a wide-band amplifier, a square-law detector, an integrator, and an indicator. The sensitivity of the receiver (the minimum detectable noise fluctuation) is generally improved as the front-end bandwidth and the integration time is increased. Sensitive receivers of this type are called total power radiometers and are frequently employed in radio astronomy. *See* RADIOMETRY.

The noise temperature (or noise figure) of an amplifier can be determined by what is commonly called the Y-factor method. The measurement utilizes two calibrated noise sources and a power receiver of the type just described. One of the noise sources is connected to the input of the amplifier, and the output power is measured with the receiver. This is repeated with the second noise source. The ratio of power received in the two cases is recorded and denoted Y. The amplifier noise temperature is then given by Eq. (21), where T_1 and T_2 (where T_2 is

$$T_{\text{amp}} = \frac{T_2 - YT_1}{Y - 1} \quad \text{K} \tag{21}$$

greater than T_1) are the noise temperatures of the calibrated terminations. It is assumed that the amplifier has sufficiently high gain that the noise contribution from the receiver can be neglected.

The total system noise temperature can be measured by using the noise-adding method. In this method an increment of noise is injected into the system and the total received noise power is compared to the noise power without the added increment. Then the total system noise power is given by Eq. (22), where Y is given by Eq. (23), and ΔT is the

$$T_{\text{syst}} = \frac{\Delta T}{Y - 1} \quad \text{K} \tag{22}$$

$$Y = \frac{\text{system noise power} + \text{injected noise power}}{\text{system noise power}} \tag{23}$$

noise temperature of injected noise.

Modern test equipment is available which greatly simplifies the measurement of the noise performance of circuits. Automated instrumentation which provides output directly in terms of noise figures is, for example, readily available. *See* ELECTRICAL NOISE GENERATOR.

Randall W. Kreutel

Bibliography. M. S. Gupta, *Noise in Circuits and Systems*, 1988; J. D. Kraus, *Radio Astronomy*, 2d ed., 1986; J. E. Midwinter, *Optical Fibers for Transmission*, 1979, reprint 1991; R. Morrison, *Noise and Other Interfering Signals*, 1991; R. Pettai, *Noise in Receiving Systems*, 1984; M. Schwartz, *Information, Transmission, Modulation and Noise*, 4th ed., 1990.

Electrical noise generator

A device that produces electrical noise for use in electrical measurements. Electrical noise generators are commonly employed to measure the noise figure or noise temperature of radio receivers. They are also used in various other tests in radar and communications systems. Celestial noise sources are used to calibrate large antennas.

Some standard types of noise generators are hot-wire, diode, gas-discharge tube, hot and cold loads (terminations), and radio-star. A hot-wire noise source consists of the filament of a lamp heated by direct current. Thermal noise having spectral density $4kTR$, where k is Boltzmann's constant and T and R are the temperature and resistance of the filament respectively, is generated across the terminals of the filament. A diode noise generator utilizes the temperature-limited shot effect to generate noise. At frequencies less than the reciprocal transit time of the diode, the noise spectral density is $2e\bar{I}$, where e is the charge on the electron and $\bar{I}$ is the average anode current. A gas-discharge noise generator, commonly

referred to as a noise tube, consists of a fluorescent light tube enclosed in a waveguide. Noise generation is essentially thermal. The noise tube is commonly employed at microwave frequencies. Hot and cold loads consist of well-matched terminations, either transmission line or waveguide, held at a given temperature by using an oven or by applying cryogenic refrigeration. Noise generation is thermal. Common temperatures for noise-generating terminations are nominally 80 and 300 K (−316 and 80°F)

Celestial radio sources (radio stars) are commonly employed as reference noise sources for evaluating the characteristics of very low-noise, high-gain space communications receiving antennas. There are a number of accurately calibrated sources available—the choice depending on system parameters (frequency, antenna gain, system noise temperature, elevation angle, and so forth) and physical location. The most common radio sources employed are Cassiopeia A, Taurus A, Cygnus A, and Orion A. The first three are classified as nonthermal sources in which radiation results from relativistic electrons interacting with an interstellar magnetic field. The electrons are rotated in a plane perpendicular to the magnetic field direction, and radiation is characterized by a component polarized parallel to that plane. The polarized component is small, however, and the major portion of the radiation is unpolarized. The nonthermal sources have flux densities which decrease with increasing frequency and, consequently, tend to have a cutoff frequency above which they are not usable. Orion A is a thermal source in which radiation occurs from a hot, ionized cloud. Orion A has a constant flux density at frequencies above 2 GHz. As a point of reference, a typical value of thermal noise received from Cassiopeia A by a 10-m-diameter (33-ft) antenna operating in the C band (4–6 GHz) would be about 150 K (−190°F). *See* POLARIZED LIGHT.

Other possible sources include the Sun, Moon, and the planets. The solar flux density is equal to that from blackbody radiation at 6000 K (10,300°F) at wavelengths less than 1 cm, but is greater than this at longer wavelengths. At radio wavelengths the radiation can be greatly enhanced during periods of sunspot activity. In general, at longer wavelengths the solar radio emission can be considered as the superposition of a slowly varying component dependent on the sunspot number and a rapidly varying component resulting from solar flares. The polarization of the former has a strong circularly polarized component, while that of the latter is generally random. *See* SUN.

At radio wavelengths, the Moon emits like a thermal blackbody with temperature in the range from 150 to 280 K (−190 to 44°F) depending on the lunar phase. At long wavelengths the lunar-phase dependence is minimal, with small (less than 10%) variations about a nominal temperature of approximately 190 K (−118°F).

It is important that variations in brightness over the source be corrected, that polarized components of flux be accounted for, and that attenuation characteristics of the intervening medium, the atmosphere, be included. *See* ELECTRICAL NOISE; RADIO ASTRONOMY.

Randall W. Kreutel

Bibliography. M. S. Gupta (ed.), *Electrical Noise: Fundamentals and Sources*, 1977; M. S. Gupta, *Noise in Circuits and Systems*, 1988; J. D. Kraus, *Radio Astronomy*, 2d ed., 1986; R. Morrison, *Noise and Other Interfering Signals*, 1991; A. Van der Ziel, *Noise: Sources, Characterization and Measurement,* 1970.

Electrical resistance

Opposition of a circuit to the flow of electric current. Ohm's law states that the current I flowing in a circuit is proportional to the applied potential difference V. It is obeyed very accurately for metallic and many other conductors over a wide range of temperatures. The constant of proportionality is defined as the resistance R. Hence, Eq. (1) holds. If V and I

$$V = IR \tag{1}$$

are measured in volts and amperes, respectively, R is measured in ohms. Microscopically, resistance is associated with the impedance to flow of charge carriers offered by the material. For example, in a metallic conductor the charge carriers are electrons moving in a polycrystalline material in which their journey is impeded by collisions with imperfections in the local crystal lattice, such as impurity atoms, vacancies, and dislocations. In these collisions the carriers lose energy to the crystal lattice, and thus Joule heat is liberated in the conductor, which rises in temperature. The Joule heat P is given by Eq. (2).

$$P = I^2R = IV = \frac{V^2}{R} \tag{2}$$

See CRYSTAL DEFECTS; ELECTRICAL CONDUCTIVITY OF METALS; ELECTRICAL RESISTIVITY; JOULE'S LAW; OHM'S LAW.

Peter A. Schroeder

Electrical resistivity

Opposition of a material to the flow of electricity. For a conductor of uniform cross section A (in square meters), length l (in meters), and resistance R (in ohms, Ω), the resistivity is defined by Eq. (1).

$$\rho = \frac{RA}{l} \qquad (\Omega \cdot m) \tag{1}$$

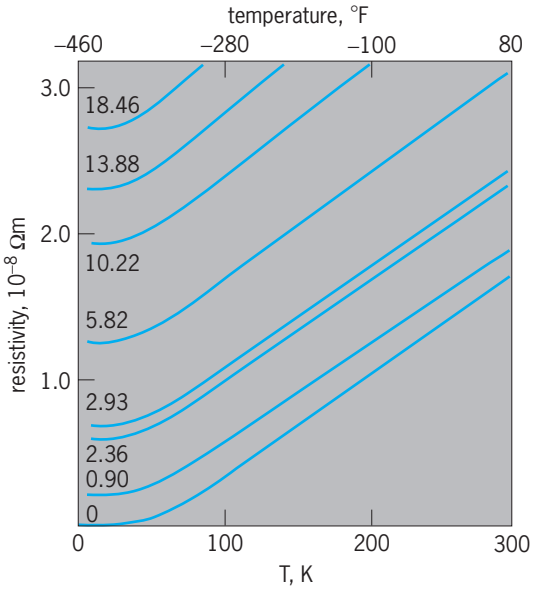
Unlike the extensive quantity R, ρ is an intensive quantity, that is, an intrinsic property of a material, independent of its amount or shape. The range of resistivity values is extremely wide (see **table**). At the extremes are conductors such as copper with a room-temperature resistivity of $1.7 \times 10^{-8}\ \Omega \cdot m$ and insulators such as quartz with a resistivity greater than $10^{16}\ \Omega \cdot m$. The range is even wider when temperature dependence (a property useful in classifying materials) is taken into account. *See* ELECTRIC INSULATOR; ELECTRICAL RESISTANCE.

Resistivities at room temperature

Material	Ω · m
Poly(tetrafluoroethylene)	10^{18}
Quartz	10^{16}
Diamond	10^{12}
Glass	10^{11}
Sodium chloride*	10^{2}
Silicon, pure	10^{3}
Germanium, pure	5
Metglass 2204 ($Ti_{50}Be_{40}Zr_{10}$)	3
Bismuth	1.2×10^{-6}
Antimony	4.2×10^{-7}
Nichrome	10^{-6}
Iron	10×10^{-8}
Potassium	7×10^{-8}
Copper	1.7×10^{-8}
Copper, pure†	10^{-12}

*At 1000 K (1340° F).
†At 4.2 K (–452° F).

Crystalline metals. For a pure metal such as copper (see **illus.**), the room-temperature resistivity is almost entirely caused by scattering of electrons by lattice vibrations, or, in quantum terms, phonons. As the temperature decreases, so does the phonon density, resulting in a decrease of resistivity. At the temperature of liquid helium (4.2 K or −452°F) this is so low that phonon scattering makes very little contribution to the resistivity, which is then determined by the concentration of imperfections in the crystal lattice. Thus, at 4.2 K and below, the resistivity may be 10^{-11} Ω · m or even lower. In an unstrained metal the imperfections responsible for the residual restivity at 4.2 K is often taken as a measure of purity. *See* CRYSTAL DEFECTS; LATTICE VIBRATIONS; PHONON.



Resistivity of the disordered copper-zinc alloy system. Numbers on the curves give the concentration of zinc in atomic percent. (*After W. E. Henry and P. A. Schroeder, The low-temperature resistivities and thermopowers of α-phase copper-zinc alloys, Can. J. Phys., 41:1076–1093, 1963*)

Many other scattering mechanisms contribute to resistance. Among these is mutual scattering of the charge carriers. Inasmuch as there are about 10^{23} electrons per cubic centimeter in a typical metal, a large contribution might be anticipated. However, the number of electrons that can interact with each other is severely restricted by energy considerations, and the T^2 variation of resistivity with absolute temperature T, typical of electron–electron scattering, is seen only at temperatures of a few kelvins where the phonon contribution to the resistivity becomes negligible. *See* ELECTRICAL CONDUCTIVITY OF METALS.

Disordered crystalline alloys. These are alloys in which the component ions are randomly distributed on the crystal lattice sites. For such an alloy the resistance can be written as in Eq. (2), where ρ_r is

$$\rho = \rho_r + \rho_i \tag{2}$$

the residual resistance produced by electron scattering by the impurity ions, and ρ_i is the resistivity contribution arising from the scattering of electrons by phonons. Matthiessen's rule states that ρ_i is the same for alloys as for the pure metal. For the more dilute alloys in systems such as copper-zinc, this is nearly true, as indicated by the parallelism of the graphs of resistivity versus temperature (see illus.), but Matthiessen's rule is at best an approximation which worsens as the concentration of the alloying constituent increases. Nichrome is an alloy with a high resistivity because of the scattering from a disordered array of the constituent ions. Because of its high resistivity and stability at high temperatures, it is frequently used as a heating element. The resistivity can become very high in an amorphous metal, such as Metglass 2204 (see table), in which all long-range order is lost. But the resistivity of a concentrated alloy may be considerably reduced if the component ions form an orderly array. This occurs in intermetallic compounds, which can be represented chemically as A_pB_q, where p and q are small integers, all the A atoms occupy one orderly array of sites, and all the B atoms occupy another. For example, the gold-gallium compound $AuGa_2$ has a residual resistivity smaller than that of many metals in their purest available form. *See* INTERMETALLIC COMPOUNDS; MATTHIESSEN'S RULE; METALLIC GLASSES.

Semimetals. Bismuth, antimony, arsenic, and graphite are classified as semimetals and have higher resistivities than iron, potassium, and copper (see table), because they have a much lower concentration of mobile electrons that can participate in the conduction process. Their resistivity, like that of metals, increases with temperature.

Semiconductors. The semiconductors silicon and germanium also have a higher resistivity than metals, again because the number of charge carriers is much reduced. They differ from semimetals in that, in their pure or intrinsic form, the number of charge carriers and the conductivity increase exponentially with absolute temperature, and consequently their resistivity decreases as the temperature increases. *See* SEMICONDUCTOR.

Ferromagnetic materials. Several transport properties of ferromagnetic alloys can be explained by assuming conduction in parallel by the spin-up and spin-down electrons. Such a model is assumed to explain the resistance of magnetic multilayers. For example, in a sample consisting of alternate layers of copper and cobalt with layer thicknesses up to 10 nanometers, the magnetizations of the cobalt layers lie in the plane of the layers and, for the appropriate thicknesses of the copper layers, the magnetizations of alternate cobalt layers are in opposite directions. When the magnetizations are made to line up by applying a magnetic field, the resistance is observed to drop so dramatically that this phenomenon is frequently referred to as giant magnetoresistance. The phenomenon is explained in terms of spin-dependent scattering. *See* MAGNETORESISTANCE.

Charge-transfer salts. A metallic conducting state with conductivity decreasing with temperature is also exhibited by organic charge-transfer salts. These consist of large organic donor molecules (cations) which can stack in various ways, and smaller inorganic acceptor molecules which arrange themselves between the stacks. TMTSF (tetramethyltetraselenafulvalene) and BEDT-TTF [bis(ethylenedithiolo)tetrathiafulvalene] are typical of complicated organic donor molecules. The TMTSF stacks in chains which give rise to quasi-one-dimensional conductivity. The BEDT-TTF salts are more two-dimensional. These materials are exciting in that the available choice of anions yield materials of low dimensionality with a multitude of interesting properties, including superconductivity. *See* ORGANIC CONDUCTOR.

Ionic conductors. The resistivity of ionic conductors such as sodium chloride (NaCl; see table) decreases with temperature. However, conduction in these materials is a thermally activated process, in which the conducting ions have to pass over the potential barriers separating them from an adjacent site. *See* IONIC CRYSTALS.

Resistance thermometry. The sensitivity of the resistivity to temperature has practical applications in thermometry. Platinum resistance thermometers are used between 14 and 330 K (−434 and 134°F); rhodium-iron alloys between 1 and 800 K (−458 and 980°F); and carbon and germanium at very low temperatures (0.01 to 100 K or −459.65 to 280°F) where their sensitivity arises from the rapid decrease in charge carrier density as the temperature is lowered. *See* LOW-TEMPERATURE THERMOMETRY; THERMOMETER.

Peter A. Schroeder

Bibliography. J. S. Dugdale, *The Electrical Properties of Metals and Alloys*, 1977; K. M. Hellwege and O. Madelung (eds.), *Landolt-Börnstein Numerical Data and Functional Relationships in Science and Technology*, *New Series*, *Group 3*: *Crystal and Solid State Physics*, vol. 15a: *Metals*: *Electronic Transport Phenomena*, 1982; C. Kittel, *Introduction to Solid State Physics*, 7th ed., 1996; K. Schröeder, *Handbook of Electrical Resistivities of Binary Metallic Alloys*, 1983.

Electrical shielding

The imposition of a metal or composite barrier between one or more sources of electrical noise and their victims with the objective of reducing or eliminating electrical interference. Examples of the barrier are the case or housing of equipment; shields covering interconnecting cables between equipment; large cabinets, racks, or consoles; shielded (screen) rooms; and entire shielded buildings or vehicles.

Shielding effectiveness. The principal measure of a shield's performance is the shielding effectiveness. It is defined by theequation below, where SE_{dB} is

$$SE_{\mathrm{dB}} = 20 \log_{10} \frac{F_b}{F_a}$$

the shielding effectiveness in decibels, F_b is the electric (or magnetic) field strength before imposition of the barrier, and F_a is the electric (or magnetic) field strength after imposition of the barrier. *See* DECIBEL.

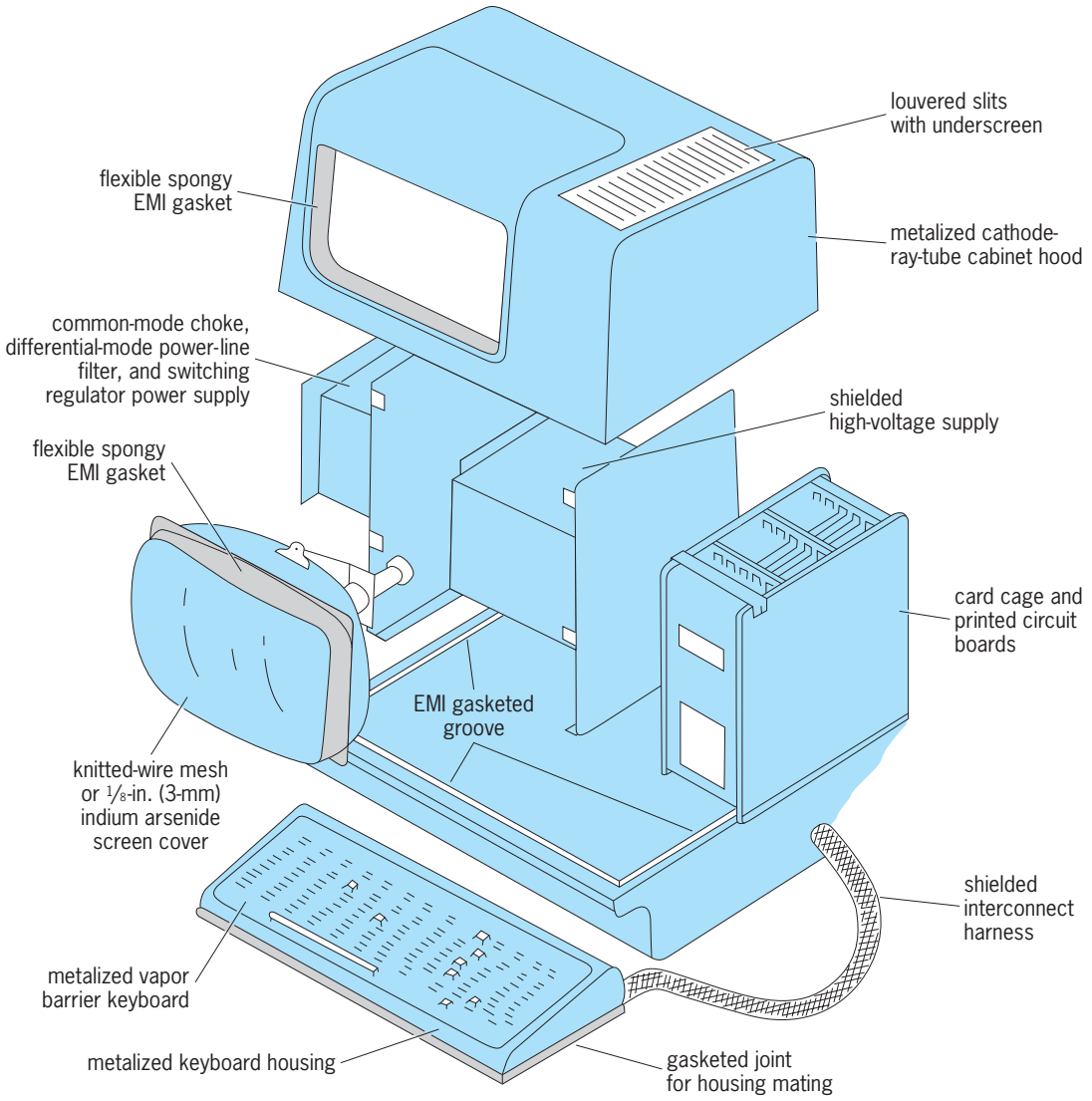
Shielding is obtained by the combination of reflection loss and absorption loss. The former is due to the impedance mismatch between the wave impedance of an oncoming wave and the surface impedance of the interposed barrier. Absorption loss corresponds to the attenuation due to skin effect at higher frequencies, and is dependent upon the frequency, conductivity, permeability, and thickness of the barrier. Shielding effectiveness is the sum of both losses. *See* ABSORPTION OF ELECTROMAGNETIC RADIATION; ATTENUATION (ELECTRICITY); RADAR-ABSORBING MATERIALS; REFLECTION AND TRANSMISSION COEFFICIENTS; REFLECTION OF ELECTROMAGNETIC RADIATION; SKIN EFFECT (ELECTRICITY).

Metal shields. Most intentional shields are made of metal to ensure high reflection losses. Even thin metals, such as household aluminum foils (with a thickness of about 1.5 mils or 0.038 mm), offer shielding effectiveness in excess of 100 dB. At low frequencies, these foils become electromagnetically transparent (that is, they do not attenuate magnetic fields). Thus, if there is a shielding problem due to the selection and makeup of the metal barrier, it is likely to occur only for low-frequency magnetic fields.

To obtain significant shielding to magnetic fields at low frequencies, the metal barrier must be very thick or composed of a highly permeable material such as Mumetal, Supermalloy, or Hypernom. Often such shields are fabricated in two or more layers, frequently laminated, to obtain good shielding properties per unit size or weight.

Leaky apertures. The major problem encountered in achieving adequate shielding at other than low frequencies is leaky apertures, not the metal choice. To provide the integrity required of the basic shield metal, all apertures must be protected or secured, or special electrical components and gaskets should be used (see **illus.**).

Shielded rooms. Shielded enclosures, ranging in size from small rooms to entire buildings, are



Some measures for controlling electromagnetic interference in video display terminals.

commercially available for performing as radiated emission and susceptibility test chambers, secure facility radiation containment, and protection for sensitive equipment from outside electromagnetic radiation. These enclosures typically provide in excess of 120 dB attenuation from 10 kHz to 1 GHz, and undergo reduced performance at lower and higher frequencies. Where extended performance is required, shielded rooms are usually made of solid seam-welded panels of galvanized steel. Special attention is given to leaky areas such as doors, and entrances where cables or wires such as power-line and telephone are brought in through filters. *See* ELECTRIC FILTER; ELECTROMAGNETIC PULSE (EMP).

Cable shields. Interconnecting cables between equipment often act as an "antenna farm" in which they behave as undesired pickup antennas and provide radiation escape from internal signals conducted along their length. Shielding such cables and harnesses is the dominant protection mechanism; other options include absorbers and filters. Cable shields may vary in complexity from one cover braid, through multibraids and composite braids or foils, to extruded tubes or conduits. *See* COMMUNICATIONS CABLE; ELECTRICAL INTERFERENCE; ELECTRICAL NOISE; ELECTROMAGNETIC COMPATIBILITY.

Donald R. J. White

Bibliography. L. H. Hemming, *Architectural Electromagnetic Shielding Handbook*, 1992; M. Mardiguian and D. R. J. White, *Electromagnetic Shielding*, vol 3: *Interference Control Technologies*, 1988; D. R. J. White, *Handbook on Electromagnetic Shielding Materials and Performance*, 2d ed., 1980.

Electrical units and standards

The process of measurement consists in finding out how many times the quantity to be measured contains a fixed quantity of the same kind, called a unit. The definitions of the units often involve complex physical theory and do not lend themselves readily to practical realization. The concrete representations

of units are known as measurement standards. In practice, measurements are made by using an instrument calibrated against a local reference standard, which itself has been calibrated either directly or by several links in a traceability chain against the national standard held by the national standards laboratory.

In order that measurements of similar quantities made in different countries may be compared, it is important that the values of the standards held by these laboratories should be known in relation to one another. International harmonization of national standards is the concern of the International Bureau of Weights and Measures (Bureau International des Poids et Mesures, or BIPM) that was set up at Sèvres, near Paris, under the Convention of the Meter (1875). The General Conference on Weights and Measures (CGPM) has overall control of the organization at the political level and meets every few years. Scientific questions and the running of the bureau are the responsibility of the International Committee of Weights and Measures (CIPM), which meets annually and is assisted by a number of consultative committees, including the Consultative Committee for Electricity and Magnetism (CCEM). The CCEM encourages research on electrical standards and arranges international comparisons of national standards, including standards of radio frequency and microwave quantities. *See* PHYSICAL MEASUREMENT.

Electrical and Magnetic Units

With the growth of the electric telegraph and supply industries in the nineteenth century, the need was recognized for accurate measurements of electrical quantities, based on generally accepted units and standards. The scientific community played an important part in developing these. From the start, the electrical units had a logical scientific basis, rather than deriving from ancient arbitrary practice.

The cgs system. A proposal by W. E. Weber in 1851 led to the absolute cgs system in which all units of quantities to be measured could be derived from the base units of length, mass, and time—the centimeter, gram, and second. This system was widely adopted although it had three weaknesses: the size of the units was inconvenient for practical use; it was difficult to realize the units from their definitions; and there were separate sets of units for electrostatic and electromagnetic quantities, based respectively on the inverse-square laws of force between electric charges and between magnetic poles. These inverse-square laws involved the permittivity ϵ_0 and the permeability μ_0 of free space (also known as the electric and magnetic constants). Ratios of the electrostatic and electromagnetic units for the same quantity involved the product $\epsilon_0\mu_0$, which J. C. Maxwell showed to be equal to the reciprocal of the square of the speed of light, c_0^{-2}. This was the first example of a relationship between the electrical units and a fundamental physical constant.

International units. The first weakness was resolved by international agreement in 1881 to fix the practical units—the volt, the ohm, and the ampere—at 10^8, 10^9, and 0.1 times the respective cgs electromagnetic units. The other weaknesses were avoided by the decisions of the 1908 International Congress in London, where realizations of these units in terms of easily reproducible standards were defined: the international ohm as the resistance of a column of mercury of specified dimensions; the international ampere as the current which would deposit silver in a silver voltameter at a given rate; and the international volt as their product, with an alternative definition in terms of the electromotive force of a Weston cell. The values were chosen to approximate as closely as the available information allowed to the exact multiples of the cgs units, but the international units were defined in their own right and their values were not changed when the results of later absolute determinations became available. They continued in use until officially abolished by the CGPM in 1948. *See* ELECTROMOTIVE FORCE (CELLS).

The mksa units. A more fundamental change resulted from a proposal by G. Giorgi in 1902. This led to the adoption of the mksa system of units, in which there are four base units: the meter, the kilogram, the second, and the ampere. Use of the meter and the kilogram instead of the centimeter and the gram gave units of a size more convenient for practical use, and use of the ampere as a base unit resolved the conflict between electrostatic and electromagnetic units while maintaining the magnitudes of the widely used practical units. This was a truly coherent system, in the sense that other units were derived from the base units without the need for factors of proportionality other than unity.

SI units. From the mksa system the present-day SI (Système Internationale) has developed, by the addition of further base units to include other fields of measurement. SI units were formally adopted by the CGPM in 1954.

The seven base units of SI are listed in **Table 1**. Of these, the first two, the kilogram and the second, are defined independently of the others. The kilogram is the mass of the international prototype of the kilogram held at the BIPM; the second is the duration of 9 192 631 770 periods of oscillation of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium-133 atom. The definitions of the other base units involve these two units. Until October 1983, the meter was also independently defined, in terms of the wavelength of a krypton lamp, but at that time

TABLE 1. SI base units

Quantity	Unit	Symbol
Mass	kilogram	kg
Time	second	s
Length	meter	m
Electric current	ampere	A
Thermodynamic temperature	kelvin	K
Luminous intensity	candela	cd
Amount of substance	mole	mol

TABLE 2. Some derived electrical units

Quantity	Unit and symbol	Derivation
Potential difference, emf	volt, V	$W \cdot A^{-1} = m^2 \cdot kg \cdot s^{-3} \cdot A^{-1}$
Resistance	ohm, Ω	$V \cdot A^{-1} = m^2 \cdot kg \cdot s^{-3} \cdot A^{-2}$
Electric charge	coulomb, C	$s \cdot A$
Capacitance	farad, F	$C \cdot V^{-1} = m^{-2} \cdot kg^{-1} \cdot s^4 \cdot A^2$
Conductance	siemens, S	$A \cdot V^{-1} = m^{-2} \cdot kg^{-1} \cdot s^3 \cdot A^2$
Magnetic flux	weber, Wb	$V \cdot s = m^2 \cdot kg \cdot s^{-2} \cdot A^{-1}$
Inductance	henry, H	$Wb \cdot A^{-1} = m^2 \cdot kg \cdot s^{-2} \cdot A^{-2}$
Magnetic flux density	tesla, T	$Wb \cdot m^{-2} = kg \cdot s^{-2} \cdot A^{-1}$
Magnetic field strength	ampere per meter	$m^{-1} \cdot A$
Current density	ampere per square meter	$m^{-2} \cdot A$
Electric field strength	volt per meter	$V \cdot m^{-1} = m \cdot kg \cdot s^{-3} \cdot A^{-1}$
Permittivity	farad per meter	$F \cdot m^{-1} = m^{-3} \cdot kg^{-1} \cdot s^4 \cdot A^2$
Permeability	henry per meter	$H \cdot m^{-1} = m \cdot kg \cdot s^{-2} \cdot A^{-2}$

the CGPM decided to take a fixed value for the speed of light and redefined the meter as the length of the path traveled by light in vacuum in 1/299 792 458 of a second. As a result, it is now possible to realize the meter with a lower uncertainty than previously. *See* LENGTH; LIGHT.

The units of other physical quantities (derived units) are derived from the base units by simple numerical relations, since SI is a coherent system. Examples are the unit of force, the newton ($N = m \cdot kg \cdot s^{-2}$); of energy, the joule ($J = N \cdot m$); and of power, the watt ($W = J \cdot s^{-1}$). Some of the more important derived electrical and magnetic units are listed in **Table 2**.

The SI base unit for electrical measurements is the ampere (A), the unit of electric current. It is defined in terms of a hypothetical experiment as that constant current which, if maintained in two straight parallel conductors of infinite length, of negligible circular cross section, and placed 1 meter apart in vacuum, would produce between these conductors a force equal to 2×10^{-7} newton per meter of length. According to electromagnetic theory, the force per unit length between two such conductors a distance d apart and carrying a current I is given by Eq. (1).

$$F = \frac{\mu_0 I^2}{2\pi d} \tag{1}$$

The definition of the ampere is thus equivalent to fixing the value of μ_0 as $4\pi \times 10^{-7}$ henry per meter.

The volt (V) is the unit of potential difference and of electromotive force. It is defined as the potential difference between two points of a conducting wire carrying a constant current of 1 ampere when the power dissipated between these points is equal to 1 watt. From the ampere and the volt, the ohm is derived by Ohm's law, and the other derived quantities follow in a similar manner by the application of known physical laws. *See* OHM'S LAW.

The following are the remaining units of electrical and magnetic quantities that have special names in the SI, together with their formal definitions, where appropriate.

Coulomb (C): The unit of electric charge, equal to 1 ampere-second. The coulomb is the quantity of electricity carried in 1 second by a current of 1 ampere.

Farad (F): The unit of capacitance, equal to 1 coulomb per volt. The farad is the capacitance of a capacitor between the plates of which there appears a potential difference of 1 volt when it is charged by a quantity of electricity of 1 coulomb.

Henry (H): The unit of inductance, equal to 1 weber per ampere. The henry is the inductance of a closed circuit in which an electromotive force of 1 volt is produced when the electric current in the circuit varies uniformly at the rate of 1 ampere per second.

Ohm (Ω): The unit of electrical resistance, equal to 1 volt per ampere. The ohm is defined as the resistance between two points of a conductor when a constant potential difference of 1 volt, applied to these points, produces in the conductor a current of 1 ampere, the conductor not being the seat of any electromotive force.

Siemens (S): The unit of electrical conductance (the reciprocal of resistance), equal to 1 ampere per volt. It was formerly known as the mho.

Tesla (T): The unit of magnetic flux density, equal to 1 weber per square meter.

Weber (Wb): The unit of magnetic flux, equal to 1 volt-second. The weber is the magnetic flux which, linking a circuit of one turn, would produce in it an electromotive force of 1 volt if it were reduced to zero at a uniform rate in 1 second.

The mechanical units of frequency (hertz), energy or work (joule), and power (watt) are frequently involved in expressing electrical and magnetic quantities. The cgs units, such as the gauss, gilbert, maxwell, and oersted, formerly used, are not part of the SI and are now obsolete. *See* UNITS OF MEASUREMENT.

Electrical Standards

Realization of the values of the electrical and other units from their SI definitions involves great experimental difficulties. For this reason, it is customary for national standards laboratories to maintain stable primary standards of the units against which other reference standards can be compared. From time to time, absolute determinations of the values of these

primary standards were made in terms of their definitions, using the methods described below.

For a number of years the ampere and the ohm were determined using a current balance and a calculable mutual inductance. The latter were superseded by the calculable capacitor and the moving-coil balance until the late 1980s, when higher accuracies became necessary. Other methods were explored, but by then the Josephson effect and the quantum Hall effect had made possible the standardization of the volt and the ohm by relation to fundamental physical constants. The recommendation by the CCEM of the values to be adopted for these constants in 1990 led to a complete change in primary electrical standards and the method of handling them; and all the major national laboratories, and the BIPM, now use this method.

For many years the primary standards maintained by most laboratories were the volt, in terms of the mean electromotive force of a group of Weston cells, and the ohm, using a group of standard resistors. A range of reference standards of other quantities are derived from these, including direct-current (dc) voltage and resistance at a variety of levels; alternating-current (ac) voltage, resistance, and power; capacitance and inductance; radio-frequency (rf) and microwave quantities; magnetic quantities and properties of materials; dielectric properties; and other quantities. These secondary standards are used for day-to-day measurements and for the calibration of local reference standards of other users in the national measurement system. *See* CAPACITANCE MEASUREMENT; ELECTRIC POWER MEASUREMENT; INDUCTANCE MEASUREMENT; MAGNETIC MATERIALS; MICROWAVE MEASUREMENTS; PERMITTIVITY; RESISTANCE MEASUREMENT; VOLTAGE MEASUREMENT.

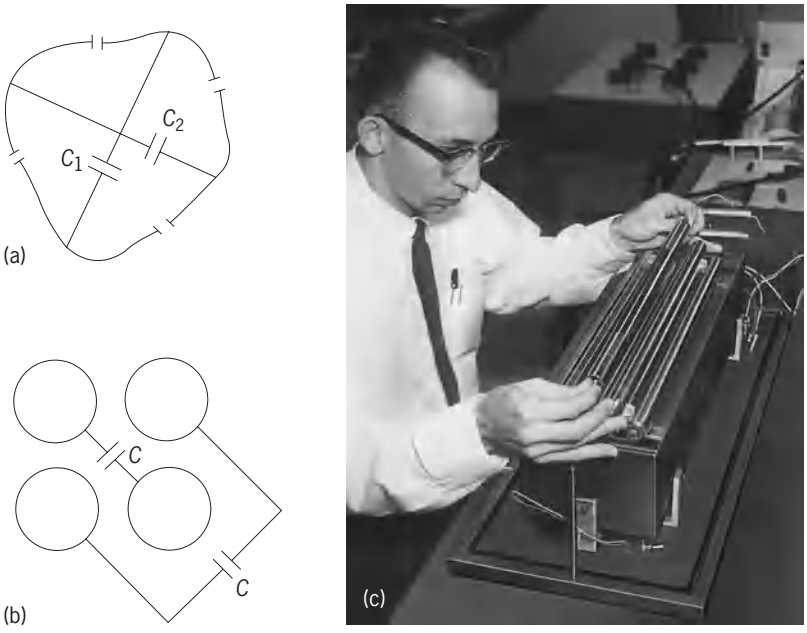


Fig. 1. Calculable capacitor. (*a*) Cross section of general cylindrical capacitor with four segments. (*b*) Cross section of symmetrical capacitor with cylindrical electrodes. (*c*) Calculable capacitor used in absolute determination of the ohm. (*U.S. National Institute of Standards and Technology*)

Absolute Determinations

It has long been the practice that national standards laboratories measure their primary electrical standards in terms of the kilogram and second, with the help of electromagnetic theory and the definition of the ampere. Two different measurements are needed to establish the values of the standards of the volt and the ohm.

For many years the current balance, which measures the force between pairs of cylindrical coils by weighing, has been a mainstay of absolute methods. The other traditional method was the determination of the ohm by comparison with a mutual inductance whose value was calculable from its dimensions. But the accuracy of both measurements has been limited to a few parts in 10^6 by difficulties in determining the dimensions of the coils involved in the experiments and the heating effects of the currents. This order of accuracy is not adequate for modern requirements, and these methods are now of historic interest only. *See* CURRENT BALANCE.

Calculable capacitor. In 1956, A. M. Thompson and D. G. Lampard published a new theorem in electrostatics which showed that if a conducting cylinder of any cross section is divided parallel to the axis into four parts (**Fig. 1***a*) the cross capacitances per unit length, C_1 and C_2, obey Eq. (2).

$$\exp \frac{-\pi C_1}{\epsilon_0} + \exp \frac{-\pi C_2}{\epsilon_0} = 1 \qquad (2)$$

This theorem has been applied as the basis for a capacitor whose change in capacitance as a guard electrode is moved can be calculated exactly from the measurement of a single length. This can be done to a very high degree of accuracy by using a laser interferometer. The problems of making many complex dimensional measurements which limited the accuracy of the current balance and the calculable mutual inductor are thus no longer a limitation. If the capacitor is made symmetrical with four cylindrical electrodes (Fig. 1*b*), then $C_1 = C_2$ and the capacitance per unit length is given by Eq. (3). This is a

$$C = \frac{\epsilon_0(\log_e 2)}{\pi} \quad \text{farads per meter} \qquad (3)$$

little less than 2 picofarads per meter.

An early capacitor of this kind is shown in Fig. 1*c*. By means of a series of bridges, the value of a 1-ohm resistor can be measured in terms of the calculated reactance of the capacitor. In this way an absolute determination of the value of a standard ohm can be made with an uncertainty (1 standard deviation) of about 5 parts in 10^8.

Electrometer. A number of laboratories have made determinations of the volt by an electrometer method, in which the electrostatic potential difference between two electrodes causes a force which can be measured mechanically. W. G. Clothier devised an electrometer in which the lower electrode is the surface of a pool of mercury. This rises against the force of gravity when the potential is applied,

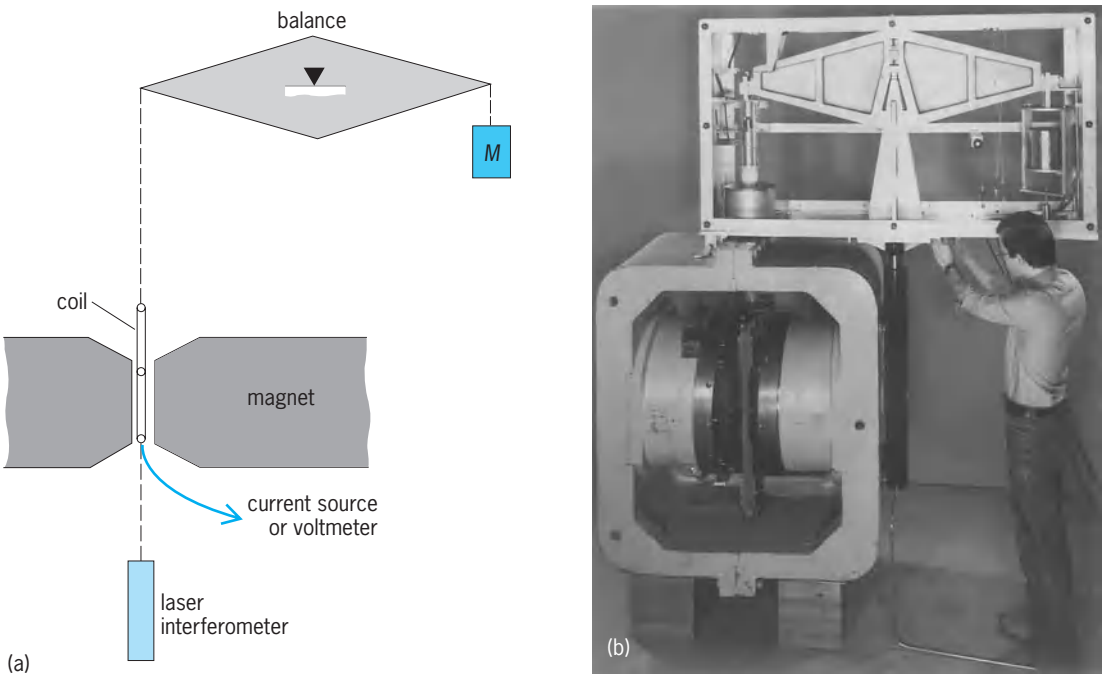


Fig. 2. Moving-coil balance. (*a*) Schematic diagram. (*b*) Equipment. (*U.K. National Physical Laboratory*)

and its movement can be measured very accurately by interferometric means. In this way an uncertainty (1 standard deviation) of about 3 parts in 10^7 has been achieved.

Moving-coil balance. B. P. Kibble developed equipment which in effect measures the electrical watt in terms of the mechanical watt (**Fig. 2**). From this the value of the volt can be determined absolutely. A coil is suspended from the arm of a balance, partly in the field of a powerful magnet. Its position in the vertical plane is measured by a laser interferometer. By using the apparatus successively in two different modes, it is not necessary to know the dimensions of the coil or the strength and distribution of the magnetic field; many errors cancel out, and the problems of accurate dimensional measurements are avoided.

In the first part of the experiment, a current I flows through the coil and the induced force is balanced by a mass M. In the second part, the coil is connected to a voltage-measuring system and made to move through the field with a velocity u, measured by the laser interferometer. The induced voltage V is measured in terms of the primary standard of the volt. Then Eq. (4) follows from the equality of the

$$VI = Mgu \tag{4}$$

electrical and mechanical units of power, where g is the acceleration due to gravity. If the current and the velocity are adjusted so that V is equal to the voltage drop when the current I flows through a resistor whose value R is known accurately in terms of the primary standard of resistance, then V is given by Eq. (5). With equipment based on this principle, an

$$V = (MguR)^{1/2} \tag{5}$$

uncertainty (1 standard deviation) of about 1 part in 10^7 has been achieved.

Josephson effect. If a Josephson junction, consisting of two superconductors separated by an insulating layer so thin that the wave functions are coupled, is irradiated with microwave energy at a frequency ν, it has a voltage–current characteristic that consists of a series of voltage steps (**Fig. 3***a*). The height of these steps is given by Eq. (6), where K_J is known as the

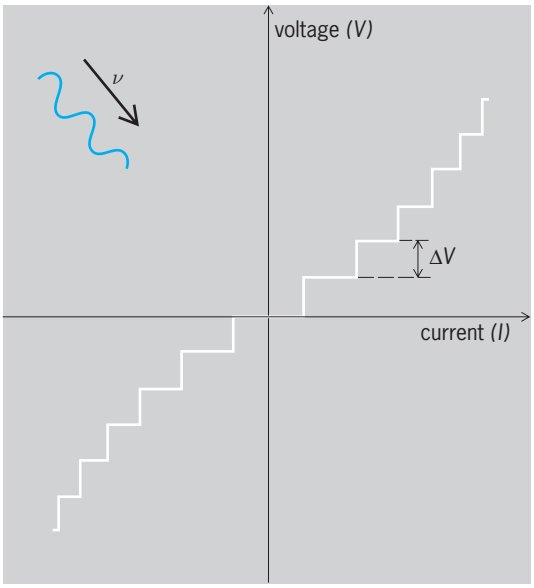
$$\Delta V = \frac{\nu}{K_J} \tag{6}$$

Josephson constant; its value is independent of the materials used and of the experimental conditions. For a practical irradiation frequency the voltage step for a single junction is in the millivolt range, but by using many junctions in series (Fig. 3*b*) it is possible to use the effect to provide a voltage standard in the region of 1 volt and thus to monitor the voltage of a standard cell. *See* JOSEPHSON EFFECT.

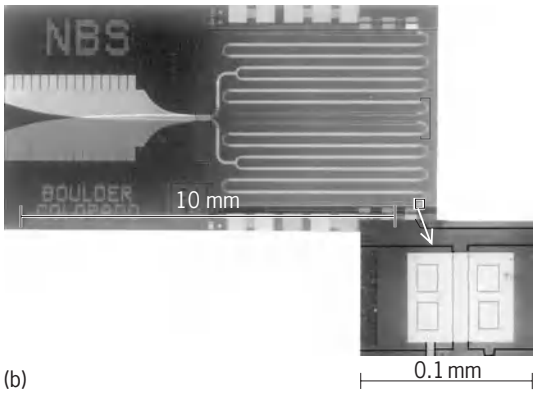
Quantum Hall effect. If an electric current I is passed through a semiconductor placed in a magnetic field of induction B_z, at right angles to the current (**Fig. 4***a*), a voltage V_H is developed across the semiconductor in the direction orthogonal to both the current and the field. This is known as the Hall effect, and the ratio R_H, given by Eq. (7), is known as the Hall resistance.

$$R_H = \frac{V_H}{I} \tag{7}$$

In 1980, K. von Klitzing demonstrated that the Hall resistance in certain semiconductors could be quantized in high magnetic fields at very low temperatures. If the semiconductor in Fig. 4*a* consists of a GaAs/AlGaAs heterostructure, in which a layer of



(a)



(b)

Fig. 3. Josephson effect. (*a*) Josephson junction characteristic. (*b*) Linear array of more than 2000 Josephson junctions in a folded microwave stripline, giving a total voltage step in the region of 1 volt. (*U.S. National Institute of Standards and Technology*)

aluminum gallium arsenide is deposited on a gallium arsenide substrate so that current flow is confined to the interface and is effectively two-dimensional, the Hall resistance takes the form shown in Fig. 4*b* and given by Eq. (8), where $i = 1, 2, 3, \ldots$, and R_K is a

$$R_H = \frac{R_K}{i} \tag{8}$$

constant having a value in the region of 25 kilohms and known as the von Klitzing constant.

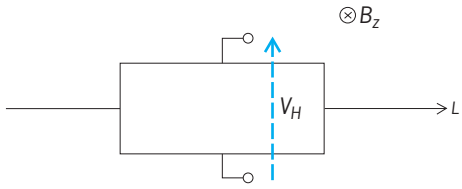
Again, the constant is independent of experimental conditions, and it is possible to set up a system working on a selected plateau of the resistance curve of Fig. 4*b* which acts as a resistance standard and can be used to monitor the values of the primary standards of resistance held by national standards laboratories. *See* HALL EFFECT.

International comparisons. For measurements made in one country to be valid in another, it is essential that national primary standards should be extremely stable and their values accurately known in relation to corresponding standards in other countries. The voltages of Weston cells and the resistances of the best wire-wound resistors drift typically by a few parts in 10^7 per year. This is significantly greater than the precision, a few parts in 10^8, with which the values of different standards can be compared. For many years the BIPM has maintained groups of standard cells and standard resistors, similar to those of the national laboratories. It has been the practice of these laboratories to compare their standards every 3 years with those of BIPM by sending traveling standards to Sèvres. In this way it has been possible to monitor variations in national standards over a period of several decades and to adjust their values when the drift has become excessive.

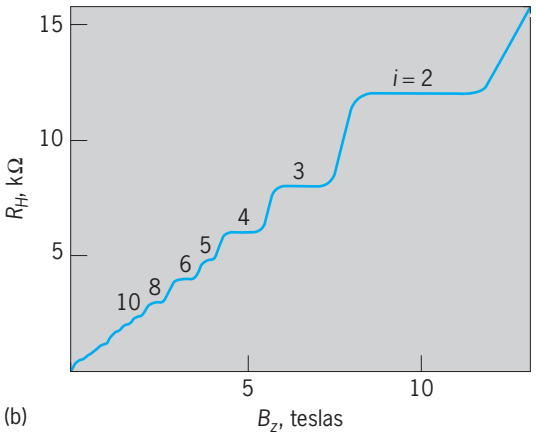
The BIPM also arranges the direct intercomparison of national standards of derived quantities, under the auspices of the CCEM. This has particular value for rf and microwave quantities.

The uncertainties associated with traveling standards and the drifts in value between comparisons have left little margin between national standards and the demands of users for improved accuracy. There are clear advantages if national standards can be referred with sufficient accuracy to some invariant natural physical constant, allowing standards in different countries to be related without transporting them. This has become possible for the volt and the ohm by using the Josephson and quantum Hall effects.

Values of constants. Theory indicates that the values of the Josephson and von Klitzing constants should be equal to $2e/h$ and h/e^2, respectively, where e is the charge on the electron and h is Planck's constant. These are further examples of relations between electrical standards and fundamental physical constants. It might be thought that the effects could be used for absolute determinations of the volt and



(a)



(b)

Fig. 4. Quantum Hall effect. (*a*) Geometry of semiconductor. (*b*) Hall resistance characteristic. i = quantum number.

the ohm. This is not so, because measurements of e and h refer to the electrical standards and neither is known independently with sufficient accuracy. It is rather the case that these relations provide a new route for the measurement of e and h.

The values of k_J and R_H have therefore to be determined in terms of the absolute values of the electrical standards. Much effort has been devoted to these measurements and to obtaining international agreement on the values to be used. *See* FUNDAMENTAL CONSTANTS.

International Agreement

As the basis for a generally accepted system of measurement, it is essential that two requirements are satisfied: stability of the primary standards and international agreement on the values to be assigned to them. The first of these has been met for electrical standards by use of the Josephson and quantum Hall effects. The CCEM made an attempt to meet the second requirement in 1972, when it recommended that the value 483 594.0 gigahertz per volt should be used for the Josephson constant. However, this proved to be premature: several major national laboratories continued to use their own, different values, and subsequent measurements have shown this value to be in error by about 8 parts in 10^6.

Recognizing this, the CCEM encouraged further research and in 1986 set up working groups that made critical assessments of the experimental results bearing on the values of the Josephson and von Klitzing constants. In September 1988 the CCEM recommended that from January 1, 1990, all laboratories should adopt the values of 483 597.9 GHz/V for the Josephson constant (to be designated K_{J-90}) and 25 812.807 Ω for the von Klitzing constant (to be designated R_{K-90}). It recommended that all laboratories should adjust the values assigned to their reference standards accordingly. To ensure consistency, it was recommended that these changes should take place on January 1, 1990, and not before. The CIPM subsequently endorsed the CCEM recommendations.

The CCEM estimated the uncertainties in the recommended values of the constants and expressed them by indicating that the Josephson effect together with the value of K_{J-90} can be used to establish a reference standard of emf having an uncertainty (1 standard deviation) relative to the SI volt of 4 parts in 10^7 and reproducibility significantly better; similarly, the quantum Hall effect, together with the value of R_{K-90}, can be used to establish a reference standard of resistance having an uncertainty (1 standard deviation) relative to the SI ohm of 2 parts in 10^7 and again a reproducibility significantly better.

Implementation of the CCEM recommendations required much work and expense on the part of national standards laboratories and national calibration services. The values of national standards of the volt and the ohm had to be adjusted by up to 10 parts per million for the volt and smaller amounts for the ohm. These changes, of course, also affected the secondary standards of other derived electrical quantities and caused consequential changes in the calibration values assigned to accurate instruments.

Much work is continuing at the BIPM and the national laboratories to improve further the accuracy and availability of electrical standards. Josephson arrays operating at 10 V dc are available, and the BIPM has conducted international comparisons of the volt and the ohm using traveling Josephson array standards and quantum Hall resistance standards. These have led to results differing by no more than 2 and 4 parts in 10^9. The BIPM also continues to carry out a variety of international comparisons of derived quantities. *See* ELECTRICAL MEASUREMENTS.

A. Earle Bailey

Bibliography. R. J. Bell (ed.), *SI: The International System of Units*, 6th ed., 1993; Comité Consultatif d'Electricité et Magnétisme, *Rapport de la 18e Session, 1988, 19e Session, 1992, 20e Session, 1995, 21e Session, 1997, 22e Session, 2000, 23e Session, 2002, 24e Session, 2005*, BIPM, Sèvres, France; B. P. Kibble, Present state of electrical units, *Proc. IEE*, 138A:187–197, 1991; B. W. Petley, Electrical units: The last thirty years, *Metrologia*, 32:495–502, 1995; B. N. Taylor and T. J. Witt, New international electrical reference standards based on the Josephson and quantum Hall effects, *Metrologia*, 26:47–62, 1989.

Electricity

Those physical phenomena involving electric charges, their motions, and their effects. The motion of a charge is affected by its interaction with the electric field and, for a moving charge, the magnetic field. The electric field acting on a charge arises from the presence of other charges and from a time-varying magnetic field. The magnetic field acting on a moving charge arises from the motion of other charges and from a time-varying electric field. Thus electricity and magnetism are ultimately inextricably linked. In many cases, however, one aspect may dominate, and the separation is meaningful. *See* ELECTRIC CHARGE; ELECTRIC FIELD; MAGNETISM.

Historical development. The earliest observations of electrical effects were made on naturally occurring substances. Magnetism was observed in the attraction of metallic iron by the iron ore magnetite. The natural resin amber was found to become electrified when rubbed (triboelectrification) and to attract lightweight objects. Both of these phenomena were known to Thales of Miletus (640–546 B.C.). Jerome Cardan in 1551 first clearly distinguished the difference between the attractive properties of amber and magnetite, thus presaging the division of electrical and magnetic effects. He also envisioned electricity as a type of fluid, a viewpoint that was developed more extensively in the late eighteenth and early nineteenth centuries. In 1600 W. Gilbert observed variations in the amounts of electrification

of various substances. He divided substances into two classes, according to whether they did or did not electrify by rubbing. The division actually is into poor and good conductors, respectively. A two-fluid theory was first proposed by C. F. duFay in 1733. A one-fluid theory of electricity was propounded in 1747 by Benjamin Franklin, who called an excess of the fluid positive electrification, and a deficiency of fluid negative electrification. This theory fell into disrepute, but the choice of positive and negative remains. Although fluid theories of electricity were superseded at the end of the nineteenth century, the concept of electricity as a substance persists.

The quantitative development of electricity began late in the eighteenth century. J. B. Priestley in 1767 and C. A. Coulomb in 1785 discovered independently the inverse-square law for stationary charges. This law serves as a foundation for electrostatics. *See* COULOMB'S LAW; ELECTROSTATICS.

In 1800 A. Volta constructed and experimented with the voltaic pile, the predecessor of modern batteries. It provided the first continuous source of electricity. In 1820 H. C. Oersted demonstrated magnetic effects arising from electric currents. The production of induced electric currents by changing magnetic fields was demonstrated by M. Faraday in 1831. In 1851 he also proposed giving physical reality to the concept of lines of force. This was the first step in the direction of shifting the emphasis away from the charges and onto the associated fields. *See* ELECTROMAGNETIC INDUCTION; ELECTROMAGNETISM; LINES OF FORCE.

In 1865 J. C. Maxwell presented his mathematical theory of the electromagnetic field. This theory proposed a continuous electric fluid. It remains valid today in the large realm of electromagnetic phenomena where atomic effects can be neglected. Its most radical prediction, the propagation of electromagnetic radiation, was convincingly demonstrated by H. Hertz in 1887. Thus Maxwell's theory not only synthesized a unified theory of electricity and magnetism, but also showed optics to be a branch of electromagnetism. *See* ELECTROMAGNETIC RADIATION; MAXWELL'S EQUATIONS.

The developments of theories about electricity subsequent to Maxwell have all been concerned with the microscopic realm. Faraday's experiments on electrolysis in 1833 had indicated a natural unit of electric charge, thus pointing toward a discrete rather than continuous charge. Thus, the groundwork for exceptions to Maxwell's theory of electromagnetism was laid even before the theory was developed. H. A. Lorentz began the attempt to reconcile these viewpoints with his electron theory in 1895. He postulated discrete charges, called electrons. The interactions between the electrons were to be determined by the fields as given by Maxwell's equations. The existence of electrons, negatively charged particles, was demonstrated by J. J. Thomson in 1897 using a Crookes tube. The existence of positively charged particles (protons) was shown shortly afterward (1898) by W. Wien, who observed the deflection of canal rays. Since that time, many particles have been found having charges numerically equal to that of the electron. The question of the fundamental nature of these particles remains unsolved, but the concept of a single elementary charge unit is apparently still valid. Of these many particles only two, the electron and the proton, exist in a stable condition on Earth. *See* BARYON; ELECTROLYSIS; ELECTRON; ELEMENTARY PARTICLE; HYPERON; MESON; PROTON; QUARKS.

A second departure from classical Maxwell theory was brought on by M. Planck's studies of the electromagnetic radiation emitted by "black" bodies. These studies led Planck to postulate that electromagnetic radiation was emitted in discrete amounts, called quanta. This quantum hypothesis ultimately led to the formulation of modern quantum mechanics. The most satisfactory fusion of electromagnetic theory and quantum mechanics was achieved in 1948 with the work of J. Schwinger and R. Feynman in quantum electrodynamics, which suppressed the particle aspect and emphasized the field. *See* HEAT RADIATION; QUANTUM ELECTRODYNAMICS; QUANTUM MECHANICS.

Sources. The sources of electricity in modern technology depend strongly on the application for which they are intended.

The principal use of static electricity today is in the production of high electric fields. Such fields are used in industry for testing the ability of components such as insulators and capacitors to withstand high voltages, and as accelerating fields for charged-particle accelerators. The principal source of such fields today is the Van de Graaff generator. *See* PARTICLE ACCELERATOR.

The major use of electricity today arises in devices using direct current and low-frequency alternating current. The use of alternating current, introduced by S. Z. de Ferranti in 1885–1890, allows power transmission over long distances at very high voltages with a resulting low-percentage power loss followed by highly efficient conversion to lower voltages for the consumer through the use of transformers. *See* ALTERNATING CURRENT; ELECTRIC CURRENT.

The development of materials that are superconducting at liquid-nitrogen temperatures may alter this concept. The cost of liquid nitrogen is trivial compared with the cost of the liquid helium previously required for maintaining superconductivity in materials. If the new materials can be formed into malleable wire, this economic factor could lead to the construction of lossless power lines and other applications in computers and in magnetically levitated transportation. *See* MAGNETIC LEVITATION; SUPERCONDUCTING DEVICES; SUPERCONDUCTIVITY.

Large amounts of direct current are used in the electrodeposition of metals, both in plating and in metal production, for example, in the reduction of aluminum ore. To avoid power transmission difficulties, such facilities are frequently located near sources of abundant power. *See* DIRECT CURRENT; ELECTROCHEMISTRY; ELECTROMETALLURGY; ELECTROPLATING OF METALS.

The principal sources of low-frequency electricity are generators based on the motion of a conducting medium through a magnetic field. The moving charges interact with the magnetic field to give a charge motion that is normal to both the direction of motion and the magnetic field. In the most common form, conducting wire coils rotate in an applied magnetic field. *See* ALTERNATING-CURRENT GENERATOR; DIRECT-CURRENT GENERATOR; GENERATOR.

The rotational power is derived from a water-driven turbine in the case of hydroelectric generation, or from a gas-driven turbine or reciprocating engine in other cases. In the latter situation the high-pressure gas that is required is obtained by heating a fluid or gas, using some thermal source, usually the burning of fossil fuels. Economic considerations, particularly the cost of natural gas and oil and the need to conserve them for petrochemical purposes, have pointed out the need for increased reliance on coal and nuclear reactors as heat sources. *See* ELECTRIC POWER GENERATION; GAS TURBINE; HYDROELECTRIC GENERATOR; NUCLEAR REACTOR.

An alternate method is magnetohydrodynamic generation. The requisite motion is that of a high-temperature gas flowing through the magnetic field. The gas consists of the combustion products of the fossil-fuel heat source with a small amount of potassium added to increase the conductivity. Extensive studies have been conducted, especially in the United States and Russia, to determine the feasibility of this method. Because the gas temperatures are very high, the exhaust gas can be used as a heat source for a conventional generator. It is estimated that a combined facility would be 50–55% efficient. This compares with an efficiency of 35–40% for the conventional facility. An added benefit appears to be a ready solution of the sulfur dioxide (SO_2) emission problem for coal-fueled systems. The potassium would efficiently interact with the sulfur dioxide and yield an easily removable solid residue. *See* ACID RAIN; ENERGY SOURCES; MAGNETOHYDRODYNAMIC POWER GENERATOR; MAGNETOHYDRODYNAMICS.

A more direct method of using fission or fusion reactors is the direct conversion of the energy released in the nuclear process into electricity. This has been achieved on a laboratory scale in the case of fission reactors.

Many high-frequency devices, such as communications equipment, television, and radar, involve the consumption of only moderate amounts of power, generally derived from low-frequency sources. If the power requirements are moderate and portability is needed, the use of ordinary chemical batteries is possible. Ion-permeable membrane batteries are a later development in this line. Fuel cells, particularly hydrogen-oxygen systems, are being developed. They have already found extensive application in earth satellite and other space systems. The successful use of thermoelectric generators based on the Seebeck effect in semiconductors has been reported in Russia and in the United States. In a particularly compact low-power device constructed in the United States, the heat needed for the operation of such a generator has been supplied by the energy release in the radioactive decay of suitably encapsulated isotopes produced in fission reactors. *See* BATTERY; FUEL CELL; ION-SELECTIVE MEMBRANES AND ELECTRODES; NUCLEAR BATTERY; THERMOELECTRIC POWER GENERATOR; THERMOELECTRICITY.

The solar battery, also a semiconductor device, has been used to provide charging current for storage batteries in telephone service and in communications equipment in artificial satellites. *See* SOLAR CELL.

Direct conversion of mechanical energy into electrical energy is possible by utilizing the phenomena of piezoelectricity and magnetostriction. These have some application in acoustics and stress measurements. Pyroelectricity is a thermodynamic corollary of piezoelectricity. *See* MAGNETOSTRICTION; PIEZOELECTRICITY; PYROELECTRICITY.

Some other sources of electricity are those in which charged particles are released with some energy and collected in some manner. Charged particles are suitably released in radioactive decay, in the photoelectric effect, and in thermionic emission, among other ways. The photovoltaic effect may also be in this group. *See* ELECTRON EMISSION; PHOTOVOLTAIC EFFECT; RADIOACTIVITY; THERMIONIC EMISSION.

The differences of work functions of various materials can be used for energy conversion. The contact potential difference may be used to convert heat directly to electricity or to provide improved collection for currents arising from some other source such as radioactivity. *See* WORK FUNCTION (ELECTRONICS).

Other possible sources of electricity arise from the existence of electrokinetic potentials in flowing fluids and of phase-transition potentials such as occur in the Workman-Reynolds effect. The possibilities of combining several effects also exist as exemplified in thermogalvanic potentials. It also appears that organic materials (as distinguished from the inorganic materials for which most of the work already described was done) merit investigation. A primitive type of organic solar battery has been developed. *See* ATMOSPHERIC ELECTRICITY; CIRCUIT (ELECTRICITY); CONDUCTION (ELECTRICITY); ELECTRIC POWER MEASUREMENT; ELECTRICAL ENGINEERING; ELECTRICAL UNITS AND STANDARDS; ELECTROKINETIC PHENOMENA; ELECTRONICS; GEOELECTRICITY.

Walter Aron

Bibliography. P. H. Abelson (ed.), *Energy: Use, Conservation and Supply*, American Association for the Advancement of Science, 2 vols., 1974, 1978; B. I. Bleaney and B. Bleaney, *Electricity and Magnetism*, 2 vols., 3d ed., 1976, paper, 1989; R. P. Feynman et al., *Feynman Lectures on Physics: The Definitive and Extended Edition*, 3 vols., 2005; R. J. Fowler, *Electricity: Principles and Applications*, 6th ed., 2004; E. M. Purcell, *Electricity and Magnetism*, 2d ed., 1985; P. B. Visscher, *Fields and Electrodynamics*, 1988.

Electrochemical equivalent

The weight of a substance, according to Faraday's law, produced or consumed by electrolysis with 100% current efficiency during the flow of a quantity of electricity equal to 1 faraday or 96,485.309 coulombs per mole (1 coulomb corresponds to a current of 1 ampere during 1 second). Electrochemical equivalents are essential in the calculation of the current efficiency of an electrode process.

The electrochemical equivalent of a substance is equal to the gram-atomic or gram-molecular weight of this substance divided by the number of electrons involved in the electrode reaction. For example, the electrochemical equivalent of zinc, for which two electrons are required in order to deposit one atom, is Zn/2 or 65.37/2 g. Thus, the faraday is equal to the product of the charge of the electron times the number of electrons (the Avogadro number) required to react with 1 atom- or molecule-equivalent of substance. The value of the faraday computed in this manner agrees with values obtained from electrochemical determinations. The relative error on the value 96,485.309 coulombs is smaller than ±0.01%. *See* COULOMETER; ELECTROLYSIS.

Paul Delahay

Electrochemical process

The principles of electrochemistry may be adapted for use in the preparation of commercially important quantities of certain substances, both inorganic and organic in nature.

Inorganic Processes

Inorganic chemical processes can be classified as electrolytic, electrothermic, and miscellaneous processes including electric discharge through gases and separation by electrical means. In electrolytic processes, chemical and electrical energy are interchanged. Current passed through an electrolytic cell causes chemical reactions at the electrodes. Voltaic cells convert chemicals into electricity. Electrothermic processes use electricity to attain the necessary temperature for reaction. For related information *see* ELECTROCHEMISTRY; ELECTROLYSIS; ELECTROLYTIC CONDUCTANCE; ELECTROMOTIVE FORCE (CELLS).

For a discussion of equipment used to convert alternating current to direct current for electrolytic plants. *See* MOTOR-GENERATOR SET; SEMICONDUCTOR RECTIFIER; SYNCHRONOUS CONVERTER.

Voltaic cells are used for the intermittent production of small amounts of electricity. When the chemicals involved are exhausted and must be replaced, the unit is called a primary cell. A special case of the primary cell is the fuel cell in which the fuel and anoxidizer are fed continuously to the cell, converted to electricity, and the products removed. If exhausted components can be revived by passing electricity backward through the unit, it is called a secondary cell, storage battery, or accumulator. Cells may be connected in parallel or in series to form a battery.

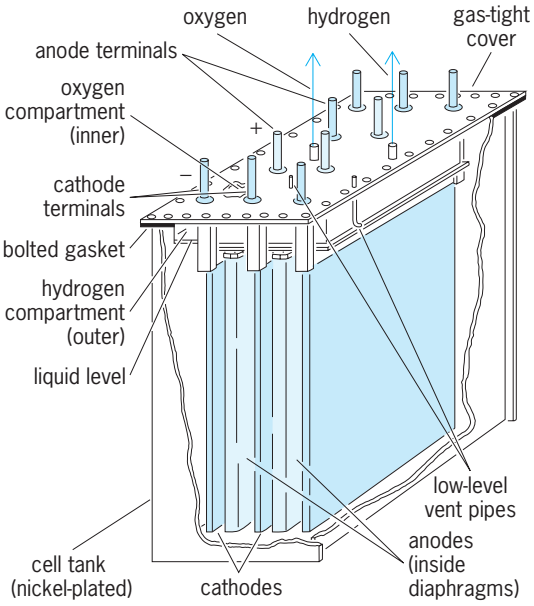


Fig. 1. Diagram of Stuart hydrogen-oxygen cell. (*Electrolyser Corp. Ltd., Toronto, Canada*)

For a discussion of the theory and description of commercial primary, secondary, and fuel cells *see* BATTERY; FUEL CELL.

Electrolysis in aqueous solutions. The electrolysis of water to form hydrogen and oxygen, according to the reaction $2H_2O \rightarrow 2H_2 + O_2$, may be considered as the simplest process for aqueous electrolytes. It does not compete with hydrogen from propane or from natural gas and with oxygen from liquid air, exceptin small installations. While simplicity, high hydrogen purity requirement, and lower capital cost (in small plants) have justified electrolytic plants, severely rising energy costs limited such applications. The electrolyte is 18–30% NaOH or KOH, the latter having a lower resistance but higher cost. The cathode is steel and the anode is nickel-plated steel separated by a diaphragm, usually of asbestos (**Fig. 1**). A cell voltage of 2.0–2.5 V is the summation of the decomposition voltage of water (1.23 V at room temperature) and the oxygen and hydrogen over voltages, plus the *IR* drop through the electrolyte, electrode contacts, and bus bars. The raw material is distilled or demineralized water. There are monopolar cells operating up to 20,000 A, bipolar or filter-press cells using 2000–5000 A, and 150 cells in series at 300–400 V overall. Cells are available which operate at 600 psi (4.1 megapascals). A. T. Kuh described 11 cells using 25–30% KOH and indicated designs operating at pressures up to 200 atm (20 MPa). *See* HYDROGEN; OXYGEN.

Heavy water, or deuterium oxide, used in moderating nuclear reactors is also a by-product of the electrolysis of water. Protium (H^1) is preferentially discharged, so that the electrolyte becomes richer in deuterium. Electrolysis must be combined with catalytic exchange or distillation processes when used

in primary or earlier stages of concentration; it is also used in final stages to produce 99–100% concentration. *See* DEUTERIUM; HEAVY WATER.

Metallurgical applications. Protective or decorative coatings on a base metal such as steel are obtained by electroplating. Plating may also be used to replace worn metal or to provide a wear-resistant surface. The final surface may require several layers of different metals or even layers of the same metal deposited under varying conditions. The metals plated are copper, cadmium, chromium, cobalt, gold, iron, lead, nickel, the platinum metals, silver, tin, andzinc, and many alloys. Electrogalvanizing is preferred over hot dipping for applying zinc to steel. Tin plate for containers is electrolytic. Factors affecting the resulting plate include pretreatment and cleaning of the metal surface, current density, concentration of metal ions, agitation, temperature, conductance of solution, pH, and addition agents. *See* ELECTROPLATING OF METALS; METAL COATINGS.

Electroforming is a method of forming or reproducing articles by electrodeposition. In contrast to electroplating, the product is removed from the base surface or mold. A nonconducting surface can be made conductive by metallizing or with graphite powder. A low-melting-point metal mold can be used, or a mold can be plated with a metal from which the final metal can be removed. The electrodeposits may be up to 0.5 in. (13 mm) thick, and after completion are removed from the mold. Phonograph records, electrotype, textile replicas, and brass instrument bells are common products of electroforming, which is also widely used in the automotive and electronic industries.

Electrodeposition of metal powders is used to produce particles in the 1–1000-micrometer range for use in powder metallurgy and metallic pigments. Powdered iron made electrolytically makes stronger parts than other types, and electrolytic powdered copper makes parts that are easier to machine and have improved wear.

In the anodizing of aluminum articles, a coating approximately 0.001 in. (25 μm) thick is applied, which can be dyed or made impervious. The article to be anodized is cleaned and made anodic in sulfuric acid solution.

Electrolytic polishing of metals is accomplished by making the article anodic in an electrolyte of mixed acids, such as phosphoric-chromic acids or phosphoric-sulfuric acids. High points on the surface apparently dissolve, to give a unique polish not achievable by mechanical polishing.

Electrolytic machining of metals is accomplished by making the metal part anodic in a suitable electrolyte. Metal dissolves at the anode and hydrogen discharges at the cathode. The cathode is contoured as a negative of the shape to be developed in the work piece. By circulating the electrolyte under very high pressure and using high current densities (30–2000 A/in.^{2} or 5–300 A/cm^2), practical machining rates are achieved. Very hard and very thin metal surfaces can be machined without changing the heat treatment.

Electrorefining is a process for purifying metals and recovering their impurities, which at times are more valuable than the original metal. Copper from its ore or scrap is purified of volatilizable impurities, cast into an anode, dissolved in an electrolyte in a cell, and deposited at a cathode as a very pure, highly conductive metal for electrical engineering purposes (**Fig. 2**). Gold, silver platinum, selenium, and tellurium are recovered as by-products. Nickel is freed of copper by using a diaphragm between anode and cathode, with cobalt and the platinum

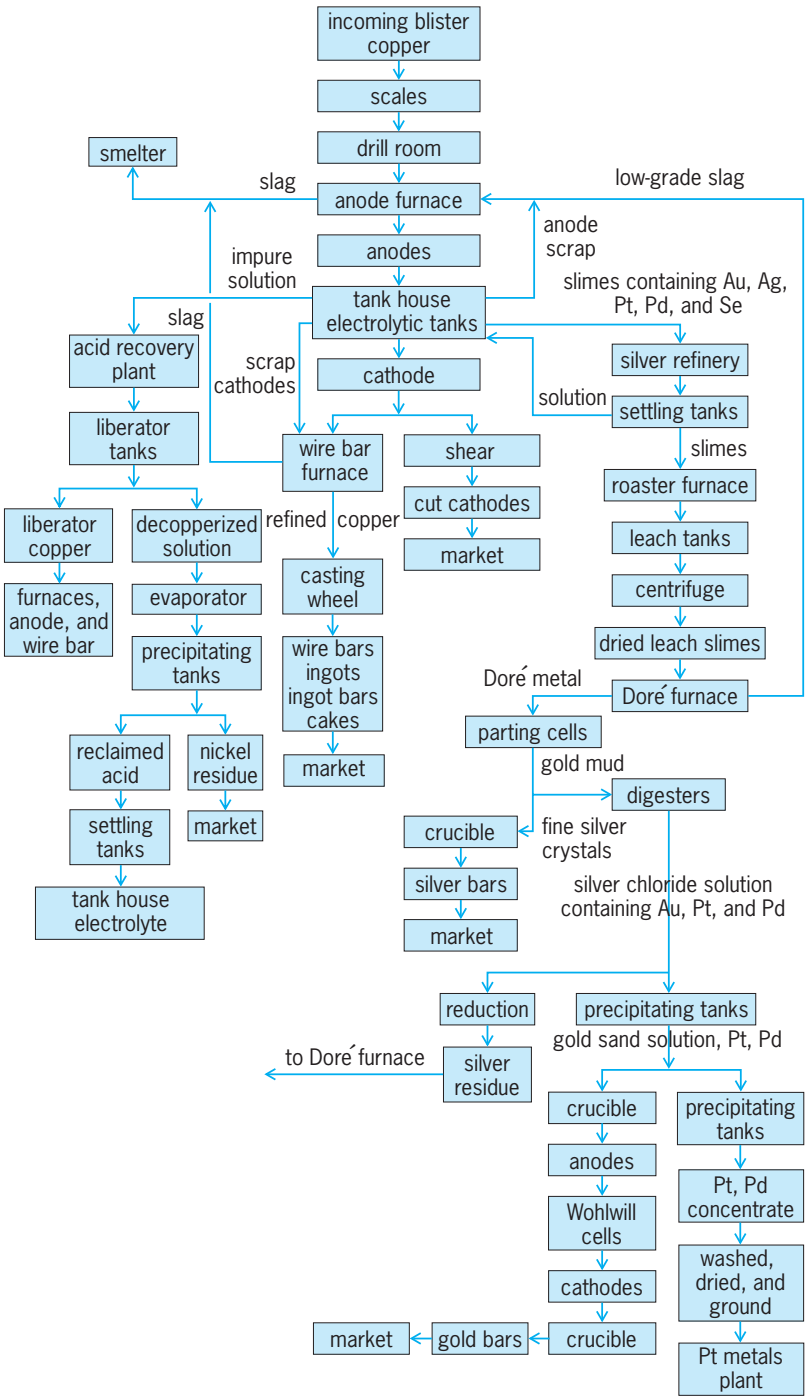


Fig. 2. Flowsheet for a copper plant and the refinery by-products resulting from a product of 99.97% conductor copper. (*Ontario Refining Co.*, *International Nickel Co. of Canada, Ltd.*)

TABLE 1. Electrolytics for various individual metals

Metal	Source	Method	Electrolyte	Application
Antimony	Crude metal	Refining	Sulfates	Alloys, batteries, chemicals
Antimony	Antimony ores	Winning	Sulfides	
Bismuth	Lead refining slimes slimes	Refining	Chlorides	Alloying agent, medicine
Cadmium	Zinc residues and slimes	Winning	Sulfates	Electroplating, alloys
Chromium	Chromium ores	Winning	Sulfates	High-temperature, alloys
Cobalt	Complex cobalt ores	Winning	Sulfates	Alloying agent, electronics
Cobalt	Complex nickel ores	Refining	Sulfates and chlorides	Alloying agent, electronics
Copper	Copper ores, complex ores	Winning	Sulfates	Electrical conductors, wire, brass alloys
Copper	Crude metal, secondary materials, waste	Refining	Sulfates	Electrical conductors, wire, brass alloys
Copper powder	Refined metal	Plating	Sulfates	Powders for powder metallurgy, oilless bearings
Copper sheet	Purified electrolyte	Plating	Sulfates	Sheet for printed circuits, electronics
Gallium	Sodium aluminate liquors	Winning	Caustic	Low-melting-point metals and alloys
Gold	Copper refining slimes	Refining	Chlorides	Jewelry, dentistry, plating
Indium	Residues, wastes, ores	Refining	Chlorides	Silver alloys, jewelry, television
Iron	Low-carbon steel	Refining	Sulfates	Powder metallurgy
Lead	Crude lead	Refining	Fluosilicates	Separation of bismuth etc., alloys, fittings
Manganese	Manganese ores	Winning	Sulfates	Stainless steels, carbon-free metal, alloys
Nickel	Crude metal, nickel matte	Refining	Sulfate-chloride	Plating, alloys, stainless steel
Silver	Copper refinery slimes, lead residues, crude metal, ore	Refining	Nitrates	Jewelry, electrical applications, alloys
Silver	Silver alloys, photo-graphic wastes	Winning	Nitrates	Jewelry, electrical applications, alloys
Solder	Waste and crudes	Refining	Fluosilicate	Metal joining, electrinic components
Tin	Crude metal	Refining	Cresol-sulfonates	Tin plate, solder, alloys
Zinc	Zinc ores, complex ores	Winning	Sulfates	Die-casting alloys, battery cups, brass, galvanizing
Zinc	Ores and residues	Winning	Caustic	Chemicals, paint

metals as by-products. Lead is freed from bismuth; tin from lead, antimony, and bismuth with silver as a by-product; and zinc from copper and lead, with cadmium as a by-product, to make commercial a 99.99% pure metal for diecasting. The last operation is usually referred to as electrowinning. This has rapidly replaced pyrometallurgy or fire processing because it can eliminate sulfur dioxide atmospheric pollution and contamination by particulates.

Electrowinning, sometimes termed aqueous electrometallurgy, involves processing of metallic ores, usually of very low metal content but large in volume, by leaching solutions, usually sulfates, to obtain metal-containing electrolytes which can be processed with insoluble anodes and metal cathodes. Examples are found in cadmium, copper, cobalt, manganese, and zinc.

Electrorefining, electrowinning, and electroforming are summarized as to the individual metals in **Table 1**, while **Table 2** gives anodes, cathodes, and diaphragms for commercially successful operation. **Table 3** lists energy consumption of aqueous electrochemical operations. *See* ELECTROMETALLURGY.

Electrolytic corrosion of metals. This occurs because some parts of the surface of metals act as

TABLE 2. Anodes, cathodes, and diaphragms used on commercial processes

Metal	Method	Anode	Anolyte-catholyte	Diaphragm	Cathode	Voltage
Antimony	Winning	Insoluble	Same	No	Steel	2.5–3
Cadmium	Winning	Insoluble	Same	No	Aluminum	4.0
Chromium	Winning	Insoluble	More acid–Acid	Yes	Hastelloy	4.2
Cobalt	Refining	Soluble	Same	No	Cobalt-stainless steel	2.5
Copper	Refining	Soluble	Same	No	Copper	0.2–0.3
Copper	Winning	Insoluble	Same	No	Copper	2–2.1
Gold	Refining	Soluble	Same	No	Gold	0.5–2.8
Lead	Refining	Soluble	Same	No	Lead	0.35–0.45
Manganese	Winning	Insoluble	Acid–Alkaline	Yes	Stainless steel	
Nickel	Refining	Soluble	Same–Pure	Yes	Nickel	2.4
Silver	Refining	Soluble	Same	Yes	Stainless steel or carbon	1.3–5.4
Tin	Refining	Soluble	Same	No	Tin	0.3
Zinc	Winning	Insoluble	Same	No	Aluminum	3.25–3.7

anodes and corrode, whereas other parts act as cathodes and do not corrode.

Cathodic protection is provided if the whole surface is made cathodic to a separate anode and sufficient voltage is available between the two electrodes. This type of protection is used to inhibit corrosion of boilers, condensers, underground pipelines, ships, and water tanks. Sacrificial anodes of zinc, magnesium, or aluminum alloy may provide the potential, or inert anodes such as graphite, stainless steel, or platinum-plated titanium may be used with power supplied from a rectifier.

Anodic protection can be used to create a passive layer on the surface of some metals, such as steel and stainless steel in some environments. It is a practical method of controlling corrosion of tanks in the chemical industry, but is not feasible for copper or brass vessels. The tank is made anodic. An inert cathode, such as platinum-clad metal, is installed in the liquid in the tank. Current is applied so as to maintain a predetermined voltage between the anodic surface and a reference electrode, such as silver–silver chloride, in the liquid. Equipment to maintain precise potential control at high current output has made anodic protection practical. For example, a current of 0.0015 A/ft^2 (0.016 A/m^2) at +0.900 V to a

TABLE 3. Energy consumption of electrochemical products*

Industry	kWh/lb	lb/kWh	Voltage/tank, cell, or furnace	Range, A/unit	Line voltage	Range, kVA/cell or furnace	kVA/line
Electrolytic refining							
Copper: multiple system,	0.09–0.2	5–13.3	0.2–0.3	6000–15, 000	80–200	1.2–4.5	480–3000
series system	0.074	13.5	120–200				
Gold (troy lb)	0.15	6.6	0.5–2.8	150–500	3–10	0.1–1.4	0.5–5
Lead	0.07–0.09	11.1–14.3	0.35–0.5	5000–6000	100–185	1.75–3.0	500–1200
Nickel	1.1	0.9	2.4–2.6	5000–6000	220–230	12–16	1300–1500
Silver, Moebius (troy lb)	0.27–0.3	3.3–3.7	2.3–2.8	400–500	45–250	1–1.4	18–125
Silver, Thum (troy lb)	0.33–0.6	1.67–3.0	1.3–5.4	150–200	200–220	0.3–1.2	30–50
Solder	0.08	12.5	0.34				
Tin	0.085	11.8	0.3–0.35	5000	100–200	1.5–2.0	500–1000
Electrowinning							
Antimony			2.5–3	1500	300	4	450
Cadmium	0.65–0.97	1.03–1.54	2.5–3.1				
Chromium	5.0–8.4	0.12–0.2	4.2–4.3	10,000	250–300	40–45	2500–3000
Cobalt	1.2–1.56	0.64–0.83	3.7–4.5				
Copper	0.89–1.34	0.77–1.12	2.0–2.12	10,000–25,000	150–200	20–30	1500–5000
Manganese	4–4.5 dc	0.22–0.25	5–5.4	6000	600	30–40	3600
	5–5.3 dc	0.19–0.2					
Silver	0.58–1.0	1–1.73	1–1.5	300		0.3–0.5	0.3–0.5
Zinc	1.4–1.6	0.61–0.71	3–3.7	5000–10,000	600–800	15–40	3000–8000
Metal melting							
Copper	0.12–0.15	6.67–8.33	85–225	12,000–30,000		6000	
Copper alloys	0.15–0.3	3.33–6.67					
Steel: cold charge, hot	0.237–0.35	2.86–4.22	80–250	30,000–100,000		25,000–33,000	
charge	0.05–0.2	5–20	80–250	30,000–100,000		25,000–33,000	
Zinc	0.045	22.2					
Metal powders							
Copper	0.3–0.5	2–3.3	0.3–0.5	10,000–15,000			
Iron	1.20	0.83	2.5				
Nickel	1.13	0.83	1.4–1.5				
Zinc	1.37	0.73	3.4				

*1 kWh/lb = 7.937 MJ/kg; 1 lb/kWh = 0.1260 kg/MJ.

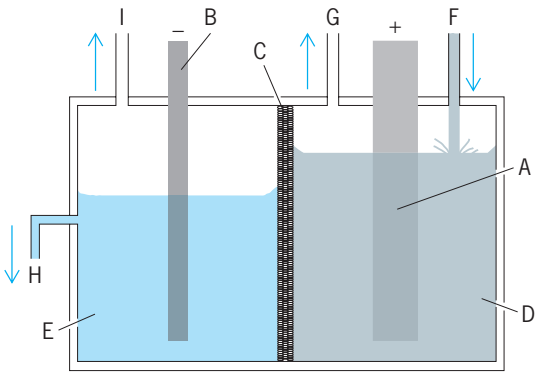


Fig. 3. Diagram of diaphragm cell for chlorine and caustic soda. A = graphite anode, B = iron screen cathode, C = asbestos diaphragm, D = anode compartment for brine and chlorine, E = cathode compartment for NaOH-NaCl cell liquor, F = brine inlet, G = chlorine outlet, H = cell liquor (caustic soda) outlet, I = hydrogen outlet.

silver–silver chloride electrode will maintain passivity of a carbon-steel tank holding 93% sulfuric acid at 27°C (80°F). *See* CORROSION.

Alkali-chlorine processes. Electrolysis of alkali-halides is the basis of the alkali-chlorine and chlorate industries. Chlorine, Cl_2, and caustic soda, NaOH (or caustic potash, KOH), are made by electrolysis of brine, a solution of sodium chloride, NaCl, in water. This is represented by reaction (1).

$$2NaCl + 2H_2O \xrightarrow{\text{electricity}} 2NaOH + Cl_2 + H_2 \qquad (1)$$

Two processes are use to prevent the products from the diaphragm cell and the mercury cell. In the diaphragm cell process (**Fig. 3**) an asbestos diaphragm is interposed between a graphite anode and an iron screen cathode. Saturated purified brine fed around the anode passes through the diaphragm to the cathode. Chlorine is formed at the anode. Hydrogen is released at the cathode, leaving NaOH as a 10–15% solution and 10–15% residual NaCl crystallizes out and is recycled. The decomposition voltage of brine to form chlorine and hydrogen is 2.3 V. At 0.75 A/in.^{2} (11.6 A/dm^2) the average voltage components of a diaphragm cell are as follows:

Anode potential	1.50 V
Cathode potential	1.25
Anolyte	0.47
Diaphragm	0.30
Conductors	0.18
Total	3.70 V

The current efficiency is 95.5–95.5%, due to some oxygen discharge at the anode and some chlorine being carried through the diaphragm. An installation is pictured in **Fig. 4**.

In the mercury cell process brine is electrolyzed between graphite anodes and a flowing mercury cathode, forming a dilute (0.2–0.4%) sodium amalgam which is decomposed in another compartment by water in contact with graphite surfaces to form H_2 and NaOH (**Figs. 5–7**). The products of the mercury cell are purer than those of the diaphragm cell. To offset the cost of mercury, a much higher current density is used in mercury cells. Typical components of voltage in a mercury cell at 5.12 A/in.^{2} (80 A/dm^2) are as follows:

Anode potential, reversible	1.34 V
Cathode potential, reversible	1.76
Decomposition voltage	3.10 V
Anode polarization	0.35 V
Cathode polarization	0.06
Electrolyte	0.60
Conductors and contacts	0.29
Total cell voltage	4.40 V

Economic factors dictate the use of higher current densities, equal to or exceeding 6.5 amperes/in.^{2} (100 A/dm^2). Cell voltage at these higher current densities can be calculated for good cell designs on the market from the equation: $V = 3.20 + 0.015\,C$, where C is the cathode current density in amperes per square decimeter. The current efficiency is approximately 95%. Inefficiency reactions are demonstrated in reactions (2) through (6). Adverse conditions can increase these inefficiencies. *See* CHLORINE.

$$2OH^- \longrightarrow {}^1/_2O_2 + H_2O + 2e \quad \text{(at anode)} \qquad (2)$$

$$H^+ + e \longrightarrow {}^1/_2H_2 \quad \text{(at cathode)} \qquad (3)$$

$$Cl_2 + 2Na(Hg) \longrightarrow 2NaCl + (Hg) \quad \text{(at cathode)} \qquad (4)$$

$$Cl_2 + H_2O \longrightarrow Cl^- + ClO^- + 2H^+ \qquad (5)$$

$$3ClO^- \longrightarrow ClO_3^- + 2Cl^- \qquad (6)$$

Fig. 4. Photograph of an installation of diaphragm alkali-chlorine cells. (*Hooker Chemical Corp.*)

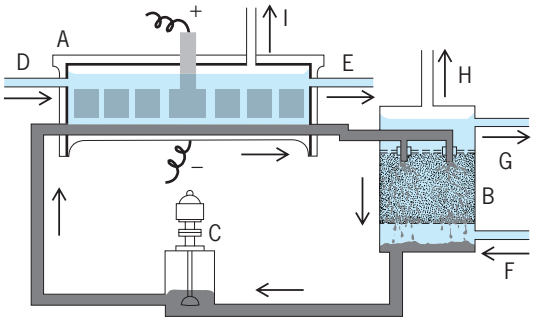


Fig. 5. Diagram of a mercury alkali-chlorine cell. A = electrolyzer, B = decomposer with graphite packing, C = mercury pump, D = feed brine, E = spent brine, F = water, G = 50% caustic soda, H = hydrogen, I = chlorine.

Two factors have been important in the development of chlorine technology: (1) The Nafion diaphragm of the synthetic resin type replacing the deposited asbestos diaphragm. (2) The dimensionally stable anode replacing the graphite anode. The dimensionally stable anode is a titanium substrate with a platinum-group coating of metals and oxides. Its use eliminates the continuously necessary voltage increase or anode-cathode spacing adustment needed because of graphite-anode wear.

In addition, there has been a swing away from mercury cells because of widespread publicity of so-called mercury poisoning by mercury discharges. These discharges have been reduced by better hydrogen cooling and recycle of metal dross, cutting mercury losses as well as permitting better "housekeeping."

Sodium hypochlorite is formed when the products of the electrolysis of brine are mixed. Electrolytic cells have been built for this purpose, but have limited or special use, such as for sterilization of swimming pools and algae control in power plant condensers. Sodium hypochlorite is usually made chemically.

Sodium chlorate is made in cells with graphite or lead peroxide anodes and steel cathodes. When mixing is encouraged, changes take place according to reactions (7)–(9). The overall reaction is (10).

$$6NaCl + 6H_2O \xrightarrow{\text{electricity}} 6NaOH + 3Cl_2 + 3H_2 \quad (7)$$

$$6NaOH + 3Cl_2 \longrightarrow 3NaClO + 3NaCl + 3H_2O \quad (8)$$

$$3NaClO \longrightarrow NaClO_3 + 2NaCl \quad (9)$$

$$NaCl + 3H_2O \xrightarrow{\text{electricity}} 3H_2 + NaClO_3 \quad (10)$$

The temperature is kept below 40°C (104°F) in cells using graphite anodes to prevent excessive attack. The optimum efficiency is at pH 6.8; hydrochloric acid is added as required. Sodium dichromate prevents reduction of chlorate at the steel cathode. The conversion of hypochlorite to chlorate is a some what slow chemical reaction, occurring partly in the cells and partly in a rundown tank. Salt is added and electrolysis is continued until the sodium chloride is down to about 100 g/liter and the chlorate has reached the desired concentration. It is then recovered by crystallization. Cells operate at 3–3.5 V, 1–3 A/dm^2, and 80–85% current efficiency. Energy consumption is about 2.5 kWh/lb (20 megajoules/kg).

Hydrochloric acid electrolysis is of interest for recovery of chlorine from HCl resulting as a by-product from organic chlorinations. A filter-press electrolyzer is used which has 30–50 unit cells with polyvinyl-cloth diaphragms and graphite electrodes. Hydrochloric acid of 30–33% concentration is fed to the anode compartment. Weak acid is withdrawn at about 20% and reconcentrated by absorbtion of HCl gas. The graphite anode is not attacked as long as the concentration is kept at 20% HCl or higher. The current efficiency is 92–96%, the loss due to electrical leakage. Energy consumption is 1800 kWh/2000 lb (7.1 MJ/kg) chlorine (direct current). The voltage balance of a unit cell is as follows:

Anode potential	1.02 V
Cathode potential	0.28
Anode polarization	0.2
Cathode polarization	0.5
Electrolyte, diaphragm	0.3
Total	2.30 V

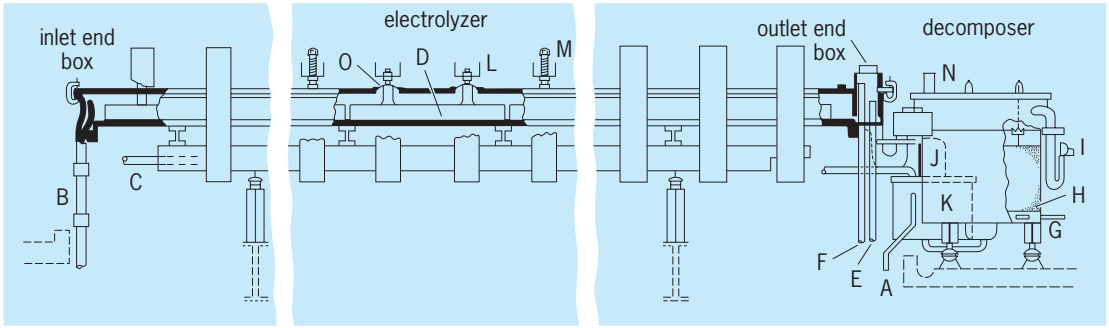


Fig. 6. Longitudinal section of Olin Mathieson E-11 mercury cell. A = dilute caustic outlet, B = brine inlet, C = mercury return, D = anode, E = brine-chlorine outlet, F = outlet end box vent, G = water inlet, H = graphite packing, I = caustic outlet, J = mercury pump, K = mercury pump sump, L = anode support bus, M = lifting screws, N = hydrogen outlet, O = anode seal. (*Olin Mathieson Chemical Corp.*)

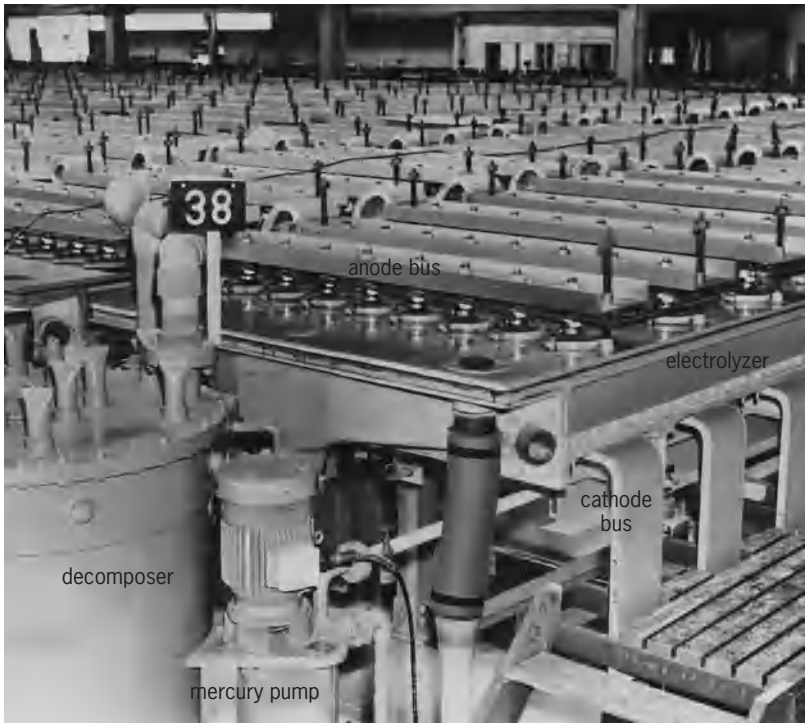


Fig. 7. Chlor-alkali mercury cell; 300,000-A capacity. (*Olin Mathieson*)

Oxidations and reductions. These reactions occur in all cells, but in a narrower sense oxidation reactions are those in which oxygen or chlorine at the anode oxidizes some material to form a new compound; reduction reactions are those in which hydrogen, liberated at the cathode, reduces a material to a new product. There are no commercial applications of inorganic electrochemical reductions by this narrow definition.

Sodium perchlorate is made by oxidation of a solution containing $NaClO_3$ at pH 6.1–6.4 by use of a platinum anode and an iron cathode, with chromate in the electrolyte. A lead dioxide anode may be used with a stainless-steel or nickel cathode, with no chromate in the electrolyte. Energy consumption is 1.4–1.6 kWh/lb (11–13 MJ/kg) $NaClO_4$ (direct current). Other perchlorates are made by metathesis with $NaClO_4$.

Persulfuric acid, $H_2S_2O_8$, is made by oxidizing sulfuric acid as an intermediate in the production of hydrogen peroxide. The reactions for the process are shown in reactions (11) and (12). Alkali persulfates

$$2HSO_4^- \xrightarrow{\text{electricity}} H_2S_2O_8 + 2e \qquad (11)$$

$$H_2S_2O_8 + 2H_2O \longrightarrow 2H_2SO_4 + H_2O_2 \qquad (12)$$

can be made in the same way. The cell has smooth platinum anodes, a porous stoneware diaphragm, and a lead cathode cooled to 30°C (86°F). The reversible potential is 2.18 V and the operating voltage 5.0–5.5 V. The energy released as heat in the cell must be removed by cooling with stoneware or glass coils in the cell. Hydrogen peroxide is recovered by distillation. *See* PEROXIDE.

Lead peroxide anodes are used in sodium chlorate, sodium perchlorate, sodium bromate, and periodate orperiodic acid regeneration cells. The material is dense and made from a lead nitrate solution. For some uses it is produced on a steel base, from which it is removed mechanically and chemically. It is also applied to tantalum, platinum-clad tantalum, or graphite.

Periodic acid is used in producing dialdehyde starch, and spent solution from oxidation can be regenerated in a cell by use of lead dioxide anode, a porous ceramic diaphragm, and an iron cathode.

Electrolytic manganese dioxide for batteries is made by electrolyzing hot $MnSO_4$ solutions at pH 6.5–7.5 by use of graphite electrodes. The MnO_2 deposited on the anode is pulverized with the graphite and separated mechanically. Energy consumption is 1 kWh/lb (8 MJ/kg) MnO_2. The quality of the MnO_2 for battery use depends on cell temperature and anode current density.

Ion-permeable membrane cells. These utilize diaphragms made of ion-exchange resins. Cation-permeable membranes permit cations to pass through but not anions, where as the reverse holds for anion-permeable membranes. **Figure 8** shows the movement of ion and water in an electric membrane stack. Purification of seawater is the most important application. Salt has been recovered from seawater that has been concentrated in this way. *See* ION-SELECTIVE MEMBRANES AND ELECTRODES; WATER DESALINATION.

Fused-salt electrolysis. Aluminum, barium, beryllium, cerium and misch metal, fluorine, lithium, magnesium, sodium, molybdenum, thorium, titanium, uranium, and zirconium are obtained by electrolysis of fused salts, because water interferes with the desired reaction. Raw materials must all be purified before addition to fuse-salt electrolytes as inaqueous electrolytes. Metallizing is a process of depositing a metal as an alloy on a substrate from a fused complex metal salt.

Aluminum is produced in steel pots, lined with carbon, graphite, or silicon carbide, containing an electrolyte of alumina dissolved in fused cryolite, $AlF_3{\cdot}3NaF$, at 950–1000°C (1740–1830°F). The pool of aluminum in the bottom of the pot is cathode. Contact is made by iron bars buried in the carbon lining or by titanium diboride contacts. Anodes are prebaked carbon blocks or the Söderberg type made from carbon paste baked in place. The arrangement of the apparatus for aluminum production is shown in **Fig. 9**. Aluminum is siphoned out periodically. Oxygen released at the carbon anode forms carbon monoxide. *See* ALUMINUM.

Barium, used in the electronics industry and in alloys, is obtained by aluminum reduction at low pressures and high temperatures.

Beryllium is made by batch electrolysis of fused salt, starting with 25% $BeCl_2$ and 75% NaCl in a chrome-iron pot, which acts as cathode to a graphite anode. Beryllium is deposited on the wall of the pot and is cleaned out and broken up when cold. Salt is

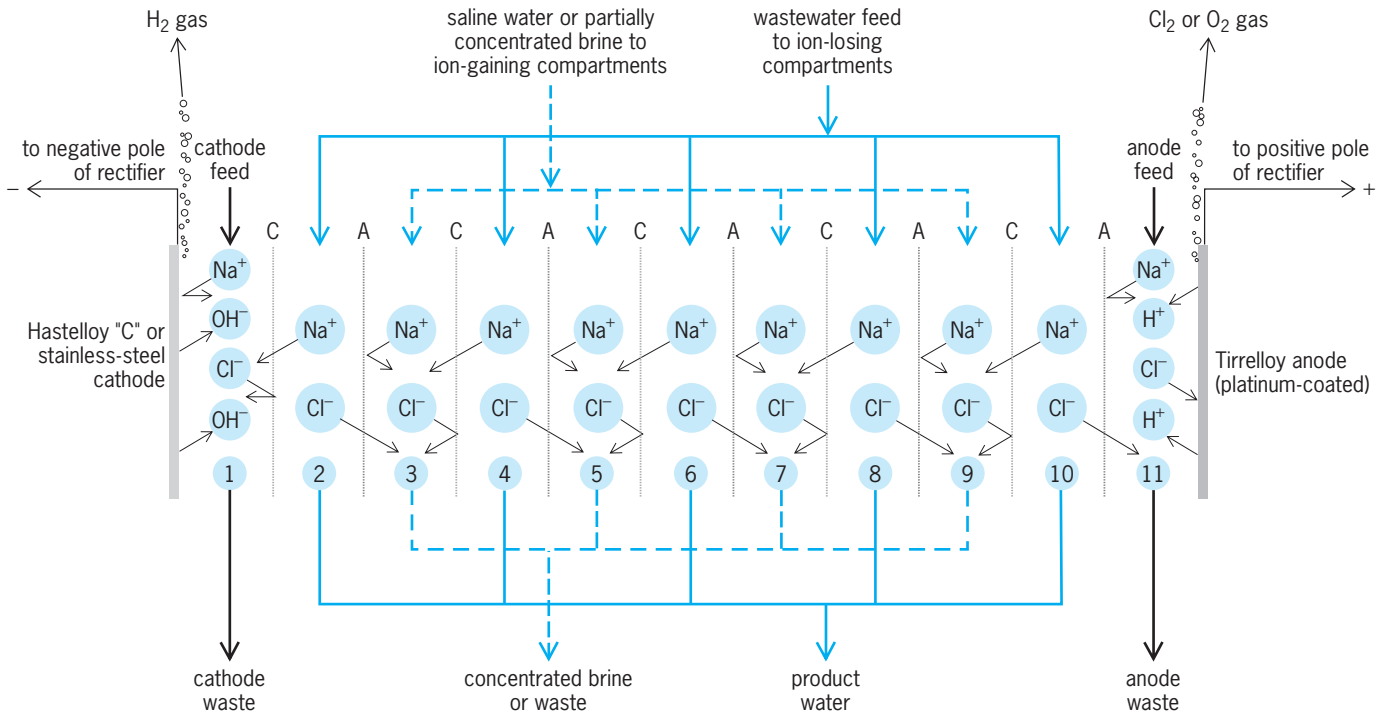


Fig. 8. Diagram of the basic ion and water flow in electric membrane stack. C = cation membrane; A = anion membrane; Na^+ = any cation, such as sodium; Cl^- = any anion, such as chloride. Numbers = compartments.

washed out with water. The metal is in the form of bright crystalline flakes. A beryllium-copper eutectic can be made by using a copper cathode. Beryllium alloys containing copper and nickel are made from BeO in arc furnaces.

Calcium has been made by electrolysis of pure fused calcium chloride about 800°C (1470°F). In this case the cathode is solid calcium, and is mechanically withdrawn from the cell as a "carrot." Calcium is now made by aluminum reduction at low pressures and high temperatures.

Cerium and misch metal are made from $CeCl_3$ or mixtures of chlorides of cerium, lanthanum, and neodymium in fused-salt electrolysis with NaCl. Misch metal is used for lighter flints.

Fluorine for separation of uranium isotopes is produced by electrolysis of 40% HF in KF between carbon anodes and steel cathodes at 88–100°C (190–212°F). A diaphragm of Monel screen keeps the products H_2 and F_2 separated. Dry HF gas bubbled continuously into the electrolyte. At a current density of 1 A/in.^{2} (15 A/dm^2) the cell operates at 9–12 V and 96% current efficiency. Energy consumption is 3.0 kWh/lb (24 MJ/kg) fluorine. The theoretical decomposition potential is 2.85 V.

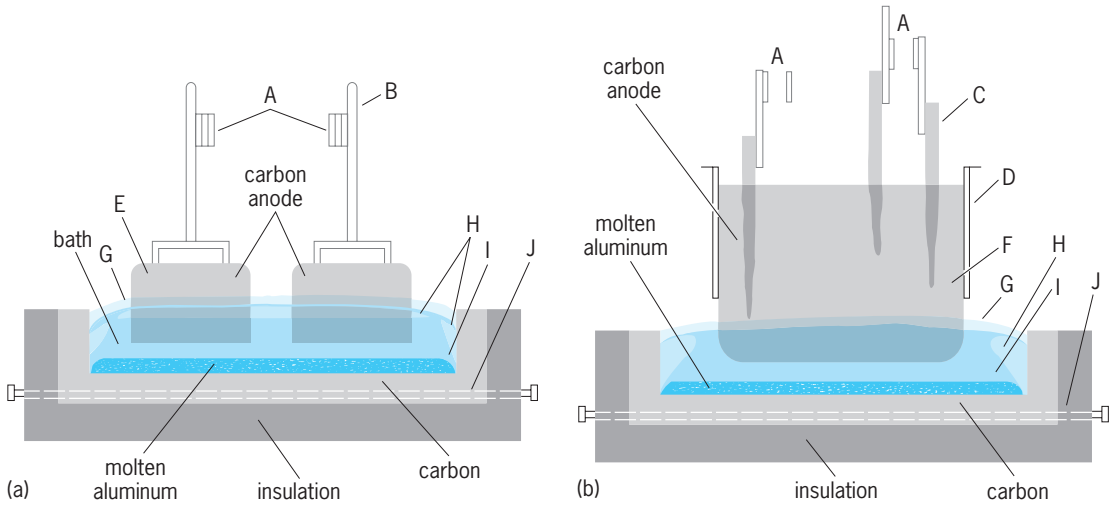


Fig. 9. Two types of aluminum cells, (*a*) utilizing prebaked carbon anode and (*b*) utilizing Söderberg carbon anode. A = anode bus, B = anode rod, C = anode stub, D = anode casing, E = prebaked carbon anode, F = Söderberg carbon anode baked in place, G = crust of frozen bath and alumina, H = frozen bath, I = bath (electrolyte), J = steel pin cathode collector.

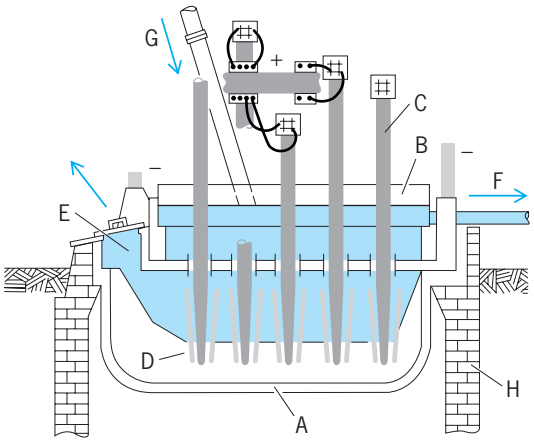


Fig. 10. Cross section of the Dow magnesium cell. A = steel container, B = ceramic cover, C = graphite anodes, D = steel cathodes, E = magnesium collecting well, F = chlorine outlet, G = magnesium chloride feed, H = furnace setting.

Lithium is made by electrolysis of fused 60% LiCl and 40% KCl at 450–500°C (840–930°F) in a cell similar to the Downs sodium cell.

Magnesium is produced by electrolysis of fused 25% $MgCl_2$ and 75% NaCl at around 700°C (1290°F). The Dow process for making magnesium from seawater uses material approximating $MgCl_2 \cdot 2H_2O$, which is fed around the graphite anodes, where dehydration occurs (**Fig. 10**). Gas from the anode compartment is wet chlorine, air, and hydrogen chloride; the latter is used to make fresh magnesium chloride from magnesium hydroxide. Magnesium metal is deposited on steel cathodes, which direct the metal to a collecting zone. The cell is a cast-steel pot in a furnace setting. Other cells uses molten anhydrous magnesium chloride feed. They have brick-lined steel bodies with graphite anodes (**Fig. 11**). Magnesium chloride, with other chlorides, is separated in vacuum crystallizers from brines, dehydrated, and melted in electric resistance cells from which molten $MgCl_2$ is tapped periodically to feed the cells. Molten magnesium is ladled from the cells

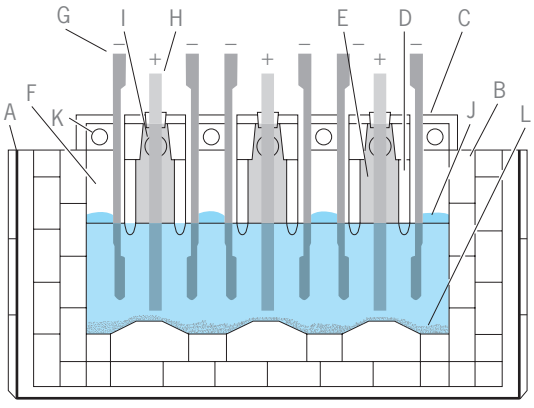


Fig. 11. Cross section of anhydrous magnesium chloride cell. A = steel box, B = ceramic lining, C = ceramic cover, D = ceramic separators, E = anode (chlorine) compartment, F = cathode (metal) compartment, G = iron cathode, H = graphite anode, I = chlorine outlet, J = magnesium, K = cathode compartment vent, L = mud.

and cast into molds. The cells operates at 6–7 V and 80–88% current efficiency and use 8–8.5 kWh/lb (63–67 MJ/kg) metal.

Sodium was once made by electrolysis of fused NaOH, but since 1929 it has been made by electrolysis of NaCl in the Downs cell. The electrolyte is 40% NaCl and 60% $CaCl_2$ at 590°C (1090°F). The cell consists of a brick-lined steel vessels. Four graphite anodes project upward from the bottom. the cathode is made of steel cylinders concentric with the anodes and supported from iron arms extending through the sides of the cell, which also conduct current. A diaphragm of 26-mesh-per-inch (10 mesh-per-centimeter) iron screen directs the sodium into an inverted trough leading to a riser pipe, which cools the metal and conducts it to a collecting tank beside the cell. Chlorine is collected in a nickel cone inverted overthe anode. Pure dry salt is fed to the cell. The reactions in the cell are Eqs. (13) and (14). The

$$2NaCl + 2e \longrightarrow 2Na + Cl_2 \qquad (13)$$

$$2Na + CaCl_2 \rightleftharpoons Ca + 2NaCl \qquad (14)$$

metal at cell temperature is 5% Ca, but as it cools, the second equation is reversed so that the cool metal is about 1% Ca. It is filtered just above the melting point of the sodium, and the final product is under 0.04% Ca. Cells of 38,000 amp operate at 7 V and 83% current efficiency. Energy consumption is about 4 kWh/lb (32 MJ/kg) metal.

Molybdenum, thorium, titanium, uranium, and zinconium can all be made by electrolysis of their complex halides, K_3MoCl_6, $ThF_4 \cdot KF$, K_2TiF_6, KUF_5, and K_2ZrF_6, in molten NaCl. Inert atmospheres are required, and the cell is usually a graphite crucible acting as anode with a graphite or molybdenum cathode, on which the metal deposits as crystals or powder. The batch of metal on the cathode is cooled in an inert atmosphere, then broken off, pulverized, and leached with water. The metal powder is used as such or melted in a vacuum arc furnace.

In the case of tantalum, pure K_2TaF_7 is heated to 900°C (1650°F) in a graphite pot acting as an anode witha removable metal cathode. When the cathode is loaded with deposited metal, it is removed and quickly replaced. The bath is replenished with K_2TaF_7. The cathode deposit is pulverized and washed with acid. The metal powder can be compacted by sintering.

Electrothermics. The manufacture of many products requires temperatures higher than can be obtained by combustion methods. Electric heat can usually be developed at, or close to, the point where it is required, so that it is relatively quick. It permits easy control of the atmosphere for oxidizing, reducing, or neutral conditions.

Products of the electric furnace include iron and steel; ferrallous; nonferrous metals and alloys; the exotic metals titanium, zircomium, hafnium, thorium, and uranium; and nonmetallic products such as calcium carbide, calcium cyanamide, sodium cyanide, silicon carbide, boron carbide, graphite,

fused alumina, magnesium oxide, quartz, silica, thoria, zirconia, lime, spinel, kyanite, sodium aluminate, dolomite, boric acid and borides, carbides of zirconium and titanium and related metals, graphite, phosphorus and phosphoric acid, and chlorides of magnesium, boron, zirconium, and titanium. An electric smelting process converts ilmenite into iron and a titanium slag for a pigment manufacture. Zinc metallurgy uses an electric furnace. Steam is generated in electric boilers where economically feasible.

Large amounts of United States steels, stainless steels, and special alloys are made in electric furnaces. **Table 4** gives values for electrode consumption, furnace voltage, secondary amperage (from transformer to electrodes), and current density.

Methods of electric heating utilize resistance, arcs, or induction. Resistance furnaces (**Fig. 12**) may use the substance being heated as a resist or, or auxiliary resistors may be used. Arc furnaces (**Fig. 13**) may have a direct arc between an electrode and the material being heated, for example, steel scrap or the arc may be between two or more electrodes. The arc may be indirect or it may be submerged in the material being melted. Induction furnaces (**Fig. 14**) use the crucible or its charge as the closed secondary circuit of a transformer with low- or high-frequency alternating current. The temperature in the carbon arc is approximately 3100°C (5600°F). Graphite or amorphous carbon are used for electrodes, depending on economics. Metals such as titanium, zirconium, and halfnium may be compacted into electrodes which are consumed by resistance or arc melting or combinations of both in a nonreactive-atmosphere or vacuum furnace. Some types of furnaces are adaptable to several products, but usually the design is developed for an individual product. *See* ELECTRIC FURNACE; ELECTRIC HEATING; PYROMETALLURGY; STEEL MANUFACTURE.

Zone refining of metals for the electronics industry, such as silicon for diodes and transistors, is accomplished by induction melting of the metal in a narrow zone and slow movement of the molten in the metal ingot from one end to the other in an evacuated or inert gas-filled enclosure. Impurities move toward the end of the ingot. The operation is repeated until the desired purity is obtained. *See* ZONE REFINING.

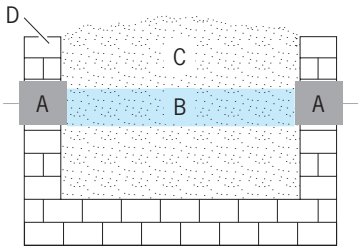


Fig. 12. Resistance furnace. A = current conductors, B = conducting core of granulated carbon, C = granular charge, D = furnace wall of brick and refractory.

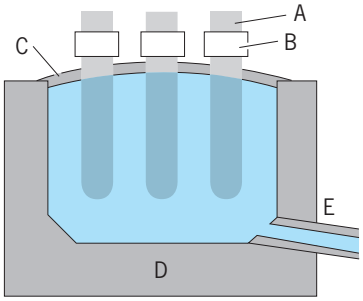


Fig. 13. Three-phase arc furnace. A = carbon or graphite electrodes, B = current clamps, C = refractory roof, D = refractory shell, E = tap hole.

High-melting point metals, carbides, oxides, and nitrides are melted with a plasma-arc torch by coating objects with molten droplets carried in a jet of inert gas passed through the torch. Raw material, such as a powder or wire, is fed through a dc arc between a tungsten cathode and water-cooled copper anode. A strong bond with the base material is obtained. *See* TORCH.

Processes in gases. Electrical discharge through gases has industrial application in ozone production and nitrogen fixation. Ozonizers consist of two metal electrodes with an air gap and a dielectric, such as gas, between them. Very dry air passed through an air gap will then contain 10–12 mg ozone per liter.

TABLE 4. Representative values for electrode consumption, voltage, current, and current density for electric furnace products

Product	Electrode consumption*		Voltage range, phase to phase	Current range, secondary amperes	Current density, A/in.^{2}
	Approx. lb/ton of product	Approx. lb/1000 kWh			
Ferrosilicon, 50%	30–40	7	130–195	35,000–55,000	35–55
Ferrosilicon, 65%	60–90	10	125–170	35,000–50,000	35–60
Ferrosilicon, 75%	90–130	12	120–170	35,000–50,000	35–60
High-carbon ferrochrome	30–50	10	110–200	25,000–40,000	25–40
Silicomanganese	80–100	23	110–140	35,000–70,000	35–70
Ferromanganese	30–50	16	85–100	30,000–60,000	35–65
Calcium carbide	40–60	17	120–200	30,000–80,000	40–70
Phosphorus	35–70	3	250–400	13,000–30,000	20–30
Refined copper	4–6	19	80–190	12,000–30,000	80–110
Nonferrous castings	4–5	15	100–110	800–5000	110–300

*1 lb/ton = 0.5 kg/metric ton. 1 lb/k Wh = 0.1260 kg/MJ.

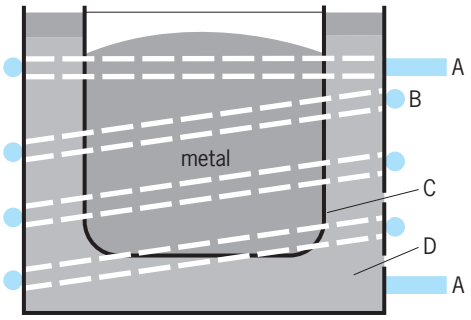


Fig. 14. Induction furnace. A = terminals of high frequency, B = copper coil, C = crucible, D = refractory.

Fixation of nitrogen by passing air through an arc furnace, thus forming oxides of nitrogen, was once practiced when power was cheap in Norway, France, and Italy, but has been replaced by conventional processes. *See* ELECTRICAL CONDUCTION IN GASES; NITROGEN FIXATION; OZONE.

Electromagnetic separation. Magnetic separation removes tramp iron from mixtures of granular solids, suspensions, and solutions. Magnetic separation is also used to separate solids of various magnetic susceptibilities, such as mineral fractions. *See* MECHANICAL SEPARATION TECHNIQUES.

Electrodialysis. This is the separation of low-molecular-weight electrolytes from aqueous solutions by migration of the electrolyte through semipermeable membranes in an electric field. It is used on an industrial scale for deashing starch hydrolyzates and whey, and in many municipalities for producing potable water from saline water. Its uses also include the concentration of liquid foods such as dairy products and citrus juices, the recovery of sulfite pulp waste and pickling acid, and the isolation of proteins. *See* COLLOID; DIALYSIS.

Electrophoretic deposition. This is the deposition of a nonconductive material in a finely divided state from a suspension in an inert medium. Electrophoresis is the migration of colloidal particles, which acquire positive or negative charges in an electric field. The process is useful in electropainting; for instance, electropainting of automobile bodies and other objects has now been adopted on a large scale. Rubber latex is an example of negatively charged colloid which can be plated on an anode. Electronic components can be coated with inorganic salts, oxides, and ceramics suspended in organic media. Bitumen can be electrodeposited out of an aqueous dispersion onto the anodic surface of steel pipe by use of an axially placed cathode. *See* ELECTROPHORESIS.

Electroendosmosis. This is the movement of a liquid with respect to an immobilized colloid in an electric field. The process is used in the dehydration of peat, dye pastes, and clay. Dies in a clay extrusion press can be "lubricated" by making the die cathodic which attracts a film of water to it from the wet clay. It is also used commercially for dewatering soils in mining, road building construction, and other civil engineering works.

Electrostatic technique. The deposition of charged particles from suspension in gases has many useful applications. The Cottrell electrostatic precipitator removes dusts and mists from gases. In one form, a fine wire is axial in a pipe and insulated from it. The pipe is grounded, and the wire is negative in a high-voltage dc circuit. Particles in the gas that is passing through the pipe become electrically charged and move to the pipe. Liquid particles drain off, while solids are periodically vibrated off. *See* ELECTROSTATIC PRECIPITATOR.

Spray painting with a high voltage between the gun and the work is particularly effective in providing an even coating with an economical use of paint or irregular and open surfaces, such as a screen. Complete coverage is achieved because an uncovered spot or pore has a higher field strength and attracts the charged droplets to it.

In xerography a sheet of plain paper is electrically sensitized in those areas corresponding to an original so that colored resin particles carrying an opposite charge are attracted and retained only on the sensitized areas, thus producing a visible image corresponding to the original.

Abrasive paper and cloth are coated with an adhesive and abrasive powers attracted to the base material in an electrostatic field. Pile fabrics can be produced in a similar manner, with the short fibers oriented by the electric field. Powdered adhesive can be applied to paper.

Plant design for electro process. The greatest changes have taken place in the electrical distribution system portion of electrolytic plants. Copper bus-bars have taken for granted for years. They show good corrosion resistance, low contact voltages, and contacts can be kept clean readily. Aluminum is not as corrosion-resistant, and contacts are poor unless specifically designed. Aluminum and copper are competitive conductors when the ratio of the cost of a given amount of aluminum to that of copper is about 1.7. Aluminum is below the equality price, copper is above. Therefore, electrolytic plants cannot afford the luxury of copper bus-bars, but still want the advantages of copper contacts—namely, the copper-surface aluminum bus-bar and the copper-shoe aluminum crossbar, achieved by plating, machined sections, explosive forming, and duplex preparation.

Electrolytic engineering encompasses cell design; electrical isolation; use of sumps instead of sewers; economic selection of bus-bars; solution circuit isolation; under-floor distribution of lines, supports, conductors, crystallizing liquors, and plastic pipe and conduits; lined and unlined cells; contacts and their inspection; and avoidance of pollution.

Early in plant development, the problems of impurities in solutions and the effects of traces of the electrolytic deposit were evident, appreciated, and solved, largely by Edisonian methods. Maximum tolerance were served up, and purification procedures were developed to a high level in copper and zinc. Crystallizing solutions were a problem for manganese. In this area, control of deposition potential,

voltage-amperage and amperage-time recording instruments gave the concept of impurity concentrations affecting the deposition potential and the current density needed to reach this potential. There was an explanation for receding or resolution of deposits. In chromium the problem was solved, in effect, by converting the ore during purification into a chemical compound purified by crystallization. Purification of electrowinning solutions of magnesium, sodium, potassium, barium, calcium, iron, aluminum, antimony, arsenic, tin, cadmium, silica, lead nickel, and cobalt by chemical methods is a common occurrence.

The reconditioning of cathodes for metal deposition and ease of product removal in the electrowinning of manganese and chromium is a repetitive, demanding, and skilled-labor operation. Lines have been designed and put in operation resembling the continuous, conditioning cleaning, pickling, and preparation lines of the electroplating phases of the automotive industry, a long-delayed cross-fertilization. When electrolytic copper sheet needed separate plants for printed electronic circuits, auxiliary electrodes were developed to control edges and ensure more uniform thickness. Perhaps manufacture of starting sheets can be mechanized by adopting continuous electroplating lines.

The concept of electrolytic cell as an insulated mechanism, supported on structures and insulated from them, with separate support for conductors, supply lines, circulating systems, all insulated and electrically broken so that they are not shunt circuits, has been well established for decades. Tanks were wood, but are now unreinforced concrete, lined or unlined, and in few cases glass fiber or cloth-reinforced plastics, such as poly(vinyl chloride) or its competitors. In general the copper industries studies have shown that the investment cost for a lead-lined concrete cell is a little more than half that of a ribbed, equivalent-strength plastic. Because of local availability, tin refining cells are poly(vinyl chloride)-lined concrete units, but chromium metal is made in plastic cells. A few foreign plants employ "painted" or thin-lined concrete.

Selenium rectifiers are used, to a greater extent than copper oxide, for small power supplies, control and signaling circuits, and for applications in which their relative immunity to voltage surges is desired.

There have been no mercury-arc rectifiers built in the United States since about 1960, and very few in Europe for electrolytic or industrial applications, except for high-voltage direct current, welding, and transmission. Monocrystalline diode rectifiers, first germanium and then silicon, have completely replaced mercury-arc rectifiers in industrial applications. Manufacturers have converted to silicon in many multianode mercury-arc rectifiers and excitron and ignition units. All new installations and expansions utilize silicon rectifiers.

The only type of power rectifiers for electrolytic loads, such as aluminum reduction, metal refining, electrowinning, chlorine and chlorate cells, and so on, are silicon rectifiers. Silicon rectifiers range in sizes up to 90,000 A and up to nearly 1000 V, with a maximum rating of 30,000 kVA.

Organic Processes

Organic electrochemistry was once regarded as a tantalizing area with many important laboratory achievements but few successes in commercial practice. This situation has changed, however, in that electroorganic processes are commercially advantageous if they can fulfill either of two conditions: (1) performance under conditions of voltage corresponding thermodynamically to the conversion of an organic group to a reduced or oxidized group, with the cell products relatively easy to isolate and purify; (2) performance of a highly selective, specific technique to make an addition at a double bond, or to split a particular bond (for example, between carbon atoms 17 and 18 of a complex molecule having 25 carbon atoms).

Selectivity and specificity are highly important in electroorganic processes for the manufacture of complicated molecules of vitamins and hormones—as well as for the medicinal products whose action on pathogenic organisms is a function of their spatial arrangement, steric forms, and resonance.

The electrolytic approach can also be competitive for some low-cost, tonnage products. Here continuous processing is important, and only a single phase should be present, that is, a solution rather than an emulsion, dispersion, or mechanical mixture. Only for fairly valuable products is it practical to find a conducting solvent and then to engineer around it.

The electrolytic oxidation and reduction of organic compounds differ from the corresponding and more familiar inorganic reactions only in that organic reactions tend to be complex and have low yields. The electrochemical principles are precisely those of inorganic reactions, while the procedures for handling the chemicals are precisely those of organic chemistry.

Most organic molecules are insoluble in the aqueous solutions that are the best electrical conductors. Unless solubility can be increased, the only other approach is to use organic solvents. These make relatively poor conductors; hence are encountered power loss, heat build-up, chemical inversion, and often, stepwise, complex reactions.

Oxidations. Commercial success in inorganic electrochemistry has come about by well-engineered combinations of organic and inorganic techniques in areas where strictly chemical methods are either impossible or inefficient, for example, in catalytic hydrogenation or oxidation.

The conventional oxidation reagents of the organic chemist are expensive. There is no market for the oxidant once it is reduced, and chemical regeneration is prohibitive in cost. Accordingly, these reagents are avoided. Electrolytic regeneration, however, can be relatively inexpensive if linked to a carefully controlled organic operation. Typical of this approach is manganese dioxide oxidation of anthraquinone with electrolytic regeneration of the oxidant. In the electrolytic oxidation of anthracene

there is a cost-efficient process that utilizes a 20% sulfuric acid suspension with a small quantity of ceric sulfate as a catalyst. Other examples include chromic acid oxidation of oleic acid to perlargonic and azelaic acids, where the oxidant is regenerated electrolytically, and the electrolytic regeneration of periodic acid (a costly oxidizing reagent) in the dialdehyde starch process. These involve savings not only in the purchase price of the oxidants but also in the disposal cost of products that cannot be marketed.

In electrolytic reactions the cell surface is the only source of reductants or oxidants, and these must be produced at highest efficiencies. The mass action effects, concentrations, temperature of reaction, reaction velocities, diffusion, and equilibria between the initial and final products of the reaction all apply to electrochemical reduction and oxidation in the same manner as do the corresponding reactions carried on outside the electrolytic cell.

Materials that are reduced absorb hydrogen at the cathode, and may be considered cathodic depolarizers; those that are oxidized absorb oxygen at the anode, and are anodic depolarizers. Oxidation reactions may involve substances other than oxygen, such as chlorine.

Anodes are selected with a high oxygen or halogen overvoltage, and cathodes are selected with a high hydrogen overvoltage. (Because the accumulation of electrolysis products at anode and cathode causes polarization, the overvoltages are needed to move the products away and keep the process going.)

Reductions. Substances that are easy to reduce may be acted on, at the interface of cathodes, with low hydrogen overvoltage. Hard-to-reduce materials may require much higher overvoltages, which are reached through either the cathode composition or the current density.

The aromatic nitrogen-containing organic compounds have been extensively studied by many researchers. As early as 1900, F. Haber and K. Elbs showed that the nitro compounds of the type RNO_2, where R is an organic radical, could either be directly reduced to the amine RNH_2 (that is, aniline), or successively reduced to the nitroso product RNO, the beta aryl hydroxylamine RNHOH, the azoxy product RN—O—NR, and the material containing an azo group RN═NR, which in turn is reduced to the hydrazoform RNH═NHR or the idene type as $H_2NR{=}RNH_2$.

Although commercial processes based on these reactions were successful, they were eventually replaced by improved nonelectrolytic processes.

Oxidation of starch. The original process for the oxidation of starch to dialdehyde starch was by periodic acid, continuously regenerated at the anode of an electrochemical cell. Dialdehyde starch is an important polymeric aldehyde used in paper manufacture and leather tanning. In this process a lead anode, coated with lead dioxide, is immersed in a suspension of starch; the cathodes are steel, enclosed in porous aluminum oxide diaphragms. One difficulty with this process is that free iodine, formed by the migration of iodine ions to the anode, tends to react with starch, despite the presence of the diaphragm.

In 1960 an improved two-stage process was devised, where just enough periodic acid is added to oxidize the starch in a straightforward chemical reaction. The iodic acid produced during the reaction is then separated from the dialdehyde starch and sent to an electrolytic cell for reconversion to periodic acid. In the electrolytic cell, this reconversion is achieved with current efficiencies of 80–90%. Materials of construction are structural plastics that are unaffected by the reagents or reactants; only the electrodes are metal. Continuous operation is achieved, with the instrumentation permitting good control of temperature, concentration, conversions per pass, gas venting or collection, external recirculation rates, and other rate factors. The wide range over which such control can be exercised allows the cell to function like a reactor or catalytic converter. In fact, the cell has many of the characteristics of the catalytic reactors, the main difference being that in the cell the variables are related to the type and surface of the electrodes, while in the catalytic reactor they are related to the surface of the catalyst and its support.

All piping is manifolded exterior to the cell (unlike the hydrogen-oxygen cell), and the feeds are valve controlled externally. The electrode spacing is prearranged, and all reaction chambers are in parallel both relative to flow and electrically. Because valve closure can shut off a fraction of the parallel streams through the cell, shut downs because of mechanical or other failures are minimized. The electrolytes are simple solutions, not emulsions or suspensions, and do not contain stabilizers or salting-in components. Gas evolution, separation, and collection are separately handled in anolyte and catholyte streams. These two streams may be blended in any proportion desired for maintenance of pH and control of concentration.

Adiponitrile production. In 1965, an electroorganic process for adiponitrile production was developed and commercialized. Power was used as a reagent in the electrolytic reductive coupling (or hydrodimerization) of acrylonitrile to adiponitrile, a component of nylon-6,6. The adiponitrile molecule is a six-carbon dimer, essentially made up of two molecules of three-carbon acrylonitrile.

In the commercial plant, each cell consists of a lead cathode and an alloy anode separated by a sulfonated polystyrene resin membrane. Spacers support the membrane and prevent it from touching the electrodes. The cells are made of polypropylene with neoprene gaskets and are grouped in banks of 24. The reactants, the catholyte and anolyte, circulate separately in an upflow direction through a bank of cells in parallel; the direct current passes through the bank in series. The catholyte contains a quanternary ammonium salt, that is, a McKee salt. This provides ions for electrical conductivity and increases the acrylonitrile's solubility above 10%; lesser concentrations would lead to the formation of proprionitrile.

The catholyte is circulated to remove the film of product at the cathode surface.

Lead alkyl production. In one electroorganic process for the electrolytic production of lead alkyls, electrochemical synthesis of tetramethyllead or tetraethyllead is achieved by employing a Grignard reagent and lead shot. The electrolyte consists of alkyl magnesium halides, which ionize in ether solutions. The solution is fed to 8000-gallon (30,280-liter) electrolytic cells, with lead pellet anodes, the walls of the cell as cathode, and a nonconducting low-permeability diaphragm. The reaction is shown in reaction (15), where R represents the methyl or ethyl groups.

$$4R^- + Pb - 4e \longrightarrow R_4Pb \qquad (15)$$

The $MgCl^+$ ions migrate to the cathode cell walls. Magnesium chloride and metallic magnesium are formed according to reaction (16).

$$4MgCl^+ + 4e \longrightarrow 2Mg + 2MgCl_2 \qquad (16)$$

In this process a variable-voltage approach is used. As the rate of reaction drops and the resistance of the cell increases, more and more voltage is applied to the cell, thus maintaining a constant rate of reaction per unit of time.

Mixed alkyl lead compounds may be made by electrolyzing ethylmagnesium chloride and adding methyl chloride in the cells; or, conversely, by starting with methylmagnesium chloride and then adding ethyl chloride.

The electrolysis cell bank is made up of 10 cells, each with an associated recirculation surge drum. In the cell the copper bus handles the low-voltage, high-amperage current loads. The lead pellet storage hopper and tubular conveyor equipment are located directly above the cell dome. The upper third of the recirculating drum is an insulated refrigerant-storage vessel that supplies coolant to carry off heat developed during the course of electrolysis. Refrigerant vapors are returned to centrifugal compressors.

Charles L. Mantell

Bibliography. A. Bard and L. R. Faulkner, *Electrochemical Methods, Fundamentals and Applications*, 2d ed., 2000; D. R. Crow, *Principles and Applications of Electrochemistry*, 4th ed., 1994; A. Kuhn (ed.), *Industrial Electrochemical Processes*, 1971; H. Lund and M. M. Baizier (eds.), *Organic Electrochemistry*: *An Introduction and a Guide*, 3d ed., 1991; D. Pletcher and F. C. Walsh, *Industrial Electrochemistry*, 2d ed., 1993; K. Scott, *Electrochemical Reaction Engineering*, 1991.

Electrochemical series

A series in which the metals are listed in the order of their chemical reactivity, the most active at the top and the less reactive or more "noble" metals at the bottom. In a broader sense such an activity series need not be limited to the metals but may be carried on through the electronegative (nonmetallic) elements as well. See the **table** for a list of common elements.

Electrochemical series of the elements*

Element	Symbol	Element	Symbol
Lithium	Li	Zinc	Zn
Potassium	K	Chromium	Cr
Rubidium	Rb	Gallium	Ga
Cesium	Cs	Iron	Fe
Radium	Ra	Cadmium	Cd
Barium	Ba	Indium	In
Strontium	Sr	Thallium	Tl
Calcium	Ca	Cobalt	Co
Sodium	Na	Nickel	Ni
Lanthanum	La	Molybdenum	Mo
Cerium	Ce	Tin	Sn
Magnesium	Mg	Lead	Pb
Scandium	Sc	Germanium	Ge
Plutonium	Pu	Tungsten	W
Thorium	Be	**Hydrogen**	**H**
Beryllium	Th		
Uranium	U	Copper	Cu
Hafnium	Hf	Mercury	Hg
Aluminum	Al	Silver	Ag
Titanium	Ti	Gold	Au
Zirconium	Zr	Rhodium	Rh
Manganese	Mn	Platinum	Pt
Vanadium	V	Palladium	Pd
Niobium	Nb	Bromine	Br
Boron	B	Chlorine	Cl
Silicon	Si	Oxygen	O
Tantalum	Ta	Fluorine	F

*According to standard oxidation potentials E° at 25° C or 77° F.

The electrochemical series as it applies to metals was first established by laboratory experiments in which the purpose was to determine which metals would displace others from solutions of their salts. Thus a clean strip of zinc immersed in a solution of copper sulfate is soon found to be covered by a deposit of copper, while zinc in turn goes into solution from the strip as zinc ions. By definition, then, zinc is a more reactive metal than copper, since it will displace copper from a solution of Cu^{2+} ions. The reaction is readily seen to be an oxidation-reduction transfer of electrons, which can be summarized by the equation below. Similarly, copper will displace

$$Cu^{2+} + Zn = Cu + Zn^{2+}$$

silver from a solution containing Ag^+ ions, depositing crystals of metallic silver and coloring the solution with Cu^{2+} ions. From these observations an activity series may be set up in the order Zn, Cu, Ag. By exhaustive experiments with other metals it becomes possible to draw up a complete list in the order of chemical activity, in which the metals at the top of the list are those which are found to give up their electrons most readily (that is, are the most electropositive elements). Such a list is shown in the table, where lithium exhibits the most reactivity as a metal.

The ease with which an isolated atom of an element gives up an electron, known as the first ionization potential, is a precise physical quantity which can be measured by electrical experiments

on gases or vapors at low pressure. The replacement experiments which determine the order of the electrochemical series take place in a very different environment, since they involve solid phases and also aqueous solutions with their consequent hydration effects. Moreover, it might well be expected that displacement reactions in solution would depend upon the concentrations of the reagents used, and also upon the presence or absence of other dissolved substances. *See* IONIZATION POTENTIAL.

To obtain a more accurate and reproducible activity series, it is best to turn to the more exact quantity called electrode potential, or oxidation-reduction potential, which is defined as the voltage developed by a sample of pure metal immersed in a solution of one of its salts (at unit activity and at 25°C or 77°F) versus a hydrogen electrode immersed in hydrochloric or sulfuric acid of equivalent concentration. For details about this measurement *see* ELECTRODE POTENTIAL

It is evident that by confining the experimental conditions to a standard concentration and temperature the hydration and concentration effects can be kept quite constant, making possible a more exact listing of metals according to their activity. Hence any present-day electrochemical series must rely on the measurements of oxidation potential and should be in agreement with the accepted values determined from such electrochemical cells. Such reliance has the further advantage that the series need not then be confined to metals but may be extended to the nonmetals, or electronegative elements. As before, those metals which will liberate hydrogen from dilute acids (such as hydrochloric or sulfuric) will stand above hydrogen in the series, while those metals and nonmetallic elements which will not liberate hydrogen from such dilute acids will stand below hydrogen in the list. Since the oxidation potentials are also related to the equilibrium constants for reversible reactions, it becomes possible to calculate oxidation potentials from other information when direct experiments are inconvenient, as in the case of the alkali metals versus aqueous solutions of their salts. *See* ELECTROCHEMISTRY; ELECTRONEGATIVITY; OXIDATION-REDUCTION.

Eugene G. Rochow

Bibliography. F. A. Cotton, G. Wilkinson, and P. L. Gaus, *Basic Inorganic Chemistry*, 3d ed., 1994; C. A. Hampel (ed.), *Encyclopedia of Electrochemistry*, 1964, reprint 1972.

Electrochemical techniques

Experimental methods developed to study the physical and chemical phenomena associated with electron transfer at the interface of an electrode and solution. The objective is to obtain either analytical or fundamental information regarding electroactive species in solution. Fundamental electrode characteristics may be investigated also.

The physical and chemical phenomena important in electrode processes generally occur very close to the electrode surface (usually within a few micrometers). Mass transfer of species involved in an electrode process to and from the bulk of solution is one important aspect. Inclusion of a large excess of inert electrolyte in most electrochemical systems eliminates electrical migration (current flow as a result of movement of ions of the substrate toward or away from the electrode) as an important means of mass transfer for electroactive species, and only convection and diffusion are considered.

Important chemical aspects of electrode processes include the oxidation or reduction occurring as a result of electron transfer, and coupled chemical reactions. Coupled reactions are initiated by production or depletion of the primary products or reactants at the electrode surface. Identification of the nature and mechanism of such coupled reactions is of particular importance in studies of electrode reactions of organic compounds, where multireaction and multielectron cascades are often found to be initiated by electron transfer to or between the electrode and the electroactive substance.

The primary experimental variables involved in electrochemical techniques are the potential E, the current I, and the time t. The potential or current at the working electrode is controlled and the other is observed as a function of time. The many ways in which either may be controlled give rise to the wide variety of controlled-potential or controlled-current techniques. In all such techniques it is necessary to specify whether mass transport of the electroactive species to the electrode is by convection or diffusion, since mathematical treatments of these two processes are quite different. Mass transport by convection is more efficient than diffusion by several orders of magnitude but is much more difficult to model mathematically.

The general scheme for electron transfer at an electrode in solution is shown in reaction (1), where O

$$O + ne^- \underset{k_b}{\overset{k_f}{\rightleftharpoons}} R \tag{1}$$

and R are the oxidized and reduced forms of the electroactive species and n is the number of electrons transferred. When k_f and k_b, the rate constants for the forward and back reaction, are very fast and O and R are not involved in preceding or following chemical reactions, the system is called reversible and the Nernst equation (2) holds. In this relation,

$$E = E^{\circ'} + \frac{0.059}{n} \log \frac{C_0}{C_R} \tag{2}$$

E is the electrode potential, $E^{\circ'}$ is the formal standard potential for the redox couple, and C_O and C_R are concentrations at the electrode surface. In the following discussions, only reversible reduction processes will be considered, although oxidations are equally applicable. *See* ELECTROCHEMICAL PROCESS; ELECTRODE; ELECTRODE POTENTIAL; ELECTROLYSIS; OXIDATION-REDUCTION.

Controlled potential. A variety of methods of this type have been developed, depending on whether the electrode potential is held constant or varied

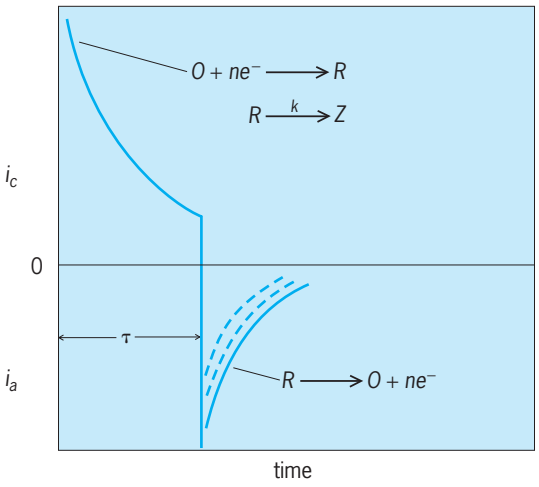


Fig. 1. Current–time behavior for double-potential-step experiment at a stationary working electrode. Broken lines in the reverse current-time curve represent cases where the rate of chemical reaction is fast compared to the time scale of the experiment. $i_a/i_c = f(k)$.

during the experiment and on whether mass transport is by convection (stirring) or diffusion.

Constant potential with convection. In this technique, known as controlled-potential electrolysis, the electrode potential is held constant as the solution is stirred or the electrode is rotated at a constant rate. The current is controlled by the concentration of electroactive substance and the stirring rate. Complete conversion of the electroactive substance takes place at a first-order rate whose constant is determined by cell geometry and stirring efficiency. With good cell geometry and efficient stirring, electrolysis is complete in a manner of minutes.

Constant potential with diffusion (chronoamperometry). When a reducing potential is imposed instantaneously on a stationary working electrode in quiescent solution, current will rise sharply and then decay as the electroactive species near the electrode is depleted by electrolysis. The magnitude of the current is proportional to the bulk concentration of electroactive species, and if the imposed potential is sufficiently negative of $E^{\circ\prime}$, the Nernst equation demands complete conversion to the reduced form. Under these conditions, the current for this chronoamperometric, or potential-step, experiment is diffusion-controlled and decays with $1/t^{1/2}$. This is shown by Eq. (3), where F is the faraday, A is the electrode

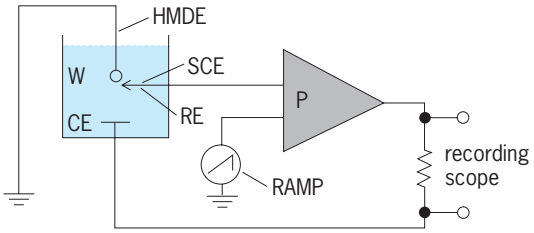
$$I = nFAC^* \left(\frac{D}{\pi t}\right)^{-1/2} \qquad (3)$$

area, D is the diffusion coefficient for the electroactive species, and C^* is the bulk concentration of electroactive species.

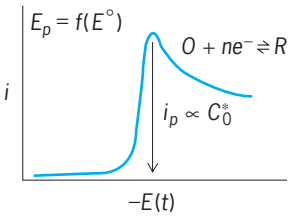
An important variant on this experiment is double-potential-step chronoamperometry, in which a second potential step is applied. During the initial step, electrolysis occurs, depleting the oxidized form O, but producing the reduced form R in the immediate vicinity of the electrode. If the potential is instantaneously switched back to the initial value after a time τ, species R will be reoxidized, and an inverted I-t curve will be obtained (**Fig. 1**). This figure also illustrates the common convention that reduction and oxidation currents are plotted in the positive (upward) and negative (downward) directions, respectively. The value of this technique for studying coupled chemical reactions is illustrated by the broken lines in Fig. 1, which demonstrate that while coupled processes have no effect upon the first step, the current after the second step is diminished by a following chemical reaction (decomposition of R to Z). The ratio of forward and back currents, measured at $t = \tau$ and $t = 2\tau$, respectively, provides a convenient experimental measure of this additional chemical complication. The degree to which the current is diminished by coupled reactions can be used to obtain both rate constants and the molecularity of such reactions. A wide dynamic range of rate constants can be measured by varying τ. *See* ELECTROCHEMISTRY; ELECTROLYSIS.

Variable potential. Several procedures of this type have been developed.

Linear sweep voltammetry (LSV). In this technique, the potential of a stationary electrode in quiescent solution is varied by applying a linear voltage ramp to the electrode. The resulting signal is recorded as a plot of current versus potential. A peak-shaped voltammogram is observed (**Fig. 2**). The decay of current after the peak arises, as in the experiment at fixed potential, because of control of the current by diffusion; indeed, the decay after the peak has a $1/t^{1/2}$ dependence. The height of the peak is dependent upon the concentration of the electroactive species, and the peak potential E_p is related to $E^{\circ\prime}$. Analytically, the experiment is applicable in the 10^{-3} M



(a)



(b)

Fig. 2. Linear sweep voltammetry. (*a*) Generalized instrumentation. (*b*) Current–voltage curve obtained on the scope. HMDE = hanging mercury drop electrode; SCE = saturated calomel reference electrode; CE = auxiliary counter electrode; P = potentiostat; W = working electrode; RE = reference electrode; and RAMP = linearly increasing voltage signal.

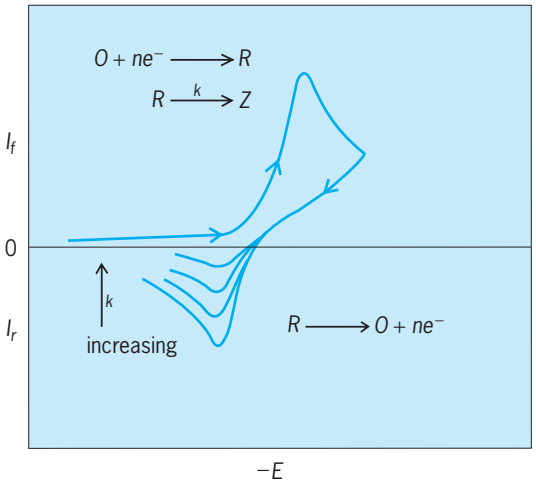


Fig. 3. Current–voltage curves in cyclic voltammetry.

range, but this can be extended to the 10^{-5} *M* range by plotting *dI/dE*, the derivative of the current with respect to potential, as a function of potential (so-called derivative LSV); to a good approximation, the background current rises constantly with potential in the region of E_p and is therefore eliminated by differentiation.

Cyclic voltammetry. The most widely used electrochemical analytical and mechanistic technique is cyclic voltammetry which bears the same relation to linear sweep voltammetry as double-potential-step chronoamperometry does to the single-step experiment. That is, at the end of the first linear ramp, the potential is swept linearly back to the initial potential. Typical cyclic voltammograms are shown in **Fig. 3** for the case of reduction of *O* to *R* and subsequent decomposition of *R* to *Z*. The figure shows a series of cyclic voltammograms corresponding to different values of the rate constant *k*. As in the double-step experiment, the appearance of the forward sweep is unaffected by the coupled chemical process. However, as *k* increases, the height of the peak on the reverse sweep due to reoxidation of *R* decreases. Sweep rates may be varied over a range of roughly 10^{-2} to 10^{2} V/s in conventional cyclic voltammetry. This may be extended considerably, as high as 10^{5} V/s or greater by using microelectrodes (electrodes with diameters in the micrometer range) and appropriate cell shielding to avoid stray currents.

Voltammetry in stirred solution. The potential in a cell configured for controlled potential electrolysis is varied linearly with time (about 5 mV/s), and the current is measured as a function of potential. The current increases sharply when the potential passes through the region of $E^{\circ\prime}$ for an electroactive species in the solution. Between reduction steps, current plateaus are established, the heights of which are proportional to the concentrations of each electroactive species. A typical voltammogram is shown in **Fig. 4***a*, where I_L is the limiting (plateau) current which is proportional to the bulk concentration of electroactive species C^*. The point designated as $E_{1/2}$ is called the half-wave potential and is approximately equal to $E^{\circ\prime}$. A small-amplitude (1–15 mV, 10–100 Hz) alternating potential (Fig. 4*b*) has been superimposed on the linear voltage ramp applied to the electrode of Fig. 4*a*. The resulting alternating-current response is peak-shaped, and the heights of the peaks are proportional to C^*. This experiment, alternating-current voltammetry, can be used to determine rates of heterogeneous electron transfer to or from the electrode to the electroactive species; in this case, it is generally done in quiescent solution and at a solid or dropping mercury electrode.

Stripping analysis. These techniques can sometimes be combined in useful ways. Analytical sensitivity is enhanced by the method known as stripping analysis. A constant reducing potential is applied to an electrode in a stirred solution containing one or more metal ions for a fixed time (about 1–60 min) sufficient to deposit substantial amounts of the metal on the electrode. After this controlled-potential electrolysis, a linear potential sweep to positive (oxidizing) potentials is applied to the electrode. Because of the preconcentration of the metal onto the electrode, the current is not limited by mass transport, and thus the stripping current is considerably larger than would have been observed in a voltammogram in stirred or quiescent solution. Sensitivity as high as 10^{-9} *M* is achieved. The technique is useful for measuring trace-metal analyses in seawater.

Square-wave polarography. This technique resembles conventional polarography at a dropping mercury electrode, except that a small-amplitude (about 1–25 mV) square wave (frequency about 225 Hz) is imposed on the controlling dc potential. The current measured is an alternating component obtained by measuring the current at times near the end of each half-cycle. This avoids the large capacitive background current associated with instantaneous electrode potential changes and allows greater

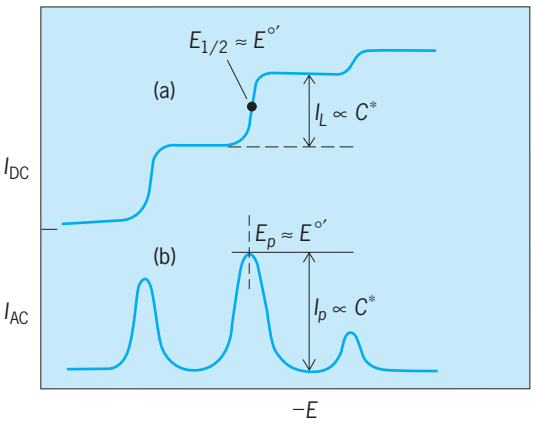


Fig. 4. Plots of the variation of current with constant applied potential in reduction processes. (*a*) Direct current versus voltage at a rotated working electrode (or stirred solution at a stationary electrode), where three successive reductions of electroactive species occur. (*b*) Alternating-current versus direct-current voltage applied at a working electrode in an alternating-current voltammetric experiment under the conditions of Fig. 4*a*.

sensitivity. Analyses over the range of 10^{-3} to 10^{-7} M are possible.

Pulse polarography. This is a variation of square-wave polarography, where a single small potential step is imposed on the controlling dc potential during the life of each mercury drop. The change in current with each step is measured and plotted as a function of the dc potential. It is convenient to make current measurements near the end of the mercury drop life, where the drop is growing slowest and the capacitative current is minimal. For this reason, the drop time is usually controlled mechanically and electronically synchronized with the current measurement. The sensitivity is similar to that of square-wave polarography. Like alternating current and square-wave polarography, a peak-shaped curve resembling Fig. 4*b* is obtained. All of these techniques may use stationary electrodes of carbon or a variety of metals in place of the dropping mercury electrode; in this case, they are referred to as voltammetry rather than polarography. *See* POLAROGRAPHIC ANALYSIS.

Controlled current. Controlled-current techniques, although easier to implement than those in which the potential is controlled, are not as selective and therefore have fallen out of favor in recent years. In chronopotentiometry, a constant current is imposed at the working electrode, and its potential is monitored with time. The electrode must assume a potential which will cause electrolysis sufficient to maintain the imposed current. Thus, the electrode initially adopts the potential of the most easily reduced species. As this species is depleted near the electrode, the potential shifts to that of the next most easily reduced species. This continues until reduction of solvent or electrolyte occurs. Characteristic curves are plotted in **Fig. 5**. The transition time τ is related to the bulk concentration of the electroactive species. However, it is difficult to establish an accurate background correction, thus rendering void the use of chronopotentiometry for analytical purposes.

Coulostatic analysis involves application of a very large, short pulse of current to the electrode, after which the cell circuit is opened and the return of the working electrode potential to its initial value is monitored. The open-circuit condition requires that the current necessary to discharge the electrode interface comes from electrolysis of electroactive species in solution. The current pulse charges up the electrode–solution interface to a new potential, and the electrode discharges, and returns to its original potential, by reducing the electroactive species in solution. The time dependence of the change in electrode potential is proportional to C^*, but use of this dependence for analytical purposes requires the independent and inconvenient measurement of the electrode–solution capacitance. The change in electrode potential ΔE versus $t^{1/2}$ results in a straight-line plot, the slope of which is proportional to concentration, as in Eq. (4), where C is the electrical capacitance associated with the electrode–solution interface.

$$\pm\Delta E = \frac{2nFD^{1/2}C^*}{\pi^{1/2}C}t^{1/2} \tag{4}$$

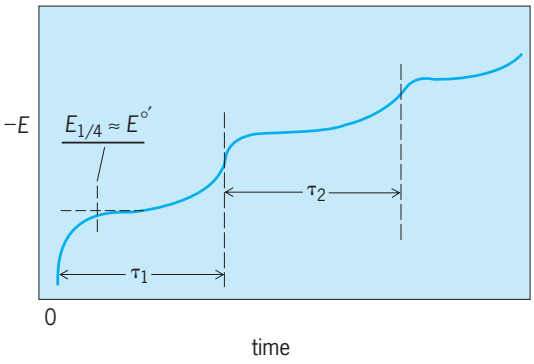


Fig. 5. Potential-time curves for chronopotentiometric experiments. Curves represent successive reduction of three different electroactive species in solution.

Digital simulation. Computer simulation of electrode reactions involving one or more electron transfer steps together with several coupled chemical reactions is a challenging task because concentration gradients of the various chemical species involved vary as a function of both time and distance from the electrode during the experiment. Major advances in the necessary mathematical and computational techniques, as well as increases in computing power even at the desktop level, have led to corresponding major advances in the ease with which electrode reactions can be digitally simulated. Many modern electrochemical investigations, especially in the areas of organic and organometallic chemistry, rely on digital simulation to confirm mechanistic postulates and to extract kinetic and thermodynamic parameters.

Thin-layer electrochemistry. Electrochemical techniques can be applied in cells where a solution only 10–100 mm thick is electrolyzed. The principal advantage is that, for experiments lasting more than about 1 s, mass transfer limitations can be ignored. Thus, correlations between observed parameters and system characteristics are more straightforward. For example, for a chronopotentiometric experiment the transition time represents complete depletion of the electroactive species. Thus, C^* is proportional to τ for each of a sequence of reduction steps. Also, if a linear potential sweep is applied, a symmetrical current peak is obtained, the height or area of which is related to the concentration of electroactive species.

Albert J. Fry

Bibliography. A. J. Bard and L. R. Faulkner, *Electrochemical Methods: Fundamentals and Applications*, 2d ed., 2000; A. J. Fry, *Synthetic Organic Electrochemistry*, 2d ed., 1989; P. T. Kissinger and W. R. Heineman (eds.), *Laboratory Techniques in Electroanalytical Chemistry*, 2d ed., 1996; K. B. Oldham and J. C. Myland, *Fundamentals of Electrochemical Science*, 1994; D. T. Sawyer, A. Sobkowiak, and J. L. Roberts, Jr., *Electrochemistry for Chemists*, 2d ed., 1995.

Electrochemistry

The science dealing with the chemical changes accompanying the passage of an electric current, or the reverse process in which a chemical reaction is used as the source of energy to produce an electric current, as in a battery or fuel cell. Electric conduction occurs through the motion of charged particles. The charged particles may be electrons (as in metals or semiconductors) or ions, which are electrically charged atoms, molecules, or molecular aggregates. Ionic conduction in electrolytes (liquid solutions, molten salts, and certain ionically conductive solids) is a phase of electrochemistry, as is corrosion. Conduction in metals, semiconductors, and gases is generally considered a portion of physics. *See* ELECTROLYTIC CONDUCTANCE.

Galvanic cells. These are better known as electric batteries. Many chemical reactions can be arranged to produce electrical energy by physically separating the reaction into half-reactions: one supplying electrons to an electrode forming the negative terminal of the cell, the other removing the electrons from the positive terminal. In the lead storage battery, electrons are supplied to the negative terminal by a half-reaction in which metallic lead is oxidized to form lead sulfate. At the positive terminal, lead dioxide is reduced to lead sulfate.

The electrons flowing in the external circuit from the negative to the positive terminal constitute the desired electric current. Charging of the lead storage battery by forcing a current to flow in the reverse direction results in the reversal of both half-reactions, and the storage of electric energy in the form of lead and lead dioxide. Such a cell is called a secondary cell, in contrast to a primary cell, such as the Leclanché cell or dry cell, which is not designed to be recharged. In the Leclanché cell, electric energy is produced by the oxidation of zinc and the reduction of manganese dioxide at a carbon electrode. The electrolyte is a moist mixture of zinc chloride, ammonium chloride, and powdered carbon. The fuel cell is designed for the continuous production of electric current through the consumption of oxidant and reductant at separate electrodes. The most common fuel cell is the hydrogen-oxygen (or air) cell with alkaline electrolyte. Many other systems, including hydrocarbon-chlorine and sodium-sulfur, have been proposed. *See* BATTERY; FUEL CELL.

Electrodeposition. The most important type of chemical reaction brought about by the passage of electric current is the deposition of a metal at a cathode from a solution of its ions. Electroplating of many metals, such as silver, cadmium, nickel, and chromium, is used for protective and decorative coatings. Electroforming is a variety of electrodeposition in which the electrically nonconductive prototype of an article to be produced is rendered conductive by spraying with a thin metallic coating, next electroplated with a metallic deposit that is then stripped from its substrate and filled with backing to reproduce the original article. Electrowinning is another application of electrodeposition used for the commercial production from molten salts of active metals such as aluminum, magnesium, and sodium, and from aqueous solution of other metals such as copper, manganese, and antimony. Electrorefining is commonly used to purify metals such as silver, lead, and copper. The impure metal is used as the anode, and purified metal is deposited at the cathode. *See* ELECTROMETALLURGY; ELECTROPLATING OF METALS.

Electrolytic processes. Many electrode reactions other than metal deposition are of commercial or scientific use. Electrolysis of brine to yield chlorine at the anode, hydrogen at the cathode, and sodium hydroxide in the electrolyte is a very important industrial process, consuming a significant fraction of the national electrical production. Many organic compounds can be prepared electrolytically. *See* ELECTROLYSIS.

Electrothermics. While not strictly electrochemical, electrothermics is generally recognized as a part of the field. It includes high-temperature processes involving electric arc or resistance furnaces. *See* ELECTRIC FURNACE.

Electroanalytical chemistry. Many electrochemical measurements are useful for analytical purposes. Electrodes are commonly used for analytical purposes through measurement of their potentials and include the glass electrode for pH measurements, and ion-selective electrodes for certain ions, such as sodium or potassium ion (special glass compositions), calcium ion (liquid membrane), and fluoride ion (doped lanthanum fluoride single crystals). Solid electrodes, especially those composed of platinum and carbon, are generally used to examine current-voltage behavior of substances in solution. Polarography involves the use of a dropping mercury electrode as one electrode of an electrolytic cell. Its use is declining except for analytical applications and those involving operation at highly negative potentials. Qualitative analysis is carried out by measurement of characteristic potentials (half-wave potentials) for electrode processes, and quantitative analysis by measurement of diffusion-controlled currents. Coulometry involves the application of Faraday's law for analytical purposes. *See* COULOMETER; ELECTRODE; ION-SELECTIVE MEMBRANES AND ELECTRODES; POLAROGRAPHIC ANALYSIS.

Several methods involving electrolysis during short periods of electrolysis permit the application of diffusion theory (in the absence of convection) to calculate mass-transport rates and to identify reactive intermediates and reaction mechanisms in multistep electrode processes. These techniques include linear sweep and cyclic voltammetry (measurement of currents with linear voltage scan) and chronopotentiometry (measurement of potential-time transients under constant current conditions). Several titration methods involve electrochemical measurements, for example, conductometric, potentiometric, and amperometric titrations. *See* ELECTROCHEMICAL TECHNIQUES; TITRATION.

Electrode kinetics. Studies of the kinetics of electrode processes are valuable not only for

understanding the mechanisms of electrode reactions but also for studying homogeneous reactions that occur in solutions preceding or following the charge-transfer step. Such studies are made by pulse or transient techniques, or by steady-state methods involving dynamic systems, such as rotating-disc electrodes or flowing solutions.

Miscellaneous phenomena. Electrochemical transport of ions through synthetic or natural membranes is important for processes such as desalination of water and electrodialysis. In biological systems, the transmittal of nerve impulses and the generation of electrical signals such as brain waves are basically of electrochemical origin. A set of related phenomena can be grouped together under electrokinetic behavior, including the motion of colloidal particles in an electric field (electrophoresis), the motion of the liquid phase relative to the stationary solid under the influence of a potential gradient (electroosmosis), and the inverse generation of a potential gradient caused by a flowing liquid (streaming potential). Alternating-current phenomena, such as dielectric behavior, double-layer charging, and faradaic rectification, may also be included in a general definition of electrochemistry. Corrosion and passivation of metals are electrochemical in nature. *See* COLLOID; CORROSION; ELECTROENCEPHALOGRAPHY; ELECTROKINETIC PHENOMENA; ELECTROPHORESIS; STREAMING POTENTIAL.

Albert J. Fry; Herbert A. Laitinen

Organic electrochemistry. Organic electrochemistry involves the study of the chemical reactions that take place when an electric current is passed through a solution containing one or more organic compounds. It is a highly interdisciplinary science because understanding and fully developing a given organic electrochemical reaction may involve not only techniques of synthesis, purification, and identification of organic products but also theory and practice of a variety of sophisticated electroanalytical techniques, surface science, cell design, electronics, engineering scaleup, and materials modification. In recent years, organic electrochemistry has been of increasing interest for industrial applications because the costs of electrical current have been rising more slowly than the costs of conventional chemical reagents and because electrochemical procedures can be environmentally less intrusive than conventional chemical processes. A number of fine chemicals are made electrochemically on scales ranging from several kilograms for expensive specialty compounds to several tons per day. A few chemicals are made on a much larger scale; the best known is adiponitrile [$NC(CH_2)_4CN$], a key intermediate in the synthesis of nylon.

Types of processes. An extremely wide range of organic chemical conversions can be carried out electrochemically. A single electrochemical cell and power supply can generate either anode potentials powerful enough to oxidize saturated alkanes or cathode potentials sufficiently negative to reduce benzene to cyclohexene. If necessary, such processes could be done simultaneously in the same electrochemical cell. An inorganic analogy illustrates this, whereby it is possible to oxidize fluoride ion to elemental fluorine at the anode of an electrochemical cell while reducing sodium ion to metallic sodium at the cathode. Of course, it is necessary to separate the two compartments so that the two products do not immediately react with each other. Almost any functional group can be modified electrochemically. *See* OXIDATION-REDUCTION.

The wide range of redox potentials that can be attained electrochemically and the precision with which this can be done represents only one aspect of organic chemical reactivity. A variety of other processes are possible in electrochemical cells. For example, it is possible to generate a chemical reagent at an electrode under carefully controlled conditions for reaction with another constituent of the medium. Reagents that have been generated in this manner include bromine, chlorine, iodine, iodine ion, hydrogen ion, hydroxide ion, solvated electrons, the powerful oxidant ruthenium oxide (RuO_4), and a variety of low- and high-valent metal ions. For many substrates, electrolysis is done at constant current, in which a fixed current is passed through the solution until the reaction is complete. In this technique, a simple and inexpensive power supply can be used. Selective electrolysis of a complex substance may require controlled-potential electrolysis, which uses a more sophisticated power source (a potentiostat) to maintain the electrode potential at a constant value. *See* ELECTROCHEMICAL PROCESS.

Methodology. The way that an electrochemical reaction is done depends upon the complexity of the process and the amount of material desired. In the laboratory, small amounts of material, from a few milligrams to a hundred grams, can be prepared in relatively simple cells or even in open beakers. Larger quantities are usually prepared in flow cells, often with many of these arranged in series or parallel. Industrial-scale electrochemical reactions are usually done in flow cells to reduce power costs and simplify operating conditions. An emerging technology is the use of microreactors for rapid (<1 s) reaction of electrochemically generated intermediates.

Electrode materials. A wide variety of substances have been used as electrodes, including the metals platinum, mercury, and iron, as well as carbon, which is proving to be the most versatile and least expensive electrode material. For many applications, the material of which the electrode is composed has little effect on the nature of the reaction taking place. The principal requirement is simply that the material be a conductor of electricity. Some applications, particularly those involving adsorbed intermediates and reagents, do take place more efficiently at electrodes made of certain materials. Electrodes made of an intimate mixture of carbon and sulfur, selenium, or tellurium have been used to introduce the latter elements into an organic substrate. Electrolysis at a platinum electrode in acidic media can effect catalytic hydrogenation of unsaturated substances. Other applications depend upon a special coating on the electrode surface to change

the nature of the electrode reactions occurring there. This coating can be a conductive polymer, a chemical redox agent, or an enzyme, where a high degree of selectivity is required. *See* ORGANIC CONDUCTOR.

Reaction mechanisms. The electrochemical behavior of thousands of organic substances has been studied. Consequently, it is possible to make generalizations about the kinds of processes likely to take place with a given substrate. Cathodic organic reactions fall into several categories: cleavage of single bonds, reduction of functional groups, and reduction of large conjugated systems such as activated alkenes and aromatic compounds. Anodic reactions are equally diverse. The oldest and best-known anodic reaction is the Kolbe reaction, in which electrochemical oxidation of carboxylate ions yields dimeric products with the evolution of carbon dioxide. A wide range of functional groups can be accommodated in this versatile reaction. Anodic oxidation of other substrates generally involves initial formation of cation radicals, which may undergo loss of a proton or alkyl group, reaction with solvent or a second component of the medium, or dimerization.

The powerful yet precisely controllable oxidizing and reducing conditions that can be achieved electrochemically are useful for generating novel organic intermediates, including carbocations, carbanions, radicals, radical ions, carbenes, nitrenes, and arynes. These reactive intermediates can frequently be trapped by other organic substances added to the medium to extend the synthetic utility of an electrode reaction. *See* ORGANIC REACTION MECHANISM; REACTIVE INTERMEDIATES.

Voltammetric methods are commonly used to provide detailed mechanistic information. In these, the electrode potential is varied in a controlled fashion in the vicinity of the redox potential of the substrate, and the current response is measured as a function of experimental variables such as scan rate, concentration, and added reagents. Other methods include identification of the products from a preparative electrolysis (often at controlled potential in mechanistic experiments); coulometry, that is, measurement of the actual amount of current passed in the electrolysis; comparison of experimental voltammetric or other responses with computer simulations of the theoretical behavior for various mechanisms; studies of changes in the structure of the electrode surface during electrolysis; and spectroscopic identification of intermediates during electrolysis. Microelectrodes, which have micrometer diameters, extend the speeds with which electrochemical measurements can be made, and thus permit measurement of even very fast rates of chemical reactions associated with electron transfers at electrodes.

Albert J. Fry

Bibliography. A. J. Fry, *Synthetic Organic Electrochemistry*, 2d ed., 1989; H. Lund and O. Hammerich (eds.), *Organic Electrochemistry*, 4th ed., 2000; T. Shono, *Electroorganic Synthesis*, 1991; S. Torii (ed.), *Novel Trends in Electroorganic Synthesis*, 1998.

Electrochromic devices

Self-contained, hermetically sealed, two-electrode electrolytic cells that change their ability to transmit (or reflect) light in response to a small bias (typically 1–2 V) applied across the two electrodes. The operation of electrochromic devices relies upon their electrochromic material content. These materials are organic or inorganic substances that are able to interconvert between two or more color states upon oxidation or reduction, that is, upon electrolytic loss or gain of electrons. The electrochromic materials that are appropriate for most practical applications are strong light absorbers in one redox state but colorless in another. Historically, the phenomenon of electrochromism based on physical phenomena, such as the Franz-Keldish and Stark effects, dates back to 1932. Since the mid-1970s, however, development has been focused on color changes carried out electrochemically, so that in modern usage the terms electrochromic devices and materials refer to devices and substances defined above. *See* STARK EFFECT.

Assembly. A typical electrochromic device is a sandwichlike structure with two glass plates and an electrolyte (**Fig. 1**). Each glass plate is coated on the inside with a transparent electrically conducting layer of indium-tin oxide, which operates as an electrode. Electrochromic mirrors include an additional reflective coating (for example, aluminum) on the outside of one of the glass plates. The electrolyte carries the ionic current inside the cell between the two electrodes, and it can be as simple as a salt (for example, sodium chloride, NaCl) dissolved in a dissociating solvent such as water. However, development has focused on gel and solid electrolytes, because they offer several advantages: they are easier to confine in the space between the electrodes; they function as

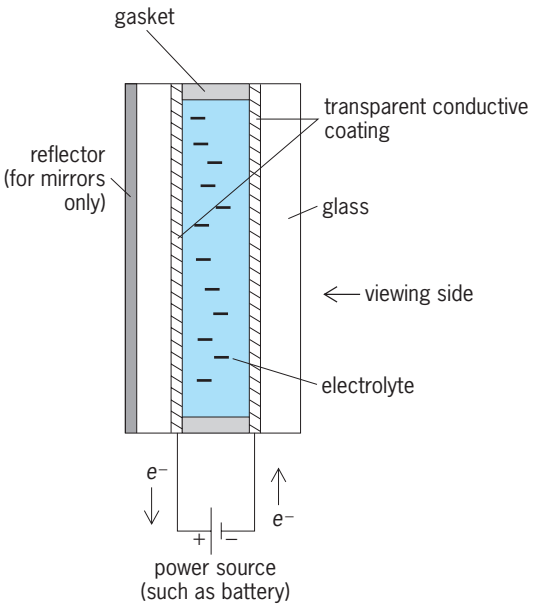


Fig. 1. Electrochromic device. The electrochromic materials can be either dissolved in the electrolyte or coated on the transparent electrodes.

laminators holding the two glass plates together; and their use minimizes the hydrostatic pressure that can cause substrate deformation and leakage problems, particularly in large-area devices such as smart windows.

State-of-the-technology electrochromic devices utilize two electrochromic materials with complementary properties: the first electrochromic material is normally reduced (ECM_1^{red}) and undergoes a colorless-to-colored transition upon oxidation (loss of electrons), while the second electrochromic material is normally oxidized (ECM_2^{ox}) and undergoes a similar transition upon reduction (gain of electrons). The electrochromic materials ECM_1^{red} and ECM_2^{ox} are selected so that they do not react with each other. The oxidation of ECM_1^{red} and the reduction of ECM_2^{ox} then are forced by the external power source (Fig. 1), which operates as an electron pump consuming energy in order to transfer electrons from one electrode to the other. Oxidation of ECM_1^{red} occurs at the positive electrode (anode) and is a source of electrons, while reduction of ECM_2^{ox} occurs at the negative electrode (cathode) and is a sink of electrons. This approach, known as complementary counterelectrode technology, has two distinct advantages. First, the long-term operating stability of the electrochromic cell is greatly enhanced, because providing both a source and a complementary sink of electrons within the same cell prevents any electrolytic decomposition of the electrolyte. Second, the reinforcing effect of two electrochromic materials changing color simultaneously enhances the contrast difference between the color states per unit charge consumed. Depending on the location of the two electrochromic materials within the electrochromic devices, three main types of such devices exist: solution, precipitation, and thin-film (**Fig. 2**). *See* ELECTRODE; ELECTROLYTE; OXIDATION-REDUCTION.

Solution type. In these devices both electrochromic materials, ECM_1^{red} and ECM_2^{ox}, are dissolved in the electrolyte, and they move to the electrodes by diffusion driven by the concentration gradients that are generated by the electrolytic depletion of ECM_1^{red} and ECM_2^{ox} in the vicinity of the electrodes (Fig. 2*a*). The main advantage of the solution-type electrochromic devices is the variety of materials that can be used; any material that can be oxidized or reduced and is electrochromic is a potential candidate. Also, unlike the other types of electrochromic devices, the maximum color intensity can be controlled easily by varying the thickness of the electrolyte layer, typically between 0.1 and 5 mm (0.004 and 0.2 in.). More than 90% of the electrochromic antiglare rearview mirrors for automobiles sold worldwide are based on solution-type devices. A typical device employs *N,N,N′,N′*-tetramethyl-*p*-phenylenediamine (TMPD) as ECM_1^{red}, and *N,N′*-diheptyl-4,4′-bipyridinium dication (HV^{2+}) as ECM_2^{ox}. Electric current forced through the electrode-solution interface is responsible for the electrolysis of the electrochromic materials to their colored forms. *See* ELECTROLYSIS.

In organic solvent-based electrolytes the blue radicals $TMPD^{\cdot+}$ and $HV^{\cdot+}$ are also soluble and diffuse away from the electrodes back into the bulk electrolyte, where they meet, react, and annihilate each other back to their respective colorless states. This property is a desirable feature that renders electrochromic rearview mirrors self-erasing, thereby providing a fail-safe system that automatically recovers the colorless, fully reflective state under power-failure conditions. However, in order for these devices to remain colored, power must be applied continuously. This property increases the energy consumption and makes these devices rather unsuitable for application in architectural glaze. Finally, another potential drawback of this approach is the speed of coloration, which is relatively slow (typically on the order of a few seconds) because it is controlled by diffusion of the electrochromic materials in the bulk electrolyte, a slow process. Implementation of gel or solid electrolytes with these devices slows the diffusion processes even further. *See* DIFFUSION.

Precipitation type. In these devices at least one of the electrochromic materials is originally dissolved in the electrolyte, but upon oxidation or reduction the colored product is electrodeposited onto the corresponding electrode (Fig. 2*b*). One of the first examples (1962) of electrolytic cells in light modulation is a precipitation device in that it employs the reversible electroplating of silver. Later (1973) it was discovered that in aqueous electrolytes the one-electron reduction product, $HV^{\cdot+}$, of the *N,N′*-diheptyl-4,4′-bipyridinium dication, HV^{2+}, forms a blue precipitate on the electrode. That discovery significantly shifted attention from physical to electrochemical electrochromism.

The switching speed of precipitation-type devices is comparable to that of solution types, because both are controlled by diffusion. But precipitation devices offer inherently higher resolution. For example, with patterned-segment electrodes they can be used potentially for high-resolution displays. However, such devices have not been commercialized because precipitation-type electrochromic materials usually suffer poor cycling lifetimes; certain recrystallization effects within the colored precipitates progressively render them electrically inaccessible by the electrode.

Thin-film type. In these devices, both electrochromic materials are immobilized and confined as thin-film coatings (0.1–5 micrometers thick) on the surfaces of the two electrodes (Fig. 2*c*). The main advantage of this arrangement is that physical separation of the two electrochromic materials prevents annihilation of the colored forms, and provides an open-circuit memory that significantly decreases the consumption of power. A typical device consumes less than 10 millicoulombs/cm^2 for full coloration. In a sense, a thin-film-type electrochromic device is a rechargeable battery, in which the color of the electrodes depends upon the state of charge. Discharging (decolorizing) can be carried out either by reversing the bias or by simply short-circuiting the two

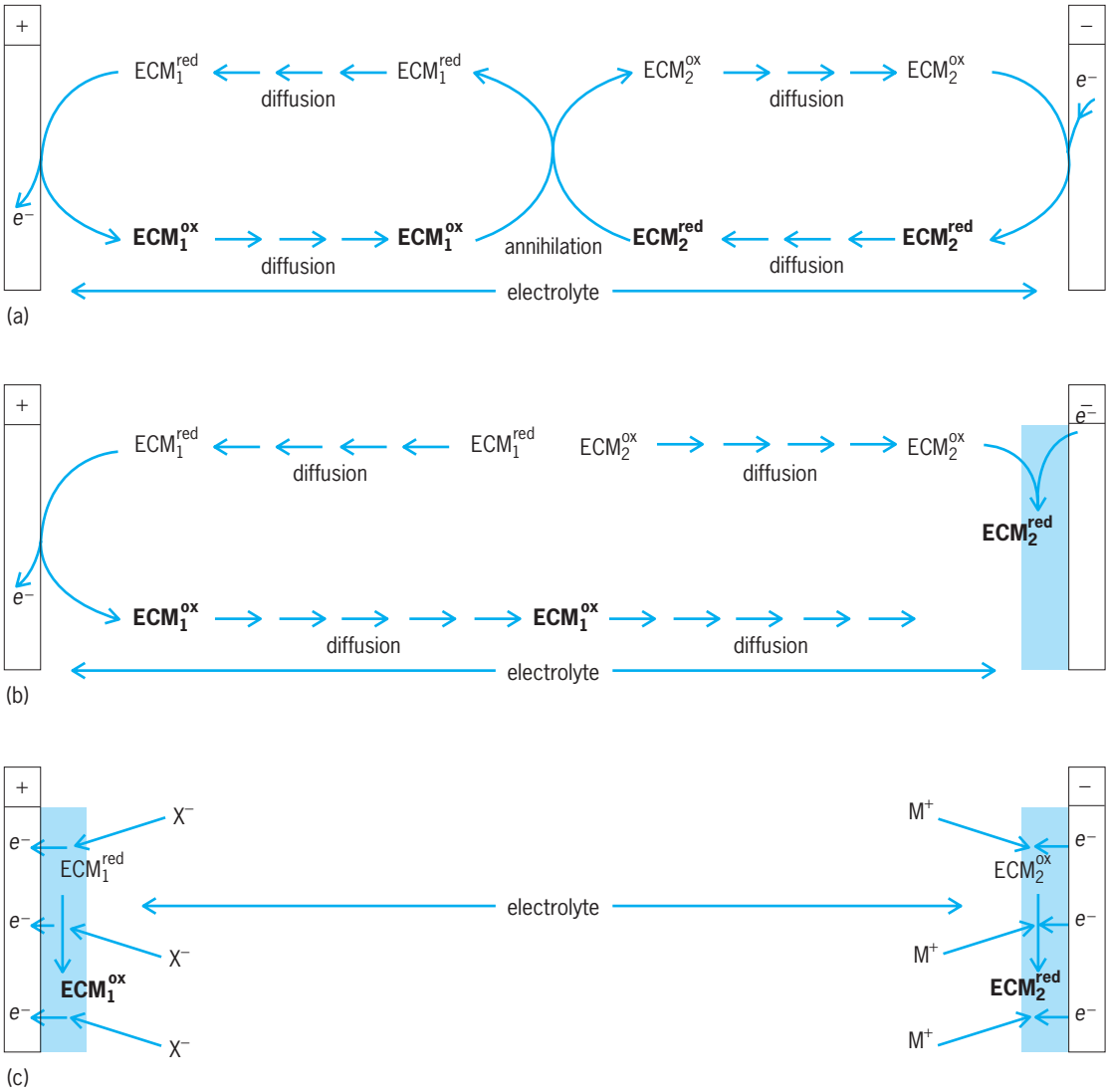


Fig. 2. Operating principles of electrochromic devices. (*a*) Solution type. (*b*) Precipitation type. (*c*) Thin-film type. The color-bearing forms of the electrochromic materials are shown in bold type.

electrodes. Furthermore, the switching speed of this type of device can be greatly improved because coloration does not depend on electrolysis of electrochromic materials dissolved in the bulk electrolyte. In practice, however, coloration depends on diffusion of ions in thin solid films, which is a much slower process than diffusion in the bulk electrolyte. Therefore, depending on the nature of the electrochromic materials, the switching speed of thin-film devices can vary from a fraction of a second to a few tens of seconds.

Several organic and inorganic materials are available for surface confinement on electrodes as thin films. A film of tungsten trioxide (WO_3), for instance, is an inorganic electrochromic material which, upon reduction, uptakes hydrogen ions (H^+) reversibly from an aqueous electrolyte [or lithium ions (Li^+) and sodium ions (Na^+) from a nonaqueous electrolyte] to form a blue tungsten bronze (M_nWO_3, where M is H, Li, or Na). Certain electrochromic rearview mirrors used in trucks employ WO_3 as ECM_2^{ox}, and a nickel hydroxide [$Ni(OH)_2$] film as ECM_1^{red}. These are manual, switch-operated devices which are preset at a level with which the driver feels comfortable. Another inorganic electrochromic material that is complementary to WO_3 and suitable for thin-film electrochromic devices is Everitt's salt, $K_4Fe_4[Fe(CN)_6]_3$, which is the colorless reduced form of Prussian blue, $Fe_4[Fe(CN)_6]_3$.

Among organic materials, conductive polymers such as poly(pyrrole), poly(aniline), and poly(thiophene) have been evaluated as thin-film electrochromic materials for oxidative coloration, and polymeric films derived from diquaternized 4,4′-bipyridine for reductive coloration. Generally, organic electrochromic polymer films are able to switch color in less than a second, and they seem to be inherently more durable than their inorganic counterparts because they better accommodate the structural changes induced upon oxidation and reduction. *See* ORGANIC CONDUCTOR.

Properties. Electrochromic devices are analogous to liquid-crystal devices in that they do not generate their own light but modulate the ambient light.

Liquid-crystal devices require use of polarizers; consequently, their viewing angle is limited, and lateral size limitations are imposed because the spacing between the electrodes (thickness) must be controlled within a few micrometers over the entire device area. Electrochromic devices do not require polarizers, thereby allowing a viewing angle approaching 180°, and contrast ratios similar to black ink on white paper (20:1 or better); moreover, control of the thickness is not important. Other desirable features of electrochromic devices include inherent color, continuous gray scale, and low average power consumption for the thin-film-type devices. Furthermore, it has been shown that electrochromic thin films can be patterned with a 2–5-μm resolution to form a large number of display elements that can be matrix-addressed. Nevertheless, even though there is no apparent intrinsic limitation, the best cycling lifetimes claimed for electrochromic materials are of the order of 10–20 million cycles, while the lifetime of liquid-crystal devices is of the order of several hundred million cycles. This long lifetime has made liquid-crystal devices a very successful technology in matrix-addressed, flat-panel displays.

The larger tolerance in thickness variation for electrochromic devices renders them better suited than liquid-crystal devices for large-area light modulation applications, such as smart windows, space dividers, and smart mirrors. Another possible application is in large-area displays that do not need frequent refreshing, such as signs and announcement boards. Reconfigurable optical recording devices (for example, disks) have been proposed as a high-resolution application that is within the presently available lifetimes of electrochromic materials. *See* ELECTROCHEMICAL PROCESS; ELECTROCHEMISTRY; ELECTRONIC DISPLAY; ELECTROOPTICS; LIQUID CRYSTALS. Nicholas Leventis

Bibliography. A. R. Kmetz and F. K. von Willisen (eds.), *Nonemissive Electrooptic Displays*, 1976; C. M. Lampert and C. G. Granqvist (eds.), *Large-Area Chromogenics: Materials and Devices for Transmittance Control*, vol. IS 4, SPIE Institute Series, 1990; N. Leventis and Y. C. Chung, New complementary electrochromic system based on polypyrrole-prussian blue composite, a benzylviologen polymer, and poly(vinylpyrrolidone)/potassium sulfate aqueous electrolyte, *Chem. Mater.*, 4:1415–1422, 1992; S. Matsumoto (ed.), *Electronic Display Devices*, 1990.

Electroconvulsive therapy

The controlled induction of a grand mal seizure under anesthesia for treating individuals suffering from certain psychiatric illnesses. Electroconvulsive therapy (ECT) was introduced as a treatment for psychiatric disorders by the Italian neurologist Ugo Cerletti in 1938. Cerletti developed a method of electrically inducing seizures in laboratory animals as a means of studying epilepsy. Aware of the observation that individuals with epilepsy and depression experienced an improvement in mood following a seizure, he postulated that inducing a seizure in a depressed individual might improve his or her condition. In 1938 he reported the first case of a psychiatric patient treated with ECT with dramatic, positive results.

Since 1938 ECT has developed considerably. The major changes in the treatment since its inception include the use of general anesthesia to modify the motor component of the seizure; the use of a pulse wave ECT apparatus, which allows careful dosing of electrical energy; the delineation of specific pathological conditions that respond to ECT; and a clearer understanding of how electrode placement and energy dosing affects both therapeutic results and side effects.

Procedure. ECT is performed under general anesthesia. While often performed on hospitalized patients, it is increasingly used in an outpatient setting. The patient is evaluated medically to determine the risk of anesthesia. The required medical workup consists of a physical examination, an electrocardiogram, and laboratory tests, usually including a blood count and determination of electrolytes concentrations. The patient is not allowed to eat or drink after midnight prior to the procedure. Both cardiovascular monitoring and electroencephalographic (EEG) monitoring are done during the procedure. An intravenous line is started, and the patient is anesthetized using a short-acting barbiturate. After the patient is fully asleep, he or she is paralyzed using a short-acting muscle relaxant. When the patient is fully asleep and paralyzed, a grand mal seizure is induced using a pulse wave ECT apparatus. The seizure typically lasts 30 to 60 s. Patients regain consciousness in 5 to 15 min. They are monitored until they are fully recovered and able to eat and drink, which usually takes an additional 30 min.

Treatment dosing. Although all ECT treatment induces a grand mal seizure, it is now known that not all seizures are alike. Both the efficacy of the treatment and the side effects are influenced by stimulus electrode placement and the quantity of energy. Bilateral and right unilateral electrode placements are the most widely used treatment placements and the most researched. Bilateral placement with a dosing slightly above seizure threshold is marginally more effective than right unilateral placement with an energy dosing 4–6 times above seizure threshold. Bilateral placement yields a slightly faster response but also slightly more cognitive side effects. Recently, bifrontal electrode placement with an energy dosing slightly above seizure threshold has been introduced with results comparable to bilateral placement. However, this placement requires more study before it is widely introduced into clinical practice.

ECT-responsive disorders. The use of ECT to treat serious psychiatric disorders has been refined over the past half century. It has shown efficacy in major depression with and without psychotic features; in both the manic and depressed phases of bipolar disorder; in the treatment of acute schizophrenia, particularly when confusion is a prominent symptom;

and in the treatment of catatonia (a psychiatric condition characterized by unresponsiveness to external stimuli) regardless of the psychiatric etiology. Due to the expansion of medication options over the last 20 years, pharmacotherapy is almost always tried before ECT is initiated. The most notable exceptions to this are in circumstances in which a fast response is needed, such as in the treatment of individuals who are dangerously suicidal or severely nutritionally impaired. ECT can show results in days rather than the weeks it typically takes with medication treatment.

The efficacy of ECT in these disorders varies. It appears to be most effective in the treatment of depressed individuals with psychotic symptoms, yielding about a 92% response rate. This compares with an approximately 62% response rate in individuals with major depression without psychotic symptoms. ECT is also extremely effective in the treatment of catatonia, with response rates of about 90%. The efficacy of medication in the treatment of mania and acute schizophrenia is so high that ECT is rarely used in these disorders currently. ECT has been used to treat obsessive-compulsive disorder (OCD) with disappointing results. However, there is a high rate of comorbid (coexisting) depression with OCD, and in this circumstance ECT can be very effective at treating the depression.

ECT has also shown efficacy in the treatment of neuroleptic malignant syndrome (NMS), a confusional state believed to result from an idiosyncratic reaction to antipsychotic medication. ECT has been shown to limit the symptoms and shorten the course of NMS, particularly when it is initiated early.

ECT also has efficacy in treating the depression of Parkinson's disease. In addition, patients with Parkinson's without depressive symptoms have shown improvement in their motor symptoms after receiving a course of ECT. *See* AFFECTIVE DISORDERS; OBSESSIVE-COMPULSIVE DISORDER; PARKINSON'S DISEASE; SCHIZOPHRENIA.

Side effects. The side effects of ECT are divided into two categories: those resulting from the induction of a grand mal seizure and those resulting from the administration of anesthesia.

Grand mal seizure. The induction of a grand mal seizure is accompanied by numerous physiologic changes that can cause side effects. When the electrical stimulus is administered, there is an immediate centrally mediated vagal nerve response, which results in sinus bradycardia (a slow heart rate). This is followed by sinus tachycardia (a fast heart rate) and increase in blood pressure during the seizure. When the seizure ends, there is a peripherally mediated vagal nerve response, which again results in bradycardia. As a consequence, the heart is subjected to rapid changes in rate and blood pressure. This can result in both ventricular and atrial arrhythmias. Typically these arrhythmias consist of several premature atrial contractions or premature ventricular contractions that are self-limited. The potential does exist for the development of significant and potentially life-threatening arrhythmias. For this reason alone, ECT is always done in the presence of an anesthesiologist who is prepared to treat any arrhythmia that might occur. *See* CARDIAC ELECTROPHYSIOLOGY.

The occurrence of a grand mal seizure is associated with short-term memory loss and confusion. The seizure itself causes both anterograde and retrograde amnesia, that is, loss of memory for events occurring both after and before the seizure, respectively. Typically this effect is cumulative so that individuals report memory problems worsening over the course of treatment. These effects are most intense during the treatment course and the several weeks surrounding it. On rare occasion individuals have reported autobiographical memory loss as far back as 6 months prior to the treatment. Evaluating the deleterious cognitive effects from ECT is confounded by the fact that depression causes significant memory problems, as do many of the medications used in its treatment. When assessing the risks and benefits of ECT in regard to memory problems, it is important to weigh the fact that tests show memory functioning—as measured by attention, concentration, and the capacity to acquire and use new information—actually improves after ECT.

The grand mal seizure also increases intracranial pressure. A minor consequence of this is headache. A potentially major consequence is that any central nervous system condition that is at risk for worsening from this increase in pressure needs to be carefully evaluated prior to the individual's undergoing treatment. For example, aneurysms can potentially rupture, or tumors could result in herniations. *See* AMNESIA; MEMORY; SEIZURE DISORDERS.

Anesthesia. Anesthesia-related side effects are similar to those associated with anesthesia used in any minor surgical procedure. The most serious risk is an idiosyncratic reaction to the anesthetic agent, which is associated with a mortality rate of approximately 1 in 100,000. More common side effects are minor headaches, muscle pain, dry mouth, and, rarely, urinary retention. *See* ANESTHESIA.

Mechanism of action. The therapeutic effects of ECT are related to the energy level, the grand mal seizure, and the electrode placement. The treatment produces several effects on neurotransmitters that can be responsible for its efficacy. There is an increase in both neurotransmitter production and release and neurotransmitter receptor sensitivity. There is also an increase in seizure threshold over the course of the treatments that is associated with a clinical response. Energy dosing appears to be important with respect to the electrode placement. Energy dosing at seizure threshold is effective with bilateral electrode placement, but the dosing needs to be four to six times above threshold when right unilateral placements are used. The reason for this is not precisely understood. There is a relative sparing of the speech and language center when right unilateral electrode placement is used, which may explain the reduction in cognitive side effects with this technique.

ECT has been shown to increase long-term potentiation (LPT) pathways in laboratory animals. LPT is

involved in the formation of neural nets related to memory formation. When this effect is blocked in laboratory animals using chemical agents, the memory side effects of ECT are reduced. Work is now under way with human subjects to see if blocking this effect can result in decreased memory side effects. Robert Ostroff

Bibliography. American Psychiatric Association Committee on Electroconvulsive Therapy, *The Practice of Electroconvulsive Therapy*, 2d ed., American Psychiatric Press, Washington, DC, 2001; S. H. Lisanby et al., The effects of electroconvulsive therapy on memory of autobiographical and public events, *Arch. Gen. Psychiat.*, 57:581–590, 2000; K. G. Rasmussen, S. M. Sampson, and T. A. Rummans, Electroconvulsive therapy and newer modalities for the treatment of medication-refractory mental illness, *Mayo Clinic Proc.*, 77:552–556, 2002; H. A. Sackeim et al., A prospective, randomized, double-blind comparison of bilateral and right unilateral electroconvulsive therapy at different stimulus intensities, *Arch. Gen. Psychiat.*, 57:425–435, 2000.

Electrode

An electrical conductor through which an electric current enters or leaves a conducting medium, whether it be an electrolytic solution, solid, molten mass, gas, or vacuum. For electrolytic solutions, many solids, and molten masses, an electrode is an electric conductor at the surface of which a change occurs from conduction by electrons to conduction by ions. For gases and vacuum, the electrodes merely serve to conduct electricity to and from the medium. *See* ELECTRODE POTENTIAL; ELECTROLYSIS; ELECTROMOTIVE FORCE (CELLS). Walter J. Hamer

Electrode potential

The equilibrium potential difference between two conducting phases in contact, most often an electronic conductor such as a metal or semiconductor on the one hand, and an ionic conductor such as an electrolyte solution (a solution containing ions) on the other. Electrode potentials are not experimentally accessible, but the differences in potential between two electronic conductors making contact with the same ionic conductor (that is, the difference between two electrode potentials) can be measured. A useful scale of electrode potentials can therefore be obtained when a particular electrode potential is set equal to zero by definition. There are several conventions, based on different definitions of the zero point on the scale of electrode potential, but all tables use the so-called standard hydrogen convention. *See* ELECTRODE; REFERENCE ELECTRODE.

Mechanism. The interfacial potential difference is usually the consequence of the transfer of some charge carriers from one conducting phase to the other. For example, when a piece of silver, which contains silver ions and free, so-called conduction electrons, is in contact with an aqueous solution of silver nitrate, the only species common to the two phases are the silver ions. Their concentration (volume density) is constant in the metal but variable (from zero to the solubility limit of the silver salt used) in the solution. When more silver ions transfer from the solution to the metal than in the opposite direction, an excess of negatively charged nitrate ions remains in the solution, which therefore acquires a negative charge. However, the metal gains more silver ions than it loses, and therefore acquires a positive charge. Such a charge separation leads to a potential difference across the boundary between the two phases. The continued buildup of such charges makes the potential of the metal more and more positive with respect to that of the solution. This effect in turn leads to electrostatic repulsion of the silver ions in the solution phase immediately adjacent to the metal; these are the very metal ions that are candidates for transfer across the boundary. Consequently, the electrostatic repulsion decreases the tendency of silver ions to move from the solution to the metal, and eventually the process reaches equilibrium, at which point the tendency of ions to transfer is precisely counterbalanced by the repulsion of the candidate ions by the existing potential difference. At that potential, there is no further net transfer of charges between the contacting phases, although individual charges can still exchange across the phase boundary, a process which gives rise to the exchange current.

With a small number of metals, especially silver and mercury, interfacial equilibrium is often established rapidly, well within milliseconds. However, with most metals, the equilibrium state is reached much more slowly, especially when the electrode process involves bond breaking or other complicated reaction mechanisms. In some cases, equilibrium is never attained, and the equilibrium potentials can be known only through calculation.

Concentration dependence. For a metal in contact with its metal ions of valence z, the potential difference E can be expressed in terms of a standard potential E° (describing the affinity of the metal for its ions) and the concentration c of these ions in solution through the Nernst equation (1), where R is the gas

$$E = E^\circ + (RT/zF) \ln c \qquad (1)$$

constant, T is the absolute temperature, and F is the Faraday.

When the metal ions in solution form a sparingly soluble salt, the solubility equilibrium can be used to convert a metal electrode responding to its own metal cations (positive ions) into an electrode responding to the concentration of anions (negative ions). Typical examples are the silver/silver chloride electrode, based on the low solubility of silver chloride (AgCl), and the calomel electrode, based on the poor solubility of calomel (Hg_2Cl_2).

An equilibrium potential difference between a metal and an electrolyte solution can also be established when the latter contains a redox couple, that is, a pair of chemical components that can be converted into each other by the addition or withdrawal of electrons, by reduction or oxidation respectively. In that case the metal often merely acts as the supplier or acceptor of electrons. When metal electrons are donated to the solution, the oxidized form of the redox couple is reduced; when the metal withdraws electrons from the redox couple, its reduced component is oxidized. Again, the buildup of a charge separation generates a potential difference, which counteracts the electrochemical charge transfer and eventually brings the process to equilibrium, a state in which the rate of oxidation is exactly equal to the rate of reduction. The dependence of the equilibrium electrode potential on the concentrations of c_{ox} and c_{red} of the oxidized and reduced forms respectively is described by a Nernst equation of the form (2), where n denotes the number of electrons

$$E = E^\circ + (RT/nF)\ln(c_{ox}/c_{red}) \qquad (2)$$

(e^-) transferred between the oxidized species (Ox) and the reduced species (Red) in the reactions $Ox + ne^- \rightleftarrows Red$. A typical example is the reduction of hydrogen ions H^+ to dissolved hydrogen molecules H_2, and vice versa, in which case the reactions are $2H^+ + 2e^- \rightleftarrows H_2$, and for which the Nernst equation is of the form (3).

$$E = E^\circ + (RT/2F)\ln(c^2_{H^+}/c_{H_2}) \qquad (3)$$

Redox potentials involving a gas are often established slowly, if at all. For determinations of such a redox potential, platinum is often used as the metal, because it is chemically and electrochemically stable. *See* OXIDATION-REDUCTION.

Electrode potentials can also be established at double phase boundaries, such as that between two aqueous solutions separated by a glass membrane. This glass electrode is commonly used for measurements of the pH, a measure of the acidity or basicity of solutions. The mechanism by which the glass electrode operates involves ion exchange of hydrogen ions at the two glass-solution interfaces. *See* ION EXCHANGE.

In all the above examples, the two contacting phases can have only one type of charge carrier in common. Usually, no equilibrium potential difference is established when more than one type of charge carrier can cross the interface, but often (depending on the nature of the metal and of the chemical components of the solution, and sometimes also depending on the geometry of the contact region) an apparently stable potential can still be obtained, which corresponds to zero net charge transfer. This can be a so-called mixed potential, important in metal corrosion, or a junction potential, which figures in most measurements of electrochemical potentials and usually limits the accuracy and precision of such measurements, including that of the pH. *See* CORROSION; ELECTROCHEMICAL SERIES.

Complications. In determining electrode potentials, there are several complications. In the first place, it follows from thermodynamics that the Nernst equation should be written in terms of activities rather than concentrations. The difference between these two parameters is often small but seldom completely negligible. The problem is that the activities of ions (in contrast to those of neutral molecules) are not experimentally accessible, although there are theoretical models that allow their estimation in sufficiently dilute solutions. *See* ACTIVITY (THERMODYNAMICS).

Second, measurements of potential differences always involve the potential difference between two metals rather than that between a metal and a solution. Therefore, electrode potentials as defined above cannot be measured either. The latter complication is circumvented by setting the electrochemical potential of one particular metal-solution interface equal to zero by definition. W. Nernst suggested that a normal hydrogen electrode be used as the common zero point; he defined this as a metal in contact with a solution saturated with hydrogen gas at 1 atm partial pressure and containing 1 mole of a monoprotic acid per liter. However, it was soon found that the nature of the acid used, and the presence of other components in the solution, would affect the potential of such an electrode—a reflection of the fact that the electrode responds to the activity rather than the concentration of the hydrogen ions. The problem here is that the activity of ions (including that of hydrogen ions) is not known, except at or near infinite dilution. Thus the standard hydrogen electrode, defined in terms of hydrogen ion activity rather than concentration, is not experimentally realizable. An alternative definition, in terms of an acid activity, has an anion sensitivity similar to the original definition in terms of concentrations. Moreover, use of the hydrogen electrode can lead to a high liquid junction potential, whereas use of the standard calomel electrode tends to reduce it. In practice, therefore, other electrodes are often used, such as silver–silver chloride electrode or a calomel electrode. In many areas of electrochemistry, the de facto reference electrode is the saturated calomel electrode, which consists of mercury plus calomel in a saturated aqueous solution of potassium chloride (KCl). The standard hydrogen electrode is used mostly for physico-chemical calculations.

Because these measurements involve a potential difference, there has been considerable confusion about the definition of that difference; that is, whether it is defined as the potential of the metal minus that of the solution, or the other way around. This is a matter of a sign convention. The problem is usually framed in terms of oxidation potentials versus reduction potentials. In 1953 the international community settled on the latter and simply called them electrode potentials.

Another zero point is often used to anchor the scale of electrode potentials of semiconductor electrodes. Here the reference point is the work function, that is, the electrical work necessary to remove an

electron from the interior of the phase to a far-away position in vacuum. This is, theoretically, a more meaningful scale. Unfortunately, the work function is not known with sufficient accuracy and precision to make this scale practical. *See* WORK FUNCTION (THERMODYNAMICS).

There are other factors besides charge transfer that can cause interfacial potential differences. Oriented dipolar molecules at interfaces can give rise to interfacial potential differences, even though they represent a charge separation within a layer of only molecular dimensions. Similarly, adsorbed molecules and ions can affect the interfacial potential difference when their adsorption involves a partial charge transfer. None of these phenomena affect the thermodynamic potentials, but they have effectively blocked attempts to define an absolute potential that is both practical and theoretically meaningful.

Applications. There are four main applications for measurements of electrode potentials: (1) in the establishment of the oxidative and reductive power of redox systems, the so-called electromotive series; (2) as concentration probes, such as in pH measurements; (3) as sources of chemical equilibrium data; and (4) as the primary (or independent) variable in studies of electrode reactions.

The electromotive series is a listing of redox couples in the order of their potentials. The most positive potentials correspond to the strongest oxidants, while most negative potentials identify the strongest reductants. This series therefore serves to organize redox systems according to their oxidative or reductive power. *See* ELECTROMOTIVE FORCE (CELLS).

A combination of two electrodes can be employed to determine the concentration of a particular species in a sample solution. One of the two electrodes is the indicator electrode, sensitive to a particular ionic species in the sample solution. The other, reference electrode typically comprises a separate compartment containing a metal in contact with its sparingly soluble salt and an excess of the anion of that salt (as in the calomel and silver/silver chloride electrodes), or with a redox couple such as the combination of iodide and triiodide ions. The solution in the reference electrode compartment then makes contact with the sample solution through a very constrictive contact, such that solution mixing is minimized. Moreover, the nature of the solution in the junction is chosen so that the potential difference across that liquid junction will be as small as possible. This can be achieved by using a high concentration of a salt with, as much as possible, equitransferent ions (ions that have near-equal mobilities, such as in KCl or NH_4NO_3). Under those conditions the liquid junction potential can often be neglected, and the Nernst equation can be written in terms of measurable potentials and concentrations. This is the basis of the measurement of pH using the combination of a glass electrode and a calomel, silver/silver chloride or iodide/triiodide reference electrode. Many other indicator electrodes are available, responding in a similar fashion to different cations and anions. There are also electrodes that respond to neutral compounds. Such electrodes are often based on the presence of an immobilized enzyme. For instance, when an enzyme reacts with its substrate, it may release or consume hydrogen ions. The electrode then responds to the resulting change in the concentration of hydrogen ions and, therefore, indirectly to the enzyme substrate. *See* ION-SELECTIVE MEMBRANES AND ELECTRODES.

Because equilibrium potentials can be correlated with chemical equilibrium constants, measurements of the potentials of electrochemical cells are a major source of numerical information about equilibrium constants. Measurements of electrode potentials not involving liquid junctions can yield data accurate to about 10 microvolts, from which salt activities (but not the corresponding ionic activities) can be calculated. Measurements involving liquid junctions (such as pH measurements using reference electrodes with liquid junctions) are much less reliable, because of the uncertainties in the liquid junction potentials, of the order of one or a few millivolts.

The rates of electrode reactions depend strongly (typically, exponentially) on the electrode potential, and most measurements of the rates of electrode reactions are therefore performed under tight control of this potential, often using a specialized instrument called a potentiostat or, in biophysical applications, a voltage clamp. *See* BIOPOTENTIALS AND IONIC CURRENTS; ELECTROCHEMISTRY. Robert de Levie

Bibliography. M. S. Antelman, *The Encyclopedia of Chemical Electrode Potentials*, 1982; A. J. Bard and H. Lund (eds.), *Encyclopedia of the Electrochemistry of the Elements and Compounds*, vols. 1–15, 1973–1984; A. J. Bard, R. Parsons, and J. Jordan (eds.), *Standard Potentials in Aqueous Solutions*, 1985.

Electrodermal response

A transient change in certain electrical properties of the skin, associated with the sweat gland activity and elicited by any stimulus that evokes an arousal or orienting response. Originally termed the psychogalvanic reflex, this phenomenon became known as the galvanic skin response. Electrodermal response (EDR) has replaced galvanic skin response as the collective term.

Mechanism. The skin of a relaxed person has a low electrical conductance (high resistance), and the skin surface is some 40 mV negative with respect to interior tissues. Sweat gland activity changes these electrical properties by increasing skin conductance and by changing the balance of positive and negative ions in the secreted fluid. Human skin is perforated by the ducts of 2–3 million eccrine sweat glands that are distributed most densely on the chest and forehead and, especially, on the palms and soles, where the orifices can be seen (under 10× magnification) as a row of tiny craters lining the dermatoglyphic ridges. On most of the body surface, sweat glands subserve the function of thermoregulation, but on the palms and soles, sweating varies with

psychological arousal. In many persons this "mental sweating" can spread to regions of the head or chest, especially if the person is highly aroused or if the ambient temperature is high.

Significance. Tonic skin conductance varies with psychological arousal, rising sharply when the subject awakens and rising further with activity, mental effort, or especially stress. Phasic skin conductance responses are wavelike increases in skin conductance that begin 1–2 s after stimulus onset and peak within about 5 s. The amplitude of the skin conductance response varies with the subjective impact of the eliciting stimulus, which in turn varies with the intensity of the stimulus, its novelty or unexpectedness for the subject, and its meaning or signal value. Aroused subjects display spontaneous skin conductance responses, generated apparently by mental events or other internal stimuli; their frequency, like the tonic skin conductance level, increases with the level of arousal. Skin potential phenomena are more complicated, and it is doubtful that there is any psychophysiological information present in skin potential level or skin potential responses that cannot be more easily obtained from measurements of skin conductance. Volar sweating, the parent phenomenon, increases the tactile sensitivity, the resistance to abrasion, and the adhesive qualities of these grasping surfaces; thus its function seems to be preparatory to manipulation.

Measurement. A low-impedance voltage source (0.5 V, 500 ohms) is applied to the skin through nonpolarizing electrodes attached typically to the ventral surface of the distal phalange, where sweat glands are most numerous. Skin contact is effected by means of an electrolyte paste containing 0.5% KCl or NaCl. A small (500-ohm) resistor in series with the subject generates a voltage that varies directly with changes in skin conductance. This signal is amplified and recorded on the moving paper chart of a polygraph. *See* LIE DETECTOR.

Applications. Electrodermal responses are measured in studies of emotion and stress, conditioning, habituation, and cognitive processing: that is, when it is desired to assess the differential or changing impact of a series of stimuli. *See* ELECTROENCEPHALOGRAPHY; SYMPATHETIC NERVOUS SYSTEM.

David T. Lykken

Electrodiagnosis

Diagnosis employing equipment that can be used clinically to measure the intrinsic electrical activity of the heart (electrocardiography), the brain (electroencephalography), peripheral nerves (nerve conduction), and skeletal muscles (electromyography). Electrodiagnostic procedures depend upon the difference in electrical potential between the interior and external surfaces of living cells. This potential difference is created by use of cellular metabolic energy to establish and maintain the concentration gradient of ions across the cell membrane. A discharge of the electrical potential (depolarization) by stimulation or injury occurs in a wave proceeding from cell to cell. The ensuing electrical currents are conducted through the tissues to the surface of the body where electrodiagnostic equipment amplifies and records the changes in potential.

In peripheral nerve conduction studies, signals elicited by electrical stimulation of a nerve are measured either over a muscle supplied by the nerve or along the nerve itself. In another electrodiagnostic test, termed electromyography, electrodes are inserted into a muscle to record its electrical activity. A set of diagnostic tests called evoked potentials records the electrical responses of the brain that follow visual or auditory stimuli.

Electrocardiography. The electrocardiogram (ECG) employs electrodes on the extremities and at various points on the chest wall to record voltage changes due to depolarization and repolarization of heart muscle (myocardium). The ECG is most useful for the diagnosis and treatment of disturbances in heart rhythm (arrhythmias) or in the conduction of electrical impulses within the heart. Both inadequate delivery of oxygen to the myocardium (myocardial ischemia) and localized death of myocardium (infarction) produce characteristic changes in the ECG pattern. The ECG also detects hypertrophy, or overgrowth (of the myocardium), and certain abnormalities of blood electrolytes. ECG recordings during a graded exercise test or over a 24-h period of normal activities (Holter monitoring) may detect ischemic changes or arrhythmias that are not present at rest. *See* CARDIAC ELECTROPHYSIOLOGY; INFARCTION.

Electroencephalography. The electroencephalogram (EEG) records waves of electrical activity from the surface of the brain through the placement of electrodes on the scalp in a standardized pattern. Electroencephalography is the most useful procedure for evaluating individuals with epileptic seizures. Because of their greater sensitivity, computerized tomography (CT) scanning and magnetic resonance imaging (MRI) have largely replaced the EEG for the detection and localization of structural brain abnormalities such as tumors or infarcts. The most important use of EEG is for evaluation of brain abnormalities due to metabolic disorders, drug intoxication, and brain infections that can produce changes in the EEG pattern. The total absence of electrical activity (noted by a flat line on the EEG) for 30 min is an important criterion for diagnosing brain death. *See* DEATH; ELECTROENCEPHALOGRAPHY.

Peripheral nerve conduction studies. Nerve conduction velocity is determined by supramaximally stimulating the surface over a nerve and measuring the time it takes for the signal to reach a recording electrode over the innervated muscle (for a motor nerve) or at another spot along the path of a sensory nerve. Nerve damage results in slowed conduction or decreased amplitude of electrical signals along a sensory nerve. Slowed conduction is a common feature of demyelinating familial neuropathies; reduced amplitudes of action potentials along sensory nerves

are observed in diabetic neuropathy. The delivery of rapid, repetitive shocks to a motor nerve results in a swift reduction in the amplitude of the electrical response of the muscle in individuals with myasthenia gravis. *See* MYASTHENIA GRAVIS.

Electromyography. Electromyography (EMG) assesses the function of muscles with measurements of their electrical activity both at rest and during contraction. The electrical activity of muscles is examined by inserting an electrode into them. The amplified signals are displayed on an oscilloscope screen or heard through a loudspeaker. Abnormal EMG patterns can result from damage to the muscles or their nerve supply. Normal skeletal muscles are electrically silent at rest, but damage to or interruption of the nerve supply to a muscle can lead to scattered electrical spikes caused by spontaneous depolarization of single nerve fibers. During their contraction, damaged muscles produce electrical signals that are diminished in amplitude and duration. Electromyography is useful in individuals with amyotrophic lateral sclerosis (Lou Gehrig's disease) and muscular dystrophy. *See* ELECTROMYOGRAPHY.

Evoked potentials. Visual, auditory, or other sensory stimulation produces small transient potentials in the brain. These evoked potentials require hundreds or thousands of rapidly repeated stimuli to generate the evoked potential or response that requires a computer to summate and average the signals. Frequently used for the diagnosis of optic neuritis or multiple sclerosis is a recording over the back of the skull of the visual evoked potential generated by the display of a light signal to one eye at a time. The auditory evoked potential recorded over the scalp in response to high-frequency clicks delivered by an earphone is also useful for the diagnosis of multiple sclerosis. *See* BIOPOTENTIALS AND IONIC CURRENTS.

Simeon Margolis

Bibliography. M. J. Aminoff, *Electrodiagnosis in Clinical Neurology*, 4th ed., Churchill Livingstone, New York, 1999; K. H. Chiappa (ed.), *Evoked Potentials in Clinical Medicine*, 3d ed., Lippincott-Raven, Philadelphia, 1997; A. L. Goldberger, *Clinical Electrocardiography: A Simplified Approach*, 6th ed., Mosby, St. Louis, 1999; A. H. Ropper and R. H. Brown, Electrophysiologic and laboratory aids in the diagnosis of neuromuscular disease, Chap. 45 in *Adams and Victor's Principle of Neurology*, 8th ed., McGraw-Hill, New York, 2005; A. H. Ropper and R. H. Brown, Special techniques for neurologic diagnosis, Chap. 2 in *Adams and Victor's Principle of Neurology*, 8th ed., McGraw-Hill, New York, 2005.

Electrodynamics

The study of the relations between electrical, magnetic, and mechanical phenomena. This includes considerations of the magnetic fields produced by currents, the electromotive forces induced by changing magnetic fields, the forces on currents in magnetic fields, the propagation of electromagnetic waves, and the behavior of charged particles in electric and magnetic fields. Classical electrodynamics deals with fields and charged particles in the manner first systematically described by J. C. Maxwell, whereas quantum electrodynamics applies the principles of quantum mechanics to electrical and magnetic phenomena. Relativistic electrodynamics is concerned with the behavior of charged particles and fields when the velocities of the particles approach that of light. Cosmic electrodynamics is concerned with electromagnetic phenomena occurring on celestial bodies and in space. *See* ELECTROMAGNETISM; ELECTRON MOTION IN VACUUM; MAXWELL'S EQUATIONS; QUANTUM ELECTRODYNAMICS; RADIO-WAVE PROPAGATION; RELATIVISTIC ELECTRODYNAMICS.

John W. Stewart

Electroencephalography

The biomedical technology and science of recording the minute electric currents produced by the brains of human beings and other animals. Electroencephalography (EEG) was discovered by Hans Berger in the late 1920s, and was found to have important clinical significance for the diagnosis of brain disease. The interpretation of EEG records has become a clinical specialty for neurological diagnosis.

Technology and physiology. The recording machine, the electroencephalograph, usually produces a 16-channel ink-written record of brain waves, the electroencephalogram. It is interpreted by an electroencephalographer. The placement of about 20 equally spaced electrodes pasted to the surface of the scalp is in accordance with the standard positions adopted by the International Federation of EEG, and is called the 10/20 system. Electrode positions are carefully measured so that subsequent EEGs from the same person can be compared. About 10 patterns or montages of combinations of electrode pairs are selected for transforming the spatial location from the scalp to the channels which are traced on the EEG pen writer.

The aggregate of synchronized neuronal activity from hundreds of thousands or millions of neurons acting together form the electrical patterns on the surface of the brain. The cellular basis of the EEG depends on the spontaneous fluctuations of postsynaptic membrane potentials between the inside and the outside of the dendritic processes of postsynaptic cells. The brain waves do not derive from the synchronized axonal spike discharges of cerebral cortical nerve cells.

Electrical voltage is transduced from the scalp by differential input amplifiers and amplified about a million times in order to drive the pens for the paper record. The recording usually takes 30–60 min during a relaxed waking state, and also during sleep when possible. Often, activating procedures are used, such as a flickering light stimulator and hyperventilation or overbreathing for about 3 min.

Normal EEG. EEG waves are defined by form and frequency. Various frequencies are given Greek

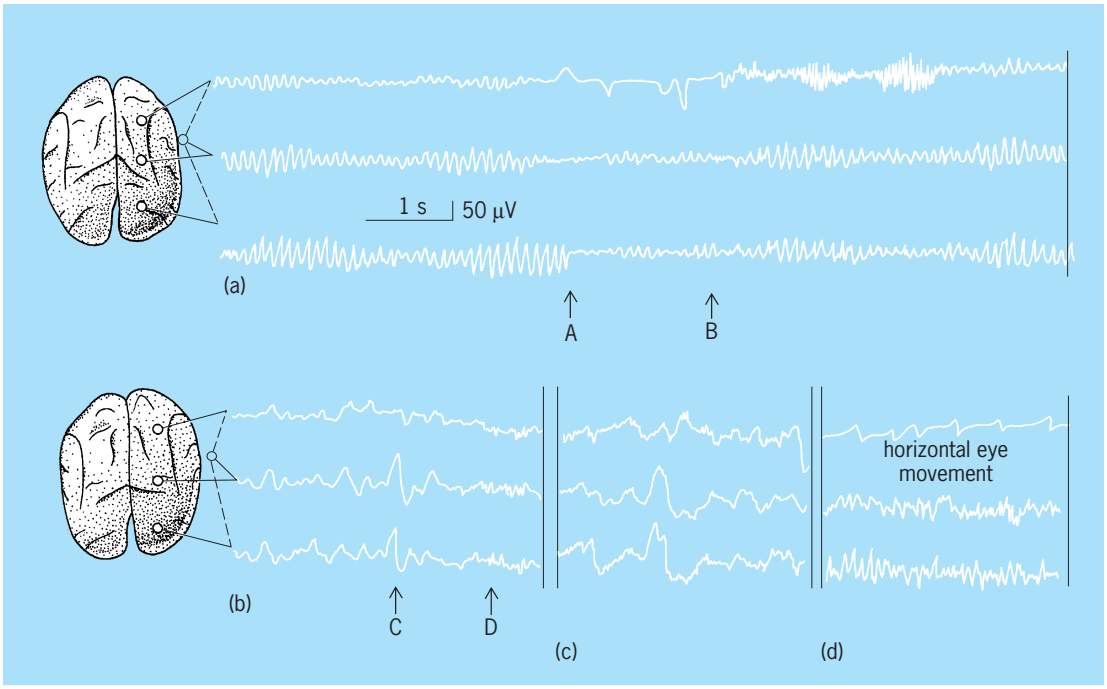


Fig. 1. Normal EEG waves. (*a*) The alpha rhythm in the waking record blocks at eye opening; eyes opened at A and closed at B. (*b*) A light-sleep EEG has theta waves and a sleep hump at C and spindles at D. (*c*) Deep sleep reveals constant, irregular slow delta waves. (*d*) In the dream state, alpha and theta activities are seen, almost like in the waking record.

letter designations. Alpha rhythm is defined as 8–12-Hz sinusoidal rhythmical waves (**Fig. 1***a*). Alpha waves are normally present during the waking and relaxed state and enhanced by closing the eyes. They are suppressed or desynchronized when the eyes are open, or when the individual is emotionally aroused or doing mental work. They may be synchronized by bright light flashes and driven over a wide range of frequencies by repetitive visual stimulation (alpha driving). They are of highest amplitude in the posterior regions of the brain. The alpha rhythm develops with age, reaching maturity by about 12 years, stabilizes, and then declines in frequency and amplitude in old age (over 65).

Beta rhythms are faster, low-voltage sinusoidal waves, usually about 14–30 Hz. They are more prominent in the frontal areas. They are often synchronized and prevalent during sedation with phenobarbital or with the use of tranquilizers and some sedative drugs.

Slower rhythms are theta and delta waves. Theta waves of 4–7 Hz usually replace the alpha rhythm during drowsiness and light sleep (Fig. 1*b*). Delta waves of 0.5–4 Hz are present during deep sleep in normal people of all ages (Fig. 1*c*), and they are the primary waves present in the records of normal infants. Delta waves are almost always pathological in the waking records of adults.

Sleep stages are objectively defined by the use of EEG records. Theta waves are present in drowsiness and light sleep (stage I). Light sleep (stage II) is characterized by negative sharp waves at the vertex (top of head), called sleep humps, and spindle bursts of 13–15-Hz waves in the central regions of the head (over the Rolandic fissures). These are intermixed with theta waves, and seen only during light sleep, sometimes called the hump-and-spindle stage of sleep.

Low-voltage, desynchronized delta waves are present during deep sleep (stage III), and higher-amplitude, slower, and more synchronized delta waves are seen in the deeper sleep stage (stage IV). Stage V sleep is called the dream stage or rapid-eye-movement (REM) stage. That is when dreams take place, accompanied by rapid horizontal eye movements, as if the sleeper were scanning a real scene. The EEG pattern is more like waking, with alpha rhythm present (Fig. 1*d*). The REM state is the one in which it is hardest to wake up the subject, so that it is sometimes called paradoxical sleep. *See* SLEEP AND DREAMING.

Clinical uses of EEG. The EEG reveals functional abnormalities of the brain, whether caused by localized structural lesions, essential paroxysmal states such as epilepsy, or toxic and abnormal metabolic conditions. The three major classes of abnormalities are asymmetries between the hemispheres, slow rhythms, and very sharp waves or spikes. Slow waves represent a depression of cerebral cortical activity or injury in the projection pathways beneath the recording electrodes. Sharp waves or spikes often indicate a hyperexcitable or irritable state of the cortex. During a full epileptic seizure attack, spikes become repetitive and synchronized over the whole surface of the brain. *See* SEIZURE DISORDERS.

The EEG is frequently used for the evaluation of comatose states. The record is slowed in all areas in coma, with delta waves predominating. If the EEG becomes isoelectric or flat for several hours, brain function is not recoverable and the coma may be

considered terminal. "Brain death" is indicated by a flat EEG, recorded at the highest gain with widely spaced electrode positions and the absence of cerebral reflexes and spontaneous respiration. *See* DEATH.

Cerebral evoked potentials. Computer advances in the analysis of EEG signals that are emitted by the brain during sensory stimulation and motor responses have led to the discovery and measurement of electrical waves known as event-related potentials or evoked potentials. These responses are averaged by a computer to enhance the small signals and increase the signal-to-noise ratio, so that they may be graphed and seen.

During the first 10 ms after a click is presented to the ear, electrical responses can be detected from the cochlea, the brainstem, and midbrain structures. Responses to 1000 clicks must be averaged to detect those small, rapid, earliest evoked responses.

This example of a far-field evoked potential is termed the brainstem auditory evoked response (BAER), usually derived from a vertex lead at the top of the scalp compared to a lead at one or the other earlobe as a reference electrode. Each ear is stimulated by clicks at a loudness level set at a certain number of decibels above the auditory threshold. The vertex electrode is far from the origin of the initial electrical activity coming from the cochlear nerve and the brainstem and the midbrain thalamic auditory nuclei, which are far-field or several inches from the active vertex electrode, which is recording field potentials generated by volume conduction rather than by cortical electrical activity from synaptic neural transmission pathways. Neuronal activity from the site of the recording electrode is termed near-field potentials. BAERs are clinically useful in diagnosing peripheral hearing loss due to damage to the structures of the ear, as distinct from hearing disorders involving the neural transmission pathways to the cortex by way of the brainstem and nuclei in the thalamus. Also, evoked potential audiometry has become an active area of clinical research for the evaluation of hearing ability in subjects who cannot respond voluntarily, such as very young children and language-impaired adults.

Somatosensory evoked responses (SERs) are far-field, fast low-voltage brainstem evoked responses derived from stimulating peripheral nerves, usually in the arms or legs, and recorded also by vertex electrodes referred to the earlobes as neutral sites. Like BAERs, the SERs are recorded to trace the somatosensory levels of neural functioning from the spinal cord to the brainstem, thalamus, and cortical areas. Abnormally long latencies or severely reduced amplitudes of the normal waves are diagnostic of neural damage at various levels of conduction of the neural impulses at various synaptic levels.

Fewer signals can be averaged to detect and measure the amplitudes and delays of the larger, slower signals that arrive at the sensory and association cortex 50–150 ms after sensory stimulation. Slower responses have been recorded from association cortex of the brain when psychological factors such as attention, decision, and surprise are indicated. In general, fast potentials relate to the normal physiological functions of direct anatomical pathways, and slower potentials relate to the elaboration of significance and meaning of the responses mediated by nonspecific pathways and association cortex.

Cerebral evoked response patterns have been determined for the major senses under a variety of stimulating conditions. Repetitive electrical stimuli to the skin activate the sensory nerves. Strong stimuli to the eyes such as repeated bright flashes or reversing light and dark squares in a checkerboard pattern, which looks like a checkerboard moving back and forth from left to right and back again, evoke visual responses from the primary and association areas of the visual cortex (VERs). Sinusoidal variations in brightness several times per second are also utilized to evoke visual responses, synchronizing occipital-area brain rhythms, the alpha rhythm, to the stimulating light frequency. The clinical applications of VERs include estimations of visual acuity in nonresponsive subjects, including young children and adults who are not capable of giving verbal or voluntary responses. Visual and other evoked responses are valuable for the diagnosis of early multiple sclerosis, often before any other definite symptoms appear. Optic nerve atrophy or lesions in the visual pathway are detected by these methods.

Evoked responses have also been discovered in the motor systems of the brain preceding voluntary movement by similar averaging methods, using muscle movement outputs as the synchronizing signals to the computer, and tracing voltage changes prior to movements or even speech.

Psychological personality types or psychiatric illnesses may be indicated by modifications of these slow components of the evoked potentials. A negative slow shift, called the contingent negative variation (CNV), is seen when the subject associates a relationship of expectancy between two sensory stimuli when a response is required. If a warning signal is given 1 s before an imperative signal to which the subject must press a button, the contingent negative variation is seen between the two stimuli. The contingent negative variation has been found to be absent or disrupted in some behavioral disorders.

Both normal EEG waves and the fast and slow evoked responses are depressed on the side of a brain injury in one hemisphere compared to the normal hemisphere. Other computer techniques involve the power spectrum analysis of the EEG, such as the Fourier transform, and also cross-correlation between equivalent channels of the two hemispheres.

Evoked potentials are quantified by computer averaging, which enhances the signal-to-noise ratio by the square root of the number of trials. Various evoked potential components are described by the electrical sign, positive or negative, and by their latency in milliseconds to peak amplitude. Factor analytic methods, such as the principal components analysis, are also used by some investigators.

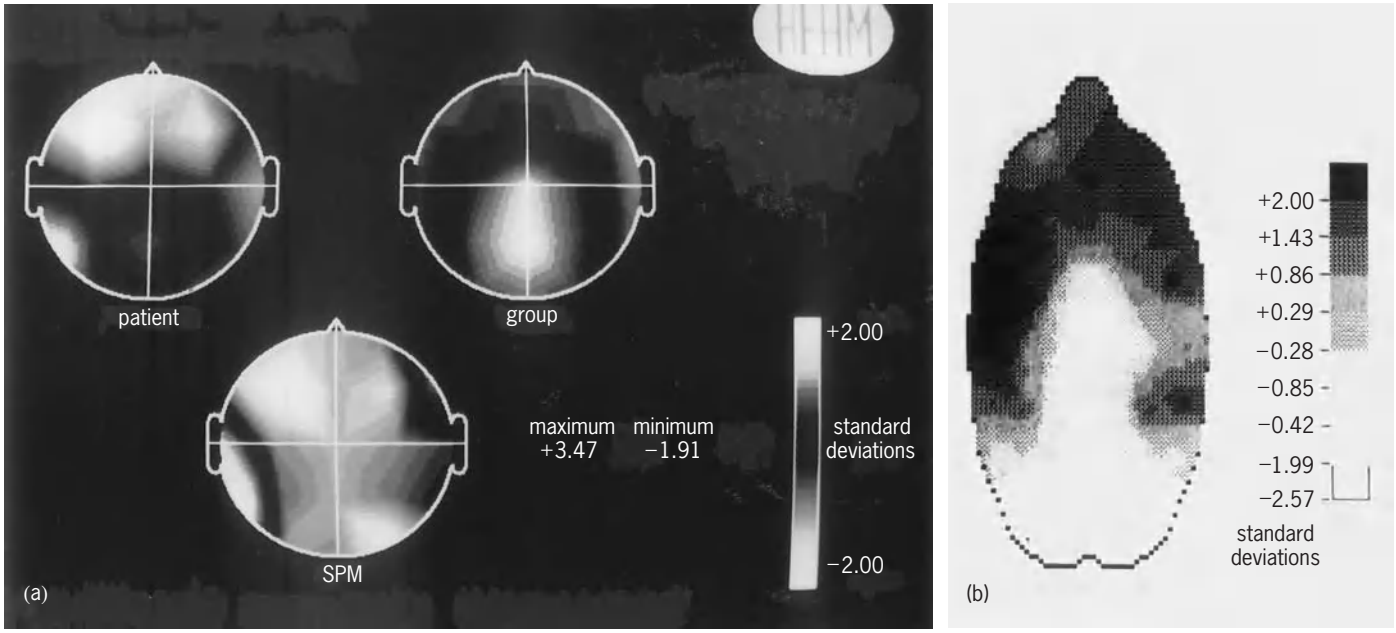


Fig. 2. Two examples of display conventions for topographic mapping. (*a*) Gray-scale technique, depicting amplitude of electrical discharge at particular points in time across the cortex. The maps display comparison of an individual's electroencephalogram with those of a control group, each normalized by subtracting the mean scores and dividing by the standard deviations based on values recorded at each electrode on 3 min of eyes-closed recording. Patient's map shows response to an auditory stimulus at time 156–192 ms after stimulus onset. Group map depicts the electrical response of a comparable control group to the same stimulus at the same point in time. Significance probability map depicts regional differences between patient and control group based upon standard deviation from mean amplitude (*courtesy of F. H. Duffy*). (*b*) Modification of the standard 10–20-electrode placement system, developed by M. Buchsbaum, using 32 electrodes and a four-point linear interpolation. Map shown is a top-down view, with a gray scale depicting brain activity within a specified frequency range.

Topographic mapping. The complexity of evoked potential and EEG analysis makes interpretation difficult in relation to where various components originate and their pattern of spread through time along the neural transmission pathways. Both of these considerations are especially important in determining the way that the brain processes cognitive information as well as sensory signals, that is, how the brain performs at work rather than at rest. In the early 1950s, scientists devised a method of displaying the distribution of EEG wave frequencies over the surface of the head. About 10 oscilloscope displays represented the various head positions and presented the dominant frequency distributions at those spatial locations. It was not until the 1980s, with the development of minicomputers and color graphics screens, that the presentation of topographic information could be analyzed in sophisticated statistical ways for research and clinical purposes by electroencephalographers and neurophysiologists. This method is best known as brain electrical activity mapping (BEAM) and is used in many research investigations of brain activity patterns in learning and language dysfunctions, psychiatric disorders, aging changes and dementia, and studies of normal and impaired child development. Difficult neurological diagnostic problems that do not show anatomical deformities by brain scan methods may often be clarified by these new electrographic proccedures.

To create topographic images of electroencephalography, numeric values are assigned to frequency bands, and these are translated into a gray scale or a rainbow scale. Data are collected from a number of electrodes, usually 20, and an interpolation algorithm is used to develop intermediate values between electrodes (**Fig. 2**).

Topographic images of evoked potentials can be made at millisecond intervals over a period of time and then viewed, like a movie or cartoon, as the computer displays these images on screen at the terminal. The spectral, temporal, and spatial analyses required in electroencephalography and evoked potential interpretation are performed by the computer. When reading the traditional electroencephalogram, the neurophysiologist may take advantage of these imaging techniques to support an interpretation. **Figure 3** provides a simple block diagram of the elements used in generating topographic images of brain electrical activity.

When viewing a topographic map, the clinician is still confronted with the critical question of whether the image (or electroencephalographic recordings associated with the image) represents the neurological activity of a normal, healthy individual or that of one with a pathological condition. A statistical technique called significance probability mapping (SPM) has been developed to aid in this discrimination. Significance probability mapping compares an individual's topographic image to an image based on neurological data of a comparable reference group, and generates a new image that depicts the deviation of the individual from the normal group.

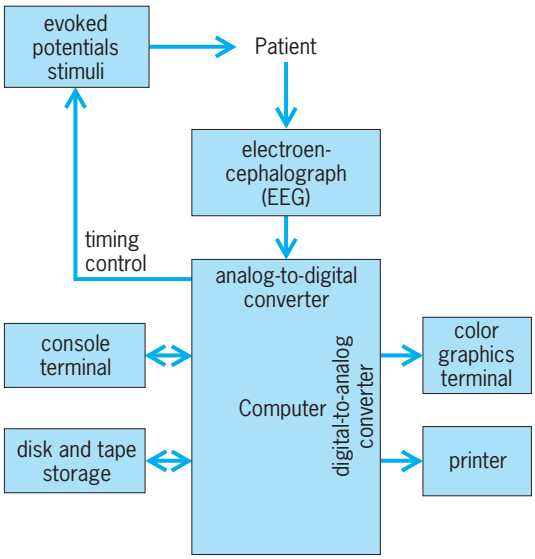


Fig. 3. Primary elements of equipment required for topographic mapping of brain electrical activity.

Such comparisons provide the final complicated step in clinical evaluation of electroencephalographic evoked potential data, and have proven extremely valuable in diagnosing neurologic abnormalities. Additionally, significance probability mapping techniques are valuable in research applications comparing two or more groups of subjects under a number of resting or activated (evoked) conditions. *See* BIOPOTENTIALS AND IONIC CURRENTS; BRAIN; NEUROBIOLOGY.

Jerome Cohen

Bibliography. F. H. Duffy, *Topographic Mapping of Brain Electrical Activity*, 1986; D. Giannitrapani and L. Murri (eds.), *The EEG of Mental Activities*, 1988; R. Johnson, Jr., et al. (eds.), *Current Trends in Event-Related Potential Research*, 1988; E. Niedermeyer, F. H. Lopes da Silva, and L. Fernando (eds.), *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*, 3d ed., 1993.

Electrokinetic phenomena

Phenomena associated with the movement of charged particles through a continuous medium or with the movement of a continuous medium over a charged surface. The four main electrokinetic phenomena are electrophoresis, electroosmosis, streaming potential, and sedimentation potential, or Dorn effect. These phenomena are related to one another through the zeta potential ζ of the electrical double layer that exists in the neighborhood of the charged surface.

Electrically charged layers. The distribution of electrolyte ions in the neighborhood of a negatively charged surface and the variation of potential ψ with distance from the surface are shown in **Fig. 1**. According to O. Stern, two different layers of ions are associated with the charged surface. The layer of ions immediately adjacent to the surface is called the Stern layer. The ions of this layer are held to the charged surface by a combination of electrostatic attraction and specific adsorption forces, such as short-range van der Waals interactions and chemical bonds. The thickness δ of this layer is assumed to be equal to the ionic radius of the adsorbed ion species. The second layer of ions is the Gouy layer. The boundary between the two layers is the limiting Gouy plane. The ions in the Gouy layer are acted upon only by electrostatic forces and thermal motions of the liquid environment (Brownian motion), and they form a diffuse atmosphere of opposite charge (positive charge in Fig. 1) to the net charge at the limiting Gouy plane. The net charge density of the diffuse ion atmosphere of the Gouy layer decreases exponentially with distance from the limiting Gouy plane. The Gouy layer forms the positive half of an electrical double layer, and the charged surface plus the Stern layer form the negative half. The effective distance of separation $1/\kappa$ between the two halves of the double layer is determined by the concentration of electrolyte (ionic strength). For an electrolyte of univalent ions in water at 25°C (77°F), the relationship for $1/\kappa$ from the Debye-Hückel theory is Eq. (1), where c is the concentration of electrolyte (moles/liter).

$$\frac{1}{\kappa} = \frac{3 \times 10^{-8}}{\sqrt{c}} \tag{1}$$

Variation of potential ψ with the distance x from the charged surface is shown by a solid curve in Fig. 1. Here ψ_0 represents the thermodynamic reversible electrode potential which is independent of the properties of the electrical double layer and dependent only on the activity of the ion which is in reversible electrochemical equilibrium with the substance of the charged surface. The potential ψ

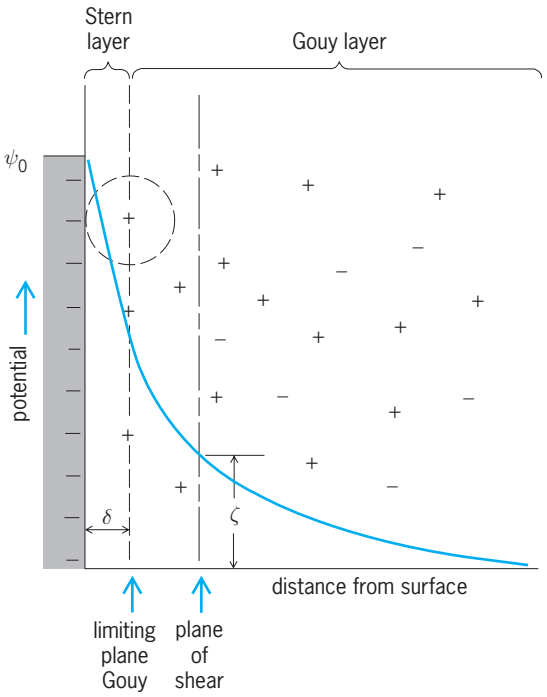


Fig. 1. Electrical double layer.

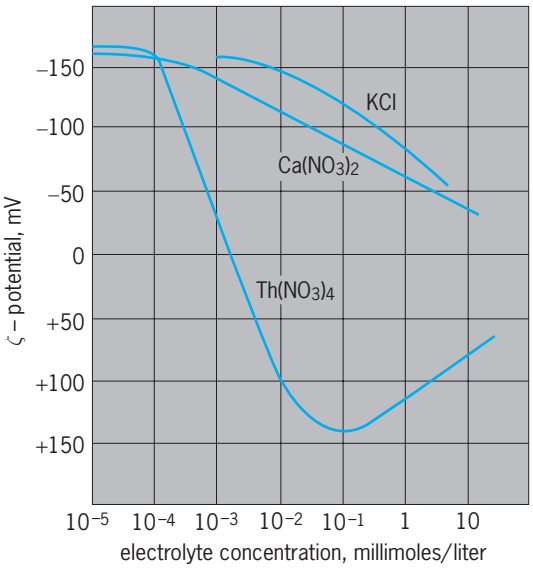


Fig. 2. Effect of electrolyte concentration on the ζ-potential of glass-water interfaces.

decreases linearly with increasing distance x in the region of the Stern layer. In the region of the Gouy layer, ψ decreases exponentially with increasing distance x.

Displacement of charged layers. In the four listed electrokinetic phenomena, a displacement occurs at some plane (plane of shear) between the charged surface and its atmosphere of ions. The position of the slipping plane in Fig. 1 is shown to be located in the Gouy layer. The potential of the plane of shear is the ζ-potential. From the theories of Gouy and Chapman, for spherical particles Eq. (2) holds. Here

$$\zeta = \frac{q}{Da}\,\frac{1}{1+\kappa a} \tag{2}$$

$1/\kappa$ is the effective thickness of the double layer, q the net charge of the particle inside the plane of shear, D the dielectric constant of the liquid, and a the particle's radius at the plane of shear. For flat surfaces

$$\zeta = \frac{4\pi e}{D\kappa} \tag{3}$$

Eq. (3) holds where e is the charge per unit area of surface. Equations (2) and (3) show that ζ-potential is determined by the net charge at the plane of shear and $1/\kappa$, the effective thickness of the ion atmosphere. In turn, ζ-potential controls the rate of transport between the charged surface and the adjacent liquid. The relationship between rate of transport v_E and ζ-potential which is valid for all four electrokinetic phenomena is Eq. (4), where v_E is the velocity

$$v_E = \frac{D\zeta E}{4\pi\eta} \tag{4}$$

of the liquid at a large distance from the charged surface, E is the field strength (V/cm), and η is the viscosity of the liquid. The conditions for validity of Eq. (4) are that the double layer thickness ($1/\kappa$) must be small compared to the radius of curvature of the surface; the substance of the surface must be nonconducting; and the surface conductance of the interface must be negligible. The equations which relate ζ-potential to electroosmotic flow rate and streaming potential may be obtained from Eq. (4) by use of Poiseuille's law for laminar flow through a capillary. For electrophoresis and sedimentation potential (Dorn effect), v_E is the velocity of the particles. E is the applied field strength for electrophoresis, whereas it is the gradient of potential developed by the sedimentation of charged particles in the Dorn effect.

Electrophoresis, electroosmosis, and streaming potential experiments have been shown to yield identical ζ-potentials for several different interfaces, particularly glass-water and protein-water systems. The sedimentation potential has not been significantly studied.

The effect of electrolytes on the ζ-potential of glass-water interfaces is shown in **Figs. 2** and **3**. As shown in Fig. 2, an increase in electrolyte concentration produces a decrease in ζ-potential, and ions of high charge of opposite sign to that of the surface can completely reverse the sign of the ζ-potential. The explanations for these two effects are given in Fig. 3, where the variation in ψ with distance from the surface is shown for low concentration of electrolyte in curve 1; moderate concentration of electrolyte in curve 2; and charge reversal by adsorption of ions (Th^{4+} on glass) in curve 3. Curves 1 and 2 show that an increase in electrolyte concentration reduces ζ-potential by reducing $1/\kappa$, as indicated by Eqs. (1), (2), and (3). Curve 3 shows that reversal of charge by ion adsorption occurs in the Stern layer and that

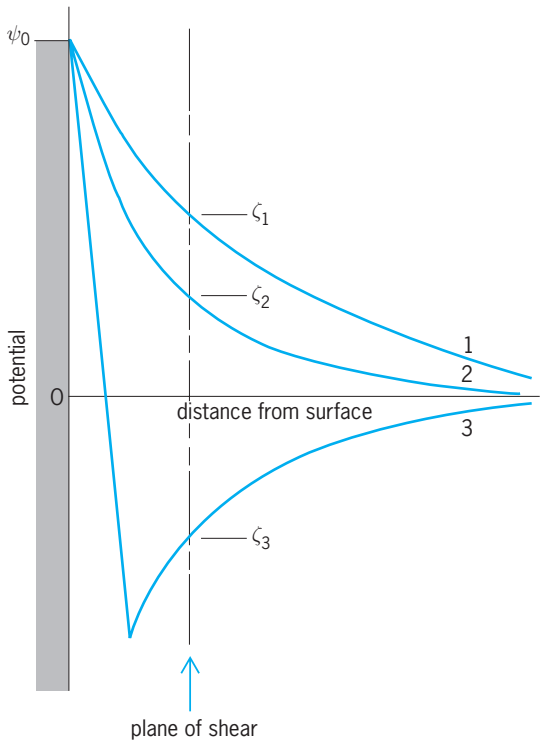


Fig. 3. Reversal of ζ-potential by ion adsorption.

this gives rise to a ζ-potential of opposite sign to the original value. *See* COLLOID; ELECTROPHORESIS; STREAMING POTENTIAL. Quentin Van Winkle

Electroless plating

A chemical reduction process which, once initiated, is autocatalytic. The process is similar to electroplating except that no outside current is needed. The metal ions are reduced by chemical agents in the plating solutions, and deposit on the substrate. Electroless plating is used for coating nonmetallic parts. Decorative electroless plates are usually further coated with electrodeposited nickel and chromium. There are also applications for electroless deposits on metallic substrates, especially when irregularly shaped objects require a uniform coating. Electroless copper is used extensively for printed circuits, which are produced either by coating the nonmetallic substrate with a very thin layer of electroless copper and electroplating to the desired thickness or by using the electroless process only. Electroless iron and cobalt have limited uses. Electroless gold is used for microcircuits and connections to solid-state components. Deeply recessed areas which are difficult to plate can be coated by the electroless process.

Nonmetallic surfaces and some metallic surfaces must be activated before electroless deposition can be initiated. Activation on nonmetals consists of the application of stannous and palladium chloride solutions. Once electroless plating is begun, it will continue to a desired thickness; that is, it is autocatalytic. The process thus differs from a displacement reaction, in which a more noble metal is deposited while a less noble one goes into solution; this ceases when the more noble deposit, if pore-free, covers the less noble substrate.

Electroless copper and gold deposits consist of very small crystals. Electroless nickel deposits are really highly supersaturated alloys containing phosphorus or boron, depending on the reducing agent used. These deposits, which are so fine-grained that they are almost amorphous, can be precipitation-hardened. Electroless nickel deposits are generally harder and more brittle than the electroplated variety. Corrosion resistance of electroless nickel depends, among other factors, on the uniformity of distribution of the phosphorus or boron. Adhesion to nonmetallic substrates is achieved primarily by mechanical means—by plating into pores which are created by selectively etching the substrate. An advantage of electroless plating over plating with current is the more uniform thickness of the surface coating. *See* ELECTROPLATING OF METALS. Rolf Weil

Electroluminescence

A general term for the luminescence excited by the application of an electric field to a system, usually in the solid state. Solid-state electroluminescent systems can be made quite thin, leading to applications in thin-panel area light sources and flat screens to replace cathode-ray tubes for electronic display and image formation. *See* LUMINESCENCE.

Destriau effect. Modern interest in electroluminescence dates from the discovery by G. Destriau in France in 1936 that when a zinc sulfide (ZnS) phosphor powder is suspended in an insulator (oil, plastic, or glass ceramic) and an intense alternating electric field is applied with capacitorlike electrodes, visible light is emitted. The phosphor, prepared from zinc sulfide by addition of a small amount of copper impurity, was later shown to contain particles of a copper sulfide (Cu_2S) phase in addition to copper in its normal role as a luminescence activator in the zinc sulfide lattice. The intensification of the applied electric field by the sharp conductive or semiconductive copper sulfide inhomogeneities is believed to underlie the mechanism of Destriau-type electroluminescence. Minority carriers are ejected from these high-field spots into the low- or moderate-field regions of the phosphor, where they recombine to excite the activator centers. This interpretation is supported by microscopic observations showing that the light is emitted from a few small spots in each phosphor grain. The structure of a Destriau-type electroluminescent cell is shown in **Fig. 1**; the light is observed through the transparent indium–tin oxide electrode.

The brightness L of lamps made with Destriau-type phosphors increases rapidly with the applied voltage V according to the equation below, where A and B

$$L = A \exp \frac{-B}{V^{1/2}}$$

are constants. The brightness increases slightly less than linearly with increase in frequency, until a saturation is reach at about 100,000 Hz. Maximum efficiency is not obtained under the same operating conditions as maximum brightness. Although flat-panel lamps were developed for general lighting application at 120 V with an active layer as thin as 1 mil (25 micrometers), the efficiency of these devices is low compared to conventional light sources; for example, electroluminescence, 3–5 lumens/watt; incandescent lamp, 16 lm/W; and fluorescent lamp,

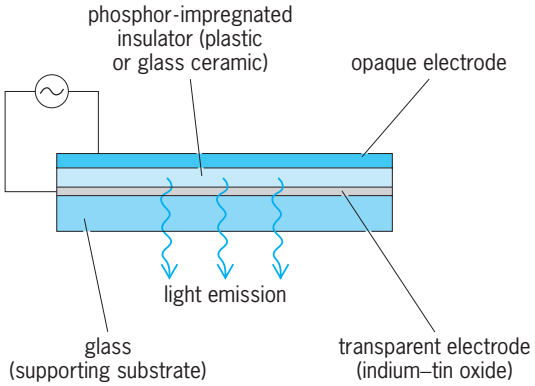


Fig. 1. Structure of powdered-phosphor (Destriau) electroluminescent cell, edge view (not to scale).

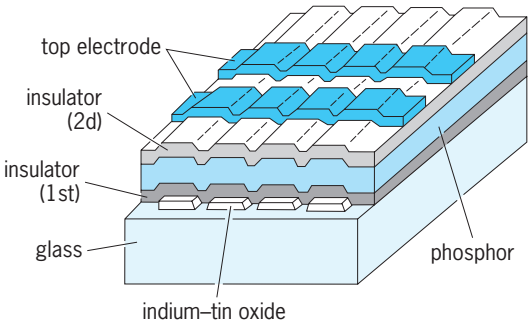


Fig. 2. Structure of a thin-film electroluminescent matrix display. (*After J. Kuwata et al., Feasibility of a 40 × 40 cm² TFEL display with 2000 × 2000 lines operating at 50 W, Society for Information Display Digest, pp. 297–300, 1988*)

80 lm/W. Electroluminescent lamps are therefore not competitive for general illumination, but they have many specialized uses. *See* LIGHT PANEL.

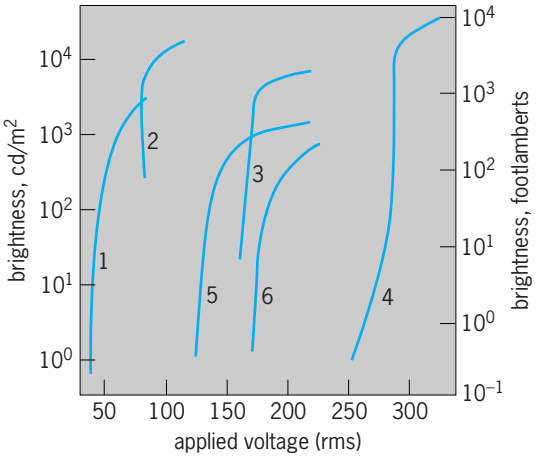
Thin-film electroluminescence. The application of electroluminescence to display and image formation received great impetus from work in the late 1960s and mid-1970s on thin-film electroluminescence (TFEL), giving rise to devices that are different in structure and mechanism from the Destriau conditions. The phosphor in these devices is not a powder but a thin (about 500 nanometers) continuous film prepared by sputtering or vacuum evaporation. The luminescence activators are manganese or rare-earth ions, atomic species with internal electronic transitions that lead to characteristic luminescence. The phosphor film does not contain copper sulfide or any other separate phase, and is sandwiched between two thin (about 200 nm) transparent insulating films also prepared by evaporative means. Conducting electrodes are applied to the outside of each insulating film; one of the electrodes is again a transparent coating of indium–tin oxide on glass, which serves as supporting substrate. If an imaging matrix is desired, both electrodes consist of grids of parallel lines, with the direction of the grid on one insulator (row) orthogonal to the other grid (column). By approximate circuitry the entire matrix can be scanned, applying voltage where desired to a phosphor element that is located between the intersection of a row and column electrode (**Fig. 2**).

A thin-film electroluminescent device is thus a symmetrical structure (conductor–insulator–phosphor–insulator–conductor). It acts like a pure capacitor at low applied voltage; no light is emitted until the voltage reaches a threshold value determined by the dielectric properties of the insulator and phosphor films (**Fig. 3**). Above this threshold the device is no longer purely capacitive, a dissipative current flows, and light emission occurs. The brightness increases very steeply with the applied voltage but is finally saturated. The light output, or average brightness, is roughly proportional to the frequency up to at least 5 kHz, and also depends on the waveform of the applied voltage.

Thin-film electroluminescent devices are believed to operate as follows. As voltage is applied, electrons trapped at the insulator–phosphor interface are released when the electric field reaches a critical value of about 5 MV/in. (2 MV/cm). The electrons are accelerated through the phosphor film, causing impact excitation of the activator centers, which then emit their characteristic luminescence. After the electrons complete their passage through the phosphor, they are retrapped at the second insulator–phosphor interface; on reversal of polarity the action is repeated and the electrons return to the original interface.

The best thin-film electroluminescent phosphor is manganese-activated zinc sulfide, which emits yellow light peaking at 585 nm. A brightness of approximately 10^4 footlamberts (3.5×10^4 candelas/m^2) has been achieved with this phosphor at saturation under 5 kHz excitation, and it exhibits a brightness of 25 footlamberts (85 cd/m^2) and an efficiency of 2 lm/W in a conventional thin-film electroluminescent matrix application at a frequency of 60 Hz. Activation of zinc sulfide and certain alkaline earth sulfides with different rare earths has yielded many other promising electroluminescent phosphors emitting blue, green, red, and white, and making full-color matrix-addressed thin-film electroluminescent displays possible. The light output of thin-film electroluminescent displays has been very reliable, with typically only 10% loss after tens of thousands of hours of operation.

Hysteresis in brightness versus voltage curves offers the possibility of using thin-film electroluminescence in memory devices. In conjunction with



Curve	Emission color	Activator	Insulator	Frequency (sine wave)
1	Yellow	Manganese	Lead titanate	5 kHz
2	Yellow	Manganese	Aluminum oxide	5 kHz
3	Yellow	Manganese	Aluminum oxide	10 kHz
4	Yellow	Manganese	Yttrium oxide	5 kHz
5	Green	Terbium	Yttrium oxide	5 kHz
6	Red	Samarium	Yttrium oxide	5 kHz

Fig. 3. Brightness-voltage characteristics of some zinc sulfide thin-film electroluminescent devices as a function of activator, insulating film, and operating frequency. (*After R. Mach and G. O. Müller, Physical concepts of high-field thin-film electroluminescent devices, Phys. Stat. Sol. (a), 69:11–66, 1982*)

photoconductors, electroluminescence can also be employed in light amplifiers and logic circuits for computers, although semiconductor logic circuits are faster. *See* ELECTRONIC DISPLAY; LIGHT AMPLIFIER.

Injection electroluminescence. Injection electroluminescence results when a semiconductor *pn* junction or a point contact is biased in the forward direction. This type of emission was first observed from silicon carbide (SiC) in England by H. J. Round in 1907. It is the result of radiative recombination of injected minority carriers with majority carriers amaterial. Such emission has been observed in a large number of semiconductors, including silicon (Si), germanium (Ge), diamond, cadmium sulfide (CdS), zinc sulfide (ZnS), zinc selenide (ZnSe), zinc telluride (ZnTe), zinc oxide (ZnO), and many of the so-called III-V compounds, such as gallium phosphide (GaP), gallium arsenide (GaAs), indium phosphide (InSb), boron phosphide (BP), boron nitride (BN), and aluminum nitride (AlN). The wavelength of the emission corresponds to an energy equal, at most, to the forbidden band gap of the material, and hence in most of these materials the wavelength is in the infrared region of the spectrum. In some cases the efficiency is high, approaching one emitted photion. In suitable structures the excitation intensity can become so high that stimulated rather than spontaneous emission predominates, and laser action results, with spectral narrowing and coherent emission. Direct electrical excitation is more convenient than optical excitation, but the beam divergence of injection lasers is normally much greater than that of gas lasers or the optically pumped solid type. *See* JUNCTION DIODE; LASER; SEMICONDUCTOR; SEMICONDUCTOR DIODE.

Other effects. If a *pn* junction is biased in the reverse direction, so as to produce high internal electric fields, other types of emission can occur, but with very low efficiency. The presence of very energetic ("hot") carriers can result in emission at energies greater than the band gap of the material (avalanche emission). The emission in this case may be correlated with small active regions called microplasmas. Light emission may also occur when electrodes of certain metals, such as Al or Ta, are immersed in suitable electrolytes and current is passed between them. In many cases this galvanoluminescence is electroluminescence generated in a thin oxide layer formed on the electrode by electrolytic action. In addition to electroluminescence proper, other interesting effects (usually termed electrophotoluminescence) occur when electric fields are applied to a phosphor which is concurrently, or has been previously, excited by other means. These effects include a decrease or increase in steady-state photoluminescence brightness when the field is applied, or a burst of afterglow emission if the field is applied after the primary photoexcitation is removed. *See* PHOTOLUMINESCENCE.

James H. Schulman; Clifford C. Klick

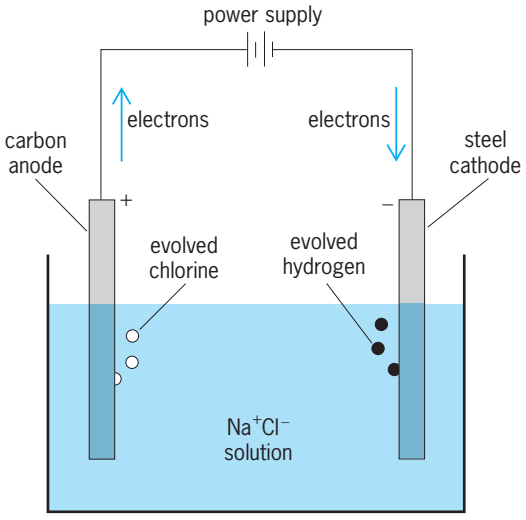
Bibliography. H. K. Henisch, *Electroluminescence*, 1962; H. F. Ivey, *Electroluminescence and Related Effects*, 1963; A. H. Kitai (ed.), *Solid State Luminescence: Theory, Materials, and Devices*, 1993; J. I. Pankove (ed.), *Electroluminescence*, 1977; S. Shionoya (ed.), *Electroluminescence*, 1989.

Electrolysis

A means of producing chemical changes through reactions at electrodes in contact with an electrolyte by the passage of an electric current. Electrolysis cells, also known as electrochemical cells, generally consist of two electrodes connected to an external source of electricity (a power supply or battery) and immersed in a liquid that can conduct electricity through the movement of ions. Reactions occur at both electrode-solution interfaces because of the flow of electrons. Reduction reactions, where substances add electrons, occur at the electrode called the cathode; oxidation reactions, where species lose electrons, occur at the other electrode, the anode. In the cell shown in the **illustration,** water is reduced at the cathode to produce hydrogen gas and hydroxide ion; chloride ion is oxidized at the anode to generate chlorine gas. Electrodes are typically constructed of metals (such as platinum or steel) or carbon. Electrolytes usually consist of salts dissolved in either water or a nonaqueous solvent, or they are molten salts. *See* ELECTROCHEMISTRY; ELECTRODE; ELECTROLYTE; OXIDATION-REDUCTION.

Applications include industrial synthesis of chemicals, electroplating of metals, metallurgical extraction and refining of metals, metal finishing and electromachining, and the production of electricity in batteries. Corrosion of metals often occurs through electrolytic processes. Electrolytic cells are used in analytical chemistry and in laboratory studies of reaction mechanisms. *See* BATTERY; CORROSION.

Principles. The flow of current, measured in amperes (where 1 A equals the passage of 1 coulomb of charge per second), represents the rate of flow of



Schematic diagram of an electrolysis cell in which the electrolyte is a solution of sodium chloride.

electrical charge through the electrolysis cell. The quantity of a substance produced or consumed in an electrode reaction is proportional to the quantity of electricity (coulombs) passed during the electrolysis a constant current of 1 A passed for 1 h (equivalent to 3600 coulombs) will produce 0.018656 mole or 1.3228 g of chlorine gas. The quantity of electricity is the integral of the current over the duration of the electrolysis and can be determined with a coulometer. *See* COULOMETER; ELECTROCHEMICAL EQUIVALENT.

In bulk electrolysis, for example, in the production of chemicals, the flow of current produces appreciable changes in the concentration of species in the electrolyte. Cells for bulk electrolysis usually employ relatively large electrodes with the liquid kept in constant motion. For analytical applications or the characterization of chemical systems, the electrolysis occurs only near the surface of the electrodes. Here, unstirred solutions and small electrodes are typically used.

When only one reaction occurs at an electrode, it is said to occur with 100% current efficiency. When two or more reactions occur at the same electrode, for example, when both oxygen and chlorine are evolved at the anode of a cell, the current efficiency of each reaction is given by the fraction of the quantity of electricity consumed for each process. For large-scale electrolytic processes, the power consumption (or energy) efficiency is of interest because it is related to the cost of electrical energy needed to produce a given amount of product. The energy efficiency is the ratio of the ideal theoretical energy for the cell reaction to the actual energy (directly proportional to the cell voltage) consumed in the electrolysis. The cell voltage depends upon a number of factors, including the thermodynamic energy requirements for the electrode reactions, the energy needed to drive the reactions at a given rate, and the resistance of the electrodes and electrolyte. *See* OVERVOLTAGE.

In analytical applications the electrode reaction at only one of the electrodes is of interest, and a third electrode, called a reference electrode, is usually introduced into the cell so that the potential of that electrode can be determined with respect to a known reference potential. The measured potential is also less perturbed by solution resistance effects in a three-electrode cell. *See* REFERENCE ELECTRODE.

Applications. There are many industrial applications for the production of important inorganic chemicals. Chlorine and alkali are produced by the large-scale electrolysis of brine (the chloralkali process) in cells carrying out the same reactions as those shown in the illustration. Other chemicals produced include hydrogen and oxygen (via water electrolysis), chlorates, peroxysulfate, and permanganate. Other processes are carried out with molten salts as solvents, because the electrode reactions of interest would be masked in aqueous solutions by the electrolysis of water to produce hydrogen and oxygen. Electrowinning of aluminum, magnesium, and sodium metals is carried out by molten salt electrolysis. Fluorine is produced by electrolysis of a 2:1 mixture of anhydrous hydrofluoric acid and potassium fluoride.

The major electrolytic processes involving organic compounds are the hydrodimerization of acrylonitrile to produce adiponitrile and the production of tetraethyllead. Many other organic compounds have been studied on the laboratory scale.

Electroplating involves the electrochemical deposition of a thin layer of metal on a conductive substrate, for example, to produce a more attractive or corrosion-resistant surface. Chromium, nickel, tin, copper, zinc, cadmium, lead, silver, gold, and platinum are the most frequently electroplated metals. Metal surfaces can also be electrolytically oxidized (anodized) to form protective oxide layers. This surface-finishing technique is most widely used for aluminum but is also used for titanium, copper, and steel. Electrolytic capacitor production involves anodization of aluminum, tantalum, and niobium. Anodic dissolution of a metal is used in electrochemical machining to produce a desired structure. It is mainly applied in the machining of very hard alloys or in producing complex structures that would be difficult to make by conventional machining methods. *See* ELECTROPLATING OF METALS.

Metals can be purified by electrorefining. Here, the impure metal is used as the anode, which dissolves during the electrolysis. The metal is plated, in purer form, on the cathode. Copper, nickel, cobalt, lead, and tin are all purified by this technique. *See* ELECTROMETALLURGY.

Electroanalysis involves the use of electrolytic processes to identify and quantitate a species. Coulometric methods are based on measuring the quantity of electricity used for a desired process. Voltammetric methods allow characterization of species through an analysis of the effect of potential and electrolysis conditions on the observed currents. Interest in photo-electrolysis is growing; this involves the utilization of radiant (for example, solar) energy to produce electricity or to drive chemical reactions in electrolytic cells. In such cells, which have not yet found practical application, the irradiated electrodes are usually made of semiconductor materials, for example, titanium dioxide, gallium arsenide, or cadmium selenide. Electrolytic cells can also be used to purify waste streams, for example, by the cathodic deposition and recovery of metals and the oxidation of organic pollutants. *See* POLAROGRAPHIC ANALYSIS.

Allen J. Bard

Bibliography. A. J. Bard and H. Lund (eds.), *The Encyclopedia of the Electrochemistry of the Elements*, 1973–1986; A. J. Bard and L. R. Faulkner, *Electrochemical Methods: Fundamentals and Applications*, 2d ed., 2000; D. R. Crow, *Principles and Applications of Electrochemistry*, 4th ed., 1994; H. Lund and O. Hammerich (eds.), *Organic Electrochemistry*, 4th ed., 2000; D. Pletcher and F. C. Walsh, *Industrial Electrochemistry*, 2d ed., 1990; M. Schlesinger and M. Paunovic (eds.), *Modern Electroplating*, 4th ed., 2000.

Electrolyte

A material that conducts an electric current when it is fused or dissolved in a solvent, usually water. Electrolytes are composed of positively charged species, called cations, and negatively charged species, called anions. For example, sodium chloride (NaCl) is an electrolyte composed of sodium cations (Na^+) and chlorine anions (Cl^-). The ratio of cations to anions is always such that the substance is electrically neutral. If two wires connected to a light bulb and to a power source are placed in a beaker of water, the light bulb will not glow. If an electrolyte, such as sodium chloride, is dissolved in the water, the light bulb will glow because the solution can now conduct electricity. The amount of electric current that can be carried by an electrolyte solution is proportional to the number of ions dissolved. Thus, the bulb will glow more brightly if the amount of sodium chloride in the solution is increased. *See* ELECTRIC CURRENT; ION.

Hydration. Water is a special solvent because its structure has two different sides. On one side is the oxygen atom, and on the other are two hydrogen atoms. Covalent molecules, such as water, are held together by covalent bonds, which are formed when two atoms share a single pair of electrons. However, when two different atoms form a covalent bond, the sharing of electrons is not always equal. An electron in a covalent bond between two different atoms might spend more time near one atom or the other. In water, the electrons from the O-H bonds spend more time near the oxygen atom than near the hydrogen atoms. As a result, the oxygen has somewhat more negative charge than the hydrogen atoms. This phenomenon is not the same as ionization, where the electron is completely transferred from one atom to the other; so to indicate the difference, the oxygen is said to have a partial negative charge (δ^-) and the hydrogen to have a partial positive charge (δ^+). *See* CHEMICAL BONDING.

Since water has two different sides with opposite partial charges, it is said that water is polar. When ions are dissolved in polar solvents such as water, the solvent molecules are organized around the ion to create the maximum amount of electrostatic attraction. Thus, if a cation is dissolved in water, the water molecules will organize themselves around the ion with the oxygen atoms directed toward the cation, because the partial negative charges on the oxygen atoms in the water molecules are electrostatically attracted to the positive charge on the cation. Likewise, water molecules will organize about anions with the partially positive hydrogen atoms directed toward the negatively charged anion. When water molecules surround dissolved ions, the ions are said to be hydrated. The electrostatic attraction associated with hydration provides the energetic driving force for dissolving ions. Since water is polar, it will also dissolve polar molecules that are not charged, such as ethanol (CH_3CH_2OH), which contains a polar O-H bond. Thus, water can dissolve both ionic substances and polar molecules that are not charged. *See* SOLVENT; WATER.

Strong and weak electrolytes. Any substance that produces ions when dissolved is an electrolyte. These substances include ionic materials composed of simple monatomic ions, such as sodium chloride, or substances composed of polyatomic ions, such as ammonium nitrate (NH_4NO_3). When these substances are dissolved, hydrated ions are generated, as in reactions (1) and (2).

$$NaCl \rightarrow Na^+ + Cl^- \quad (1)$$

$$NH_4NO_3 \rightarrow NH_4^+ + NO_3^- \quad (2)$$

A special type of electrolyte is an acid, in which the cation is H^+. When acids are dissolved in water, H^+ ions are produced along with an anion, which can be either a monatomic ion, as in hydrochloric acid (HCl), or a polyatomic ion, as in nitric acid (HNO_3) or acetic acid ($HC_2H_3O_2$), as in reactions (3)–(5).

$$HCl \rightarrow H^+ + Cl^- \quad (3)$$

$$HNO_3 \rightarrow H^+ + NO_3^- \quad (4)$$

$$HC_2H_3O_2 \rightarrow H^+ + C_2H_3O_2^- \quad (5)$$

Electrolytes such as sodium hydroxide (NaOH) that yield the hydroxyl (OH^-) anion when dissolved in water are called bases [reaction (6)]. Some

$$NaOH \rightarrow Na^+ + OH^- \quad (6)$$

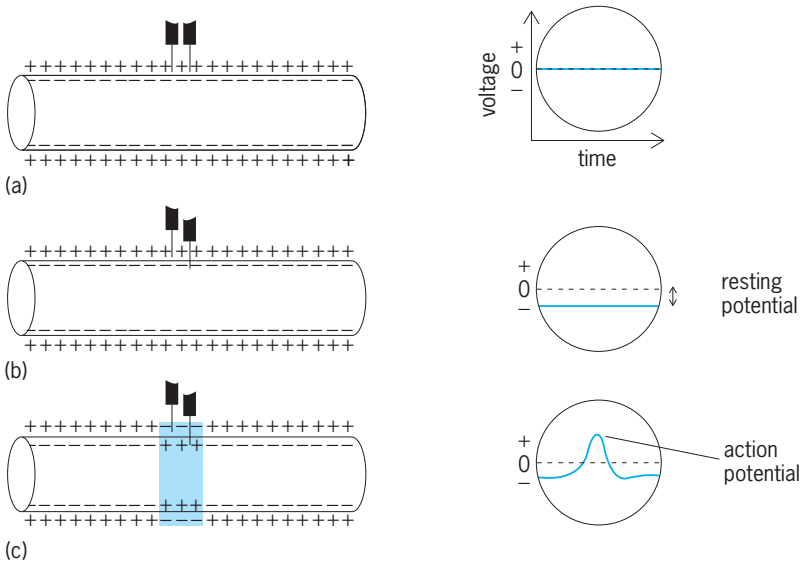
molecules that do not contain ions, such as ammonia (NH_3), generate OH^- ions when dissolved in water as in reaction (7); these bases are also electrolytes.

$$NH_3 + H_2O \rightarrow NH_4^+ + OH^- \quad (7)$$

Polar covalent molecules, such as ethanol, dissolve in water but do not generate any ions and are called nonelectrolytes.

Many electrolytic substances completely dissociate into ions when they are dissolved in water. In these cases, the reactions shown above would proceed completely to the right, leaving only dissolved ions and no associated, electrically neutral molecules. For example, when sodium chloride is dissolved in water, all of the dissolved material is present as Na^+ and Cl^- ions, with no dissolved NaCl molecules. Substances that completely dissociate are called strong electrolytes, because every molecule dissolved generates ions that contribute to the electrical conductivity. Ionic substances, such as NaCl and NH_4NO_3, are strong electrolytes. Acids and bases that completely dissociate are called strong acids and strong bases, and these substances are also strong electrolytes. Hydrochloric acid (HCl) and HNO_3 are strong acids, so every molecule of HCl or HNO_3 that dissolves generates a H^+ cation and a Cl^- or NO_3^- anion, respectively.

Some substances dissolve in water but do not dissociate completely. For example, when acetic acid is dissolved in water, some molecules dissociate to form H^+ and $C_2H_3O_2^-$ ions, while others remain



Measurement of the electric potential across a neuronal membrane by using an oscilloscope. (*a*) With both electrodes outside the membrane, no potential is measured. (*b*) With one electrode inside the membrane and the neuron in the resting state, the resting potential is measured. (*c*) When a nerve impulse travels along the neuron, the charge is reversed, and an action potential that is the opposite of the resting potential is measured for a brief period. The shaded region represents the portion where the polarity is reversed. (*After H. Curtis and N. S. Barnes, Biology, 5th ed., Worth, 1989*)

associated as $HC_2H_3O_2$ units, which contain polar O-H bonds and are therefore readily soluble in water. Acids such as acetic acid that do not dissociate completely are called weak acids. Similarly, the reaction of ammonia with water in reaction (7) proceeds to only a small extent. Thus, some molecules of ammonia react with water to generate OH^- and NH_4^+ ions, but many simply remain as NH_3. Since each dissolved molecule of ammonia does not generate an OH^- anion, ammonia is said to be a weak base. The electrical conductivity of an electrolyte solution depends on the number of ions present, so a solution of acetic acid or ammonia will not conduct as well as a solution of an equal amount of a strong electrolyte. Thus, when a light bulb is connected to a solution of acetic acid, it will not glow as brightly as a light bulb connected to a solution of sodium chloride. In fact, when 100 acetic acid molecules are dissolved in water, 99 exist as associated $HC_2H_3O_2$ molecules, while only one dissociates to form H^+ and $C_2H_3O_2^-$ ions. Therefore, in order to get an equal amount of electrical conductivity as a solution of a strong electrolyte, 100 times more acetic acid must be dissolved. For this reason, substances that dissociate somewhat but not completely are called weak electrolytes. Weak acids and weak bases are the most common types of weak electrolytes. *See* IONIC EQUILIBRIUM.

Physiological roles. In biological systems, electrolytes play important roles in regulating kidney function and the retention of water. Electrolytes are also vital for providing the electric current needed for nerve impulses in neurons. When the ratio of cations to anions is different on two sides of a cell membrane, the electric charge on either side of the membrane is different, leading to the establishment of an electric potential. The cell membranes of neurons are impermeable to anions, but they contain special proteins that maintain a higher concentration of cations outside the neuron by either actively transporting sodium (Na^+) and potassium (K^+) ions into and out of the cell or by changing the permeability of the membrane to these cations. Since anions cannot follow the cations out of the neuron, a negative charge builds up inside the neuron and a positive charge builds up outside the neuron, generating an electric potential. When no impulse is being transmitted by the neuron, it is said to be in its resting state and the resulting electric potential is called the resting potential. If two electrodes are placed outside the neuron in its resting state, no potential is measured (**illus.** *a*); however, if one electrode is inside the neuron and the other electrode is outside, the resting potential can be measured (illus. *b*). *See* ELECTRODE; OSMOREGULATORY MECHANISMS.

When a nerve impulse is generated, a localized section of the neuron suddenly becomes permeable to Na^+ cations. Since the concentration of cations is higher outside the neuron, Na^+ ions rapidly flow into the neuron because of attraction to the excess negative charge inside and the concentration gradient. An excess of positive charge is built up inside the neuron, and the electrical charge is reversed. As a result, a new action potential that is opposite of the resting potential is generated (illus. *c*). When the Na^+ ions enter the neuron, the membrane becomes permeable to K^+ cations, which rapidly flow out of the neuron, restoring the resting potential. The action potential therefore persists for only a half millisecond. This process is repeated at adjacent sites in a single direction along the neuron, propagating the nerve impulse. *See* BIOPOTENTIALS AND IONIC CURRENTS; SYNAPTIC TRANSMISSION.

H. Holden Thorp

Bibliography. A. J. Bard and L. R. Faulkner, *Electrochemical Methods: Fundamentals and Applications*, 2000; R. Chang, *Chemistry*, 1991; H. Curtis and N. S. Barnes, *Biology*, 1989; L. Stryer, *Biochemistry*, 4th ed., 1995; G. Wulfsberg, *Principles of Descriptive Inorganic Chemistry*, 1992.

Electrolytic conductance

The transport of electric charges, under electric potential differences, by particles of atomic or larger size. This phenomenon is distinguished from metallic conductance, which is due to the movement of electrons. The charged particles that carry the electricity are called ions.

Positively charged ions are termed cations; the sodium ion, Na^+, is an example. The negatively charged chloride ion, Cl^-, is typical of anions. The negative charges are identical with those of electrons or integral multiples thereof. The unit positive charges have the same magnitude as those of electrons but are of opposite sign. Colloidal particles, which may have relatively large weights, may be ions, and may carry many positive or negative

charges. Electrolytic conductors may be solids, liquids, or gases. Semiconductors have properties that are intermediate between the metallic conductors and insulators.

Measurement. Conductances are usually reported as specific conductances κ, which are the reciprocals of the resistances of cubes of the materials, 1 cm in each dimension, placed between electrodes 1 cm square, on opposite sides. These units are sometimes called mhos, that is, ohms spelled backward. Conductances of solutions are usually measured by Friedrich Kohlrausch's method, in which a Wheatstone bridge is employed. Such a bridge is shown diagrammatically in **Fig. 1**. The resistances R_3 and R_4 (usually of the same value) form two arms of the bridge. Resistance R_2 is adjustable, and the remaining arm is the cell holding the electrolytic conductor, or as is usually stated, solution of electrolyte. Direct current and the usual galvanometers cannot be used because of an apparent failure of Ohm's law. Passage of direct current produces chemical reactions and a back electromotive force (emf) is generated by the galvanic action of the products. By using an alternating current, the electrochemical reactions occurring when the current is briefly passed in one direction may be reversed when the direction of the current is changed. When a small alternating-current input signal is used, practically all the electric charge passed during each half cycle is stored in the electric double layer, which acts as a capacitor. The electrodes are usually made of platinum and are platinized, that is, coated with finely divided platinum. The surface area, and hence the electrode capacitance, is thereby greatly increased. By making measurements at several frequencies and extrapolating to infinite frequency, the effect of electrode reactions can be eliminated. For less exact measurements, a fixed frequency of 60–1000 Hz is commonly used.

To determine the conductance C, that is, the reciprocal resistance $1/R$ of the cell of Fig. 1, the resistance R_2 of the bridge is adjusted until a minimum of sound is heard in the telephone.

Greater sensitivity may be obtained by electronic amplification of the off-balance signal and by using an oscilloscope to detect the point of balance. When the bridge is in balance, the conductance is given by the relation $C = R_4/R_2R_3$. From this the specific conductance κ may be obtained from the equation $\kappa = KC$, in which K is the cell constant. Occasionally,

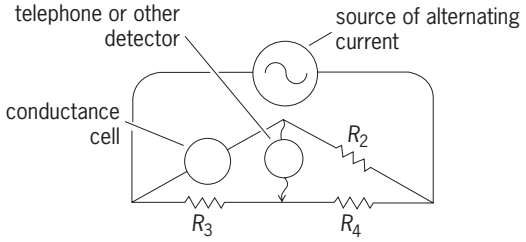


Fig. 1. Wheatstone bridge circuit for the measurement of electrolytic conductance. R_3 and R_4 are fixed resistances; R_2 is a variable resistance.

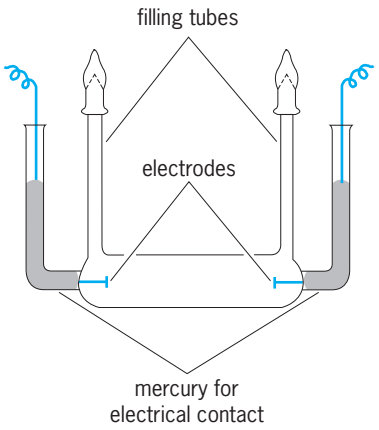


Fig. 2. Diagram of conductance cell.

this constant can be computed from the dimensions of the cell. Usually, however, it is determined by using a solution whose κ value is accurately known from measurements in such a cell, or by comparison with the specific conductance of mercury.

For precision work, care must be taken to avoid errors due to electrical reactances. This has been done in specially designed bridges. A typical, properly designed conductance cell is shown in **Fig. 2**. The cell is filled with solution through the center tubes. Electrical contact is made with the electrodes by platinum wires sealed through the glass wall. These connect the mercury in the outside tubes, which are widely spread to avoid errors due to electrical capacity.

Equivalent conductance. Although many substances and mixtures show electrolytic conductance, the greater part of the research on the subject has been on aqueous solutions of salts, acids, and bases. There have been considerable data accumulated for solutions of such electrolytes in nonaqueous solvents, such as alcohols. The data are usually given in terms of equivalent conductance Λ, which is defined by Eq. (1), in which κ is the

$$\Lambda = \frac{1000\kappa}{c} \tag{1}$$

specific conductance and c is the concentration in equivalents per liter. Values of Λ change with the concentration and, in general, increase as the solutions measured are made more dilute, that is, as c is decreased. A plot of values of the equivalent conductance λ against $\sqrt{c}$ for some typical electrolytes is shown in **Fig. 3**. Svante Arrhenius, who was the first to assume that electrolytic conductance is due to freely moving charged ions, explained the decrease of Λ with increasing c by assuming that the number of ionic carriers gets smaller as the concentration increases, and he computed a degree of dissociation α by formula (2). The term Λ_0 is obtained by

$$\alpha = \frac{\Lambda}{\Lambda_0} \tag{2}$$

determining Λ at a series of low concentrations and extrapolating to a limiting value, termed the equivalent conductance at infinite dilution. Though

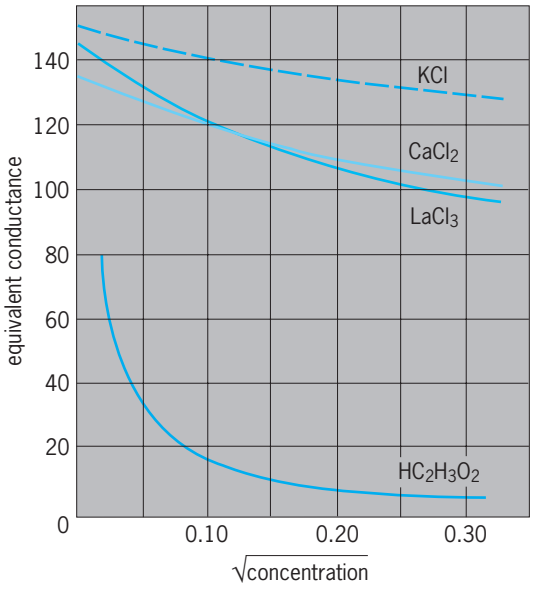


Fig. 3. Equivalent conductance at infinite dilution for some typical electrolytes.

Eq. (2) has been shown by later work to give nearly the right values of α for certain poorly conducting solutions, it is now considered to be much in error for the so-called strong electrolytes. These include most salts, such as potassium chloride, KCl, and sodium sulfate, Na_2SO_4, and inorganic acids and bases, such as hydrochloric acid, HCl, and sodium hydroxide, NaOH. For an electrolyte which yields two types of ion, it can be shown that Eq. (3) holds,

$$\Lambda = F\alpha\,(U^+ + U^-) \tag{3}$$

in which F is the faraday and U^+ and U^- are the mobilities, or speeds, under unit potential difference of the positive and negative ions, respectively. For Eq. (2) to hold, these mobilities must be constant from the equivalent concentration c at which Λ is measured to infinite dilution.

Since the advent of the Debye-Hückel theory of interionic attractions, strong electrolytes have been considered to be completely dissociated, that is, the term α of Eq. (3) is equal to unity for these substances. The decreases observed in the values of the equivalent conductances Λ, with increases in concentration, are assumed to be due to reductions in the values of the ionic mobilities U^+ and U^-. According to the theory of P. Debye and E. Hückel, the ion possesses an ionic atmosphere distributed with radial symmetry around the ion as center. This is due to the fact that interionic attractions and repulsions, together with thermal vibrations, tend to produce a slight preponderance of negative ions around a positive ion, and vice versa. The presence of this atmosphere leads to the lowering of ionic mobilities with increasing ion concentrations. The adaptation of the Debye-Hückel theory for conducting solutions is due to Lars Onsager. His equation for very dilute uni-univalent electrolytes, such as sodium chloride, is shown in Eq. (4), in which θ and σ are given for uni-univalent electrolytes by Eqs. (5) and (6), in which D

$$\Lambda = \Lambda_0 - (\theta\Lambda_0 + \sigma)\sqrt{c} \tag{4}$$

$$\theta = \frac{8.16 \times 10^5}{(DT)^{3/2}} \tag{5}$$

$$\sigma = \frac{8.28}{\eta(DT)^{1/2}} \tag{6}$$

is the dielectric constant at the absolute temperature T and η is the viscosity.

Equation (4) yields accurate values of the data for strong electrolytes up to concentrations of about 0.001 M, above which there are small deviations. Modifications of Eq. (4) for solutions of salts of higher valence types, such as calcium chloride, $CaCl_2$, and lanthanum chloride, $LaCl_3$, are available and have also been found to agree with the data for dilute solutions.

Onsager, in his derivation of Eq. (4), treated the ions as point charges. Later Raymond Fuoss and Onsager extended the theory to include the radii of the ions and also the effects of higher concentrations. The ion sizes obtained from conductance data agree closely with those calculated from activity measurements. *See* ACTIVITY (THERMODYNAMICS).

Equation (4), or its empirical and theoretical extensions, can be used to obtain values of Λ_0 from the data on equivalent conductances of dilute solutions. Some typical figures for limiting equivalent conductances Λ_0 of typical strong electrolytes in aqueous solution at 25°C (77°F) are listed below.

HCl	426.16	NH_4Cl	149.86	$MgCl_2$	129.40
NaOH	247.8	KNO_3	144.96	NaI	126.94
KBr	151.9	$CaCl_2$	135.84	NaCl	126.45
KI	150.38	Na_2SO_4	129.9	LiCl	115.03

Ionic conductances. The fraction of the current carried by a given type of ion is known as the transference number, t, which depends on the relative mobilities of the ions. This relation is shown in Eqs. (7). The ionic mobilities, and therefore the transference numbers, vary somewhat with concentration, but the limiting equivalent conductances Λ_0 for various electrolytes can be accurately represented as the sum of the limiting ionic conductances. Thus, for potassium chloride, $\Lambda_{0,KCl} = \lambda_{0,K^+} + \lambda_{0,Cl^-}$, and the value of λ_{0,Cl^-} is the same whether it is derived from measurements on HCl, NaCl, or KCl solutions. This additive relation is known as Kohlrausch's law of the independent mobility of ions. However, it is necessary to obtain the value of λ_0 for at least one ion constituent independently in order to establish the ion conductances of the other ions. The relation used is shown in Eqs. (8), in which λ_0 is the

$$t^+ = \frac{U^+}{U^+ + U^-} \qquad t^- = \frac{U^-}{U^+ + U^-} \tag{7}$$

$$t_0{}^+\Lambda_0 = \lambda_0{}^+ \quad \text{or} \quad t_0{}^-\Lambda_0 = \lambda_0{}^- \tag{8}$$

limiting equivalent conductance of an electrolyte and t_0^+ and t_0^- are the limiting transference numbers of the positive and negative ion constituents, respectively.

The same value of λ_{0,Cl^-}, within 0.02%, is obtained from precision conductance and transference measurements on solutions of hydrogen, lithium, sodium, and potassium chlorides. Values of the limiting ionic conductance at 25°C (77°F) are given below for some ions.

H^+	349.82	$^1/_2Ba^{2+}$	63.64
OH^-	198	Ag^+	61.92
SO_4^{2-}	79.8	$^1/_2Ca^{2+}$	59.50
Br^-	78.4	$^1/_2Mg^{2+}$	53.06
I^-	76.8	Na^+	50.11
Cl^-	76.34	HCO_3^-	44.48
K^+	73.52	$CH_3CO_2^-$	40.9
NH_4^-	73.4	$CH_2ClCO_2^-$	38.7
NO_3^-	71.44	Li^+	38.69

Nonaqueous systems. In addition to the study of water solutions of electrolytes, considerable study has been given to electrolytes in nonaqueous and mixed solvents. In general, the same principles as those outlined above apply to the interpretation of the results. However, fewer of the electrolytes are completely dissociated, and the degrees of dissociation of the weaker acids and bases are lower. This is due to the fact that, in general, the dielectric constants of nonaqueous solvents are smaller than those of water, so that the attractions between positive and negative ions are greater.

It will be observed that in this article discussion is confined to quite dilute solutions of electrolytes. For concentrated solutions few generalizations of any value can be given.

Molten salts exhibit a wide range of conductivities, depending upon their structures. Salts of alkali and alkaline earth metals usually are largely ionic in character and are highly conductive in the molten state, whereas heavy metal salts may be essentially covalent and exhibit little or no conductivity. Thus the conductivities of liquid arsenic chloride ($AsCl_3$) and bismuth chloride ($BiCl_3$) near their melting points are approximately 10^{-6} and 0.44 ohm^{-1} cm^{-1}, respectively, reflecting the more ionic structure of $BiCl_3$.

If, instead of using quite low potentials in the measurement of electrolytic conductances, voltages of the order of 100,000 are employed, the conductances observed are no longer constant but tend to increase with the potential used. Under these conditions Ohm's law evidently is not valid. This increase of conductance with high potentials is called the Wien effect. This effect is in accord with the interionic attraction theory. When the velocity of the ions becomes sufficiently great, the ion atmospheres do not have time to form to their full extent, so that both the electrophoretic and time of relaxation effects exert less influence on the conductance. However, a large Wien effect is also found for weak acids and bases. It would appear that the high potentials produce, temporarily, additional ionization of these substances. This explanation has been proposed and discussed theoretically by Onsager. If very high frequencies are used in the measurements, an increase in the conductance, termed the Debye-Falkenhagen effect, is observed. This can also be explained by the interionic attraction theory.

Solids. Electrical conduction in solids can range from purely electronic (in metals and semiconductors) to purely ionic (in solids that are electronic insulators but have ions with appreciable mobilities). Examples of purely ionic solid conductors are silver iodide and silver sulfide, in which all of the current is carried by silver ions. In other cases, such as metallic oxides, both ions may serve as charge carriers. Ionic conduction must lead to transport of matter and to chemical reactions at the electrodes, whereas electronic conduction leads only to a transfer of charge. *See* ELECTROCHEMISTRY; ELECTROMOTIVE FORCE (CELLS).

Herbert A. Laitinen

Bibliography. J. L. Copeland, *Transport Properties of Ionic Liquids*, 1974; D. Inman and D. G. Lovering (eds.), *Ionic Liquids*, 1981; R. F. Snipes, *Statistical Mechanical Theory of the Electrolytic Transport of Non-Electrolytes*, 1974.

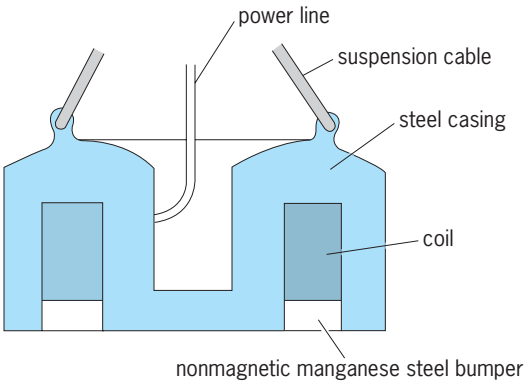
Electromagnet

A soft-iron core that is magnetized by passing a current through a coil of wire wound on the core. Electromagnets are used to lift heavy masses of magnetic material and to attract movable magnetic parts of electric devices, such as solenoids, relays, and clutches.

The difference between cores of an electromagnet and a permanent magnet is in the retentivity of the material used. Permanent magnets, initially magnetized by placing them in a coil through which current is passed, are made of retentive (magnetically "hard") materials which maintain the magnetic properties for a long period of time after being removed from the coil. Electromagnets are meant to be devices in which the magnetism in the cores can be turned on or off. Therefore, the core material is nonretentive (magnetically "soft") material which maintains the magnetic properties only while current flows in the coil. All magnetic materials have some retentivity, called residual magnetism; the difference is one of degree. *See* MAGNETIZATION.

A magnet, when brought near other susceptible material, induces magnetic poles in the susceptible material and so attracts it. A force is developed in the susceptible material that tends to move it in a direction to minimize the reluctance of the flux path of the magnet. The reluctance force may be expressed quantitatively in terms of the rate of change of reluctance with respect to distance. *See* MAGNETISM.

In an engineering sense the word electromagnet does not refer to the electromagnetic forces incidentally set up in all devices in which an electric current



Cross section of circular lifting electromagnet.

exists, but only to those devices in which the current is primarily designed to produce this force, as in solenoids, relay coils, electromagnetic brakes and clutches, and in tractive and lifting or holding magnets and magnetic chucks.

Electromagnets may be divided into two classes: traction magnets, in which the pull is to be exerted over a distance and work is done by reducing the air gap; and lifting or holding magnets, in which the material is initially placed in contact with the magnet. For examples of the first type *see* BRAKE; CLUTCH; RELAY; SOLENOID (ELECTRICITY).

Examples of the latter type are magnetic chucks and circular lifting magnets. The **illustration** shows a typical circular lifting magnet. The outer rim makes up one pole and the inner area is the opposing pole. The coil is wound cylindrically around the center pole. Manganese steel, used as a protective cover plate for the coil, is nonmagnetic and thus forces the flux through the magnetic member being lifted.

The mechanical force between two parallel surfaces is given by Maxwell's equation, shown below,

$$F = B^2A/2\mu_0 \text{ (newtons)}$$

where B is the flux density (in webers/m^2), A is the cross-sectional area (in m^2) through which the flux passes, and μ_0 is the permeability of free space ($4\pi \times 10^{-7}$ henry/m). When two poles are active, the force produced by each is calculated to find the total force. An interesting result of this relation is that the force is not simply the result of the total flux (BA) but also of the flux density. Thus if the same flux can be forced through one-half the cross-sectional area, the net pull will be doubled. In practice it is difficult, if not impossible, to calculate the actual lifting capacity of the magnet by using Maxwell's equation since the capacity varies with the shape and kind of material lifted, how the material is stacked, and other factors. Therefore, lifting magnets are usually rated on their all-day average lifting capacity.

Since currents are large and the circuit is highly inductive, control of a lifting magnet is a problem. If the line switch were simply opened, a destructive arc would result due to the coil's inductance not allowing the current to change instantaneously to a zero value. Therefore, the controller employed with a lifting magnet usually does the following things automatically: (1) reduces magnet current after an initially high value to a lower holding value to reduce heating of the magnet, (2) introduces a shunt discharge resistor across the magnet coil before allowing the line to be opened when the operator turns the magnet off, and (3) causes a reduced current of reverse polarity to flow in the magnet coil for a short time after the operator turns the switch off. Thus the residual magnetism is canceled and scraps and small chunks that might have continued clinging to the magnet are released. Jerome Meisel

Bibliography. D. G. Fink and H. W. Beaty (eds.), *Standard Handbook for Electrical Engineers*, 14th ed., 2000; G. R. Jones, M. A. Laughton, and M. G. Say (eds.), *The Electrical Engineers' Reference Book*, 15th ed., 1993; E. M. Purcell, *Electricity and Magnetism*, 2d ed., 1984.

Electromagnetic compatibility

The situation in which electrical and electronic devices and systems work as intended, both within themselves and in their electromagnetic environment.

Electromagnetic interference (EMI) is said to exist when unwanted voltages or currents are present so that they adversely affect the performance of a device or system. Such voltages or currents may reach the victim circuit or device by conduction or by nonionizing radiation. In all cases, electromagnetic interference arises because of a combination of three factors: a source, a transmission path, and a response, at least one of which is unplanned. Electromagnetic interference control refers to the process of making design changes or adjustments of signal or noise levels in order to achieve electromagnetic compatibility (EMC). Examples of electromagnetic interference often experienced by the public are herringbone patterns on the television screen and buzzing and clicking sounds in the radio.

Sources of interference. Designers usually do not intend that their devices be sources of interference. However, what is a desired signal in one path may be an undesired signal, that is, noise, in a path into which it is inadvertently coupled. Interference may be unintentional, or nonfunctional, such as an arc discharge, radiation from a lightning stroke, a corona discharge from high-tension power lines, or a noise caused by a sudden change in current flow in a conductor. Functional interference, while not usually produced for interference purposes, often includes sine waves, computer clock pulses, speech or video waves, or pulses forming data trains. An example of functional interference is signal leakage from cable television systems. Fluorescent lamps, commutators, automotive ignition systems, and industrial, scientific, and medical equipment such as diathermy machines all constitute sources of interference, as does the electromagnetic pulse (EMP) which accompanies a nuclear detonation. Short-wave radio

listeners often experience interference from high-frequency radar from Russia. Interference also exists across national borders because of incomplete international agreement on the use of the radio-frequency spectrum. *See* ELECTRICAL INTERFERENCE; ELECTROMAGNETIC PULSE (EMP); RADIO SPECTRUM ALLOCATION.

Interference coupling. Electromagnetic interference always starts with a flow of current through a conductor (which may include a conducting gas), and always appears at the victim in the form of a current or a voltage. The coupling path, however, may be a conduction or a radiation path. Thus the actual paths include common wiring, capacitance between devices, mutual inductance between adjacent wiring, nonionizing radiation, or wires in an electromagnetic field. Such coupling is aided by the fact that all conductors, whether wires, printed circuit board traces, or the conductors within an integrated circuit, exhibit both resistance and inductance. Likewise, between wires, mutual inductance and capacitance exist, resulting in a variety of possible coupling paths. *See* CAPACITANCE; COUPLED CIRCUITS; INDUCTANCE.

Grounds and bonds. Grounding is the establishment of an electrically conductive path between an electrical or electronic element of a system and a reference point or plane known as the ground. Grounding thus is a circuit concept, whereas bonding refers to the physical implementation of that concept. Grounding may also refer to an electrical connection made to the earth.

Grounds for currents whose wavelength is long compared with circuit dimensions are best made directly to a single point in a system. For currents whose wavelength is comparable to, or shorter than, circuit dimensions, however, the multipoint ground must be used to avoid the situation in which the ground lead might be a significant fraction of a quarter-wavelength, and thus not serve as a low-impedance conductor. *See* ELECTRONIC EQUIPMENT GROUNDING; GROUNDING.

Shields. The purpose of shielding is to confine radiated energy to a specific region, or to prevent radiated energy from entering a specific region. The most effective shield is a solid metallic enclosure, made of a permeable metal (for example, iron or steel) if frequencies below 100 kHz are to be shielded, or of any metal if higher frequencies are to be shielded. However, the solid shield does not permit light, air, water, or other substances to be passed through it, so shields with holes, including screens, braids, and honeycomb arrangements, as well as conductive glass may be needed. The widespread use of plastic enclosures has made thin film shields vital in achieving the needed shielding effectiveness in the use of such enclosures. *See* ELECTRICAL SHIELDING.

Filters. An electrical filter offers relatively little opposition to the passage of certain frequencies or direct current (dc) while blocking the passage of other frequencies. Accordingly, filters play a significant role in reducing conducted interference to the extent that such interference has a spectral content different from that of the desired signals.

A filter may be either reflective or lossy. Reflective filters present an impedance mismatch to unwanted frequencies, thereby returning them to the input, whereas lossy filters absorb unwanted frequencies. A filter may be designed on a time-domain basis as well as on a frequency-domain basis. *See* ELECTRIC FILTER; IMPEDANCE MATCHING.

Mathematical models. The complexity of electronic circuitry no longer allows the use of simple trial-and-error methods in achieving an electromagnetically compatible system design. System behavior can be expressed in terms of equations that describe source outputs, transmission or coupling characteristics, and susceptor responses. Such models have been incorporated into system analysis programs which not only predict system compatibility but also can be used to evaluate the effect of parameter changes on overall system performance, as may be needed in analyzing a proposed subsystem specification waiver, or in generating interface specifications for use within an overall system design. *See* SIMULATION.

Standards. Numerous standards have been developed for military as well as for industrial and consumer products. For military equipment, MIL-STD-461B, *EMI Characteristics Requirements for Equipment*, defines the emission and susceptibility limits applicable to units under various test conditions; MIL-STD-462, *Measurement of EMI Characteristics*, describes how the measurements are to be made.

Such organizations as the Institute of Electrical and Electronics Engineers (IEEE), the Society of Automotive engineers (SAE), the Electronic Industries Association (EIA), and the American National Standards Institute (ANSI) are active in promulgating standards for electronic equipment, including electromagnetic compatability aspects.

The Federal Communications Commission's (FCC) *Rules and Regulations*, Part 15, as amended by FCC Docket 20780, establishes emission limits on "computing devices," defined as any electronic devices or systems that use digital techniques. A class A computing device is destined for commercial, industrial, or business use. It must meet certain radiation limits (for frequencies of 30 to 100 MHz) at a 96-ft (30-m) distance, and conduction limits (for frequencies of 0.45 to 30 MHz) with respect to its power leads. A class B computing device is destined for residential use. This device must meet certain radiation limits at a 10-ft (3-m) distance. Both its radiation limits and its conduction limits are more stringent than those of a class A device. *See* DIGITAL COMPUTER; MICROCOMPUTER.

Mutual effects. Digital systems, such as computers, tend to interfere with analog systems, such as voice and video communications, more readily than analog systems interfere with digital systems. Therefore, data streams to be transmitted over analog voice circuits are converted to a quasianalog tone form first. Computer clocks also may have to be shielded and

their output circuits may have to be filtered to prevent interference to communication equipment. In addition, personal computers must be connected to television receivers so that the video output of the personal computer does not reach the television receiving antenna, which then would radiate such signals. *See* ELECTRICAL NOISE. Bernhard E. Keiser

Bibliography. J. Goedbloed, *Electromagnetic Compatibility*, 1993; B. E. Keiser, *Principles of Electromagnetic Compatibility,* 3d ed., 1987; C. R. Paul, *Introduction to Electromagnetic Compatibility*, 1992; D. A. Weston, *Electromagnetic Compatibility Principles and Applications*, 1991.

Electromagnetic field

A changing magnetic field always produces an electric field, and conversely, a changing electric field always produces a magnetic field. This interaction of electric and magnetic forces gives rise to a condition in space known as an electromagnetic field. The characteristics of an electromagnetic field are expressed mathematically by Maxwell's equations. *See* ELECTRIC FIELD; ELECTROMAGNETIC RADIATION; ELECTROMAGNETIC WAVE; MAXWELL'S EQUATIONS. Jesse W. Beams

Electromagnetic induction

The production of an electromotive force either by motion of a conductor through a magnetic field in such a manner as to cut across the magnetic flux or by a change in the magnetic flux that threads a conductor.

Motional electromotive force. A charge moving perpendicular to a magnetic field experiences a force that is perpendicular to both the direction of the field and the direction of motion of the charge. In any metallic conductor, there are free electrons, electrons that have been temporarily detached from their parent atoms.

If a conducting bar (**Fig. 1**) moves through a magnetic field, each free electron experiences a force due to its motion through the field. If the direction of the motion is such that a component of the force on the electrons is parallel to the conductor, the electrons will move along the conductor. The electrons will move until the forces due to the motion of the conductor through the magnetic field are balanced by electrostatic forces that arise because electrons collect at one end of the conductor, leaving a deficit of electrons at the other. There is thus an electric field along the rod, and hence a potential difference between the ends of the rod while the motion continues. As soon as the motion stops, the electrostatic forces will cause the electrons to return to their normal distribution.

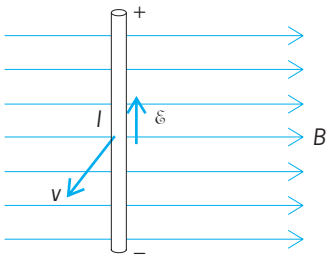


Fig. 1. Flux density *B*, motion *v*, and induced emf $\mathscr{E}$ when a conductor of length *l* moves in a uniform field. ***(After R. L. Weber, M. W. White, and K. V. Manning, Physics for Science and Engineering, McGraw-Hill, 1957)***

From the definition of magnetic induction (flux density) B, the force on a charge q due to the motion of the charge through a magnetic field is given by Eq. (1), where the force F is at right angles to a

$$F = Bqv \sin\theta \tag{1}$$

plane determined by the direction of the field, and the component $v \sin\theta$ of the velocity is perpendicular to the field. When B is in webers/m^2, q is in coulombs, and v is in m/s, the force is in newtons.

The electric field intensity E due to this force is given in magnitude and direction by the force per unit positive charge. The electric field intensity is equal to the negative of the potential gradient along the rod. In motional electromotive force (emf), the charge being considered is negative. Thus, Eqs. (2)

$$E = \frac{F}{-q} = -Bv \sin\theta = -\frac{\mathscr{E}}{l}$$
$$\mathscr{E} = Blv \sin\theta \tag{2}$$

hold. Here l is the length of the conductor in a direction perpendicular to the field, and $v \sin\theta$ is the component of the velocity that is perpendicular to the field. If B is in webers/m^2, l is in meters, and v is in m/s, the emf $\mathscr{E}$ is in volts.

This emf exists in the conductor as it moves through the field whether or not there is a closed circuit. A current would not be set up unless there were a closed circuit, and then only if the rest of the circuit does not move through the field in exactly the same manner as the rod. For example, if the rod slides along stationary tracks that are connected together, there will be a current in the closed circuit. However, if the two ends of the rod were connected by a wire that moved through the field with the rod, there would be an emf induced in the wire that would be equal to that in the rod and opposite in sense in the circuit. Therefore, the net emf in the circuit would be zero, and there would be no current.

EMF due to change of flux. When a coil is in a magnetic field, there will be a flux Φ threading the coil the magnitude of which will depend upon the area of the coil and its orientation in the field. The flux is given by $\Phi = BA \cos\theta$, where A is the area of the coil and θ is the angle between the normal to the plane of the coil and the magnetic field. Whenever there is a change in the flux threading the coil, there will be an induced emf in the coil while the change is taking place. The change in flux may be caused by

a change in the magnetic induction of the field or by a motion of the coil. The magnitude of the induced emf, Eq. (3), depends upon the number of turns of

$$\mathscr{E} = -N\frac{d\Phi}{dt} \tag{3}$$

the coil N and upon the rate of change of flux. The negative sign in Eq. (3) refers to the direction of the emf in the coil; that is, it is always in such a direction as to oppose the change that causes it, as required by Lenz's law. If the change is an increase in flux, the emf would be in a direction to oppose the increase by causing a flux in a direction opposite to that of the increasing flux; if the flux is decreasing, the emf is in such a direction as to oppose the decrease, that is, to produce a flux that is in the same direction as the decreasing flux. *See* FARADAY'S LAW OF INDUCTION; LENZ'S LAW.

Consider the case of a flat coil of area A rotating with uniform angular velocity ω about an axis perpendicular to a uniform magnetic field of flux density B. For any position of the coil, the flux threading the coil is $\Phi = BA\cos\theta = BA\cos\omega t$, where the zero of time is taken when θ is zero and the normal to the plane of the coil is parallel to the field. Then the emf induced as the coil rotates is given by Eq. (4).

$$\mathscr{E} = -N\frac{d\Phi}{dt} = -NBA\frac{d(\cos\theta)}{dt} = NBA\omega\ \sin\omega t \tag{4}$$

The induced emf is sinusoidal, varying from zero when the plane of the coil is perpendicular to the field to a maximum value when the plane of the coil is parallel to the field.

Self-induction. If the flux threading a coil is produced by a current in the coil, any change in that current will cause a change in flux, and thus there will be an induced emf while the current is changing. This process is called self-induction. The emf of self-induction is proportional to the rate of change of current. The ratio of the emf of induction to the rate of change of current in the coil is called the self-inductance of the coil.

Mutual induction. The process by which an emf is induced in one circuit by a change of current in a neighboring circuit is called mutual induction. Flux produced by a current in a circuit A (**Fig. 2**) threads or links circuit B. When there is a change of current in circuit A, there is a change in the flux linking coil B, and an emf is induced in circuit B while the change is taking place. Transformers operate on the principle of mutual induction. *See* TRANSFORMER.

The mutual inductance of two circuits is defined as the ratio of the emf induced in one circuit B to the rate of change of current in the other circuit A. For a detailed discussion of self- and mutual inductance *See* INDUCTANCE.

Coupling coefficient. This refers to the fraction of the flux of one circuit that threads the second circuit. If two coils A and B having turns N_A and N_B, respectively, are so related that all the flux of either threads both coils, the respective self-inductances are given by Eqs. (5) and the mutual inductance of the pair is

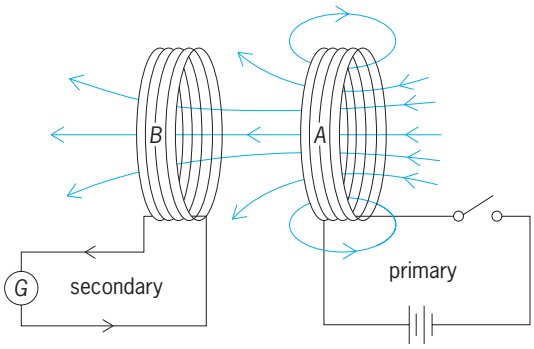


Fig. 2. Mutual induction. An emf is induced in the secondary when the current changes in the primary. (*After R. L. Weber, M. W. White, and K. V. Manning, Physics for Science and Engineering, McGraw-Hill, 1957*)

given by Eq. (6). Then Eqs. (7) hold.

$$L_A = \frac{N_A\Phi_A}{I_A} \qquad L_B = \frac{N_B\Phi_B}{I_B} \tag{5}$$

$$M = \frac{N_A\Phi_B}{I_B} = \frac{N_B\Phi_A}{I_A} \tag{6}$$

$$M^2 = \frac{N_AN_B\Phi_A\Phi_B}{I_AI_B} = \frac{N_A\Phi_A}{I_A}\frac{N_B\Phi_B}{I_B} = L_AL_B \qquad M = \sqrt{L_AL_B} \tag{7}$$

In general, not all the flux from one circuit threads the second. The fraction of the flux from circuit A that threads circuit B depends upon the distance between the two circuits, their orientation with respect to each other, and the presence of a ferromagnetic material in the neighborhood, either as a core or as a shield. It follows that for the general case that Eq. (8) holds.

$$M \le \sqrt{L_AL_B} \tag{8}$$

The ratio of the mutual inductance of the pair to the square root of the product of the individual self-inductances is called the coefficient of coupling K, given by Eq. (9). The coupling coefficient has a maximum value of unity if all the flux threads both circuits, zero if none of the flux from one circuit threads the other. For all conditions, K has a value between 0 and 1.

$$K = \frac{M}{\sqrt{L_AL_B}} \tag{9}$$

Applications. The phenomenon of electromagnetic induction has a great many important applications in modern technology. For example, *see* COUPLED CIRCUITS; GENERATOR; INDUCTION HEATING; MICROPHONE; MOTOR; SERVOMECHANISM.

Kenneth V. Manning

Bibliography. W. J. Duffun, *Electricity and Magnetism*, 4th ed., 1990; D. Halliday and R. Resnick, *Physics*, pt. 2, 4th ed., 1992; A. Mottershead, *Introduction to Electricity and Electronics*, 2000; E. M. Purcell, *Electricity and Magnetism*, 2d ed., 1984.

Electromagnetic pulse (EMP)

A transient electromagnetic signal produced by a nuclear explosion in or above the Earth's atmosphere. Though not considered dangerous to people, the electromagnetic pulse (EMP) is a potential threat to many electronic systems.

Discovery. The existence of a nuclear-generated EMP has been known for many years. Originally predicted by scientists involved with the early development of nuclear weapons, it was not considered to be a serious threat to people or equipment. Then in the early 1960s some of the high-altitude nuclear tests conducted in the Pacific led to some strange occurrences many miles from ground zero. In Hawaii, for example, some 800 mi (1300 km) from the Johnston Island test, EMP was credited with setting off burglar alarms and turning off street lights. In later tests that were conducted in Nevada, significant EMP-induced signals were coupled to cables. These experimental results gave credibility to the potential military use of EMP and encouraged investigators to more accurately describe its origin, electromagnetic characteristics, and coupling to systems, and to develop an affordable method for system protection.

Initial nuclear radiation. In a typical nuclear detonation, parts of the shell casing and other materials are rapidly reduced to a very hot, compressed gas, which upon expansion gives rise to enormous amounts of mechanical and thermal energy. At the same time the nuclear reactions release tremendous amounts of energy as initial nuclear radiation (INR). This INR is in the form of rapidly moving neutrons and high-energy electromagnetic radiation, called x-rays and prompt gamma rays. About a minute after detonation, the radioactive decay of the fission products gives rise to additional gamma rays and electrons (or beta particles), known as residual nuclear radiation (RNR). The distribution of the total explosive energy of a hypothetical fission detonation in the atmosphere below an altitude of 6 mi (10 km) is 50% blast, 35% thermal, 10% RNR, 5% INR. At higher altitudes where the air is less dense, the thermal energy increases and the blast energy decreases proportionately. *See* BETA PARTICLES; GAMMA RAYS; NUCLEAR FISSION; NUCLEAR REACTION; RADIOACTIVITY.

EMP is associated with the INR output, which is a small percentage of the total explosive energy. Nevertheless, EMP is still capable of transferring something of the order of 0.1–0.9 joule/m^2 (0.007–0.06 ft-lbf/ft^2) onto a collector, more than enough to cause upset or damage to normal semiconductor devices.

Early research first developed the physics of high-altitude EMP. Then in 1978 a consistent explanation was set forth of how EMP is related to the total gamma source strength regardless of the detonation height above ground. **Figure 1** identifies six contributions to the gamma source strength for a hypothetical 1-megaton surface burst. As the detonation height is elevated, gamma source contributions from neutrons interacting with the ground and air decrease; for a high-altitude detonation (higher than

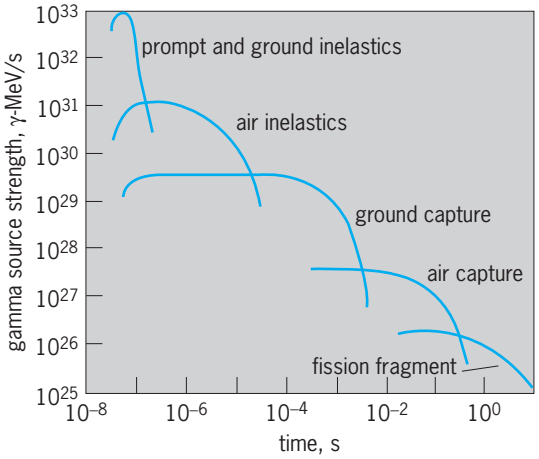


Fig. 1. Total gamma source strength versus time for nominal 1-megaton surface burst. Both horizontal and vertical scales are logarithmic. (*After C. L. Longmire, On the electromagnetic pulse produced by nuclear explosions, IEEE Trans. Antennas Propag., AP-26(1):3–13, 1978*)

37 mi or 60 km) the gamma source strength becomes essentially the prompt gammas from the nuclear burst. The gamma source strength wave shape then approaches a smooth curve, approximating a double exponential, and since EMP is generated from the gamma source strength as described below, it too approaches a double exponential wave shape.

EMP generation in a high-altitude burst. As the prompt gammas move away from a high-altitude nuclear detonation (**Fig. 2**), those gamma rays moving toward the Earth penetrate a more dense region of the atmosphere called the source or deposition region. In this 6-mi thick (10-km) region, approximately 15–21 mi (25–35 km) above the Earth, the highly energetic gamma rays interact with the air molecules to form Compton electrons (with energies starting at 1 MeV) and less energetic gamma rays, which then proceed in the same general direction as the original gamma rays. The fast Compton electrons eventually slow down by stripping other electrons from air molecules to form secondary electron-ion pairs. (Though these secondary electrons and ions

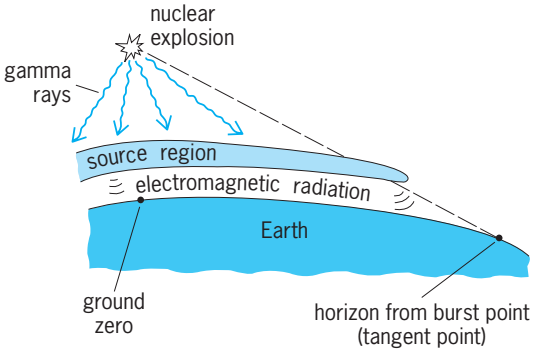


Fig. 2. Schematic representation of the EMP in a high-altitude burst. The extent of the source region varies with the altitude and the yield of the explosion. (*After S. Glasstone and P. J. Dolan, eds., The Effects of Nuclear Weapons, U.S. Department of Defense and the Energy Research and Development Administration, 3d ed., 1977*)

do not contribute to the generation of the EMP, they do cause the region to become highly conductive, and therefore play an important role in determining the EMP wave shape and amplitude.) While slowing down the very intense, short-duration flux of Compton electrons is also deflected by the Earth's geomagnetic field, according to Eq. (1), where the

$$\vec{F} = -q(\vec{v} \times \vec{B}) \qquad (1)$$

deflection force $\vec{F}$ is perpendicular to the geomagnetic field $\vec{B}$ and the velocity $\vec{v}$ of a Compton electron of charge q. The Compton electrons then spiral about the geomagnetic lines, radiating electromagnetic energy in the form of EMP until they eventually recombine with local, positively charged ions. *See* COMPTON EFFECT; SYNCHROTRON RADIATION.

Characteristics for a high-altitude burst. For a high-altitude nuclear detonation, the radiated EMP observed at large distances from the source region can be represented in time t as an electric field $E(t)$ and a magnetic field $H(t)$, given by Eqs. (2) and (3), with

$$E(t) = E_0(e^{-\alpha t} - e^{-\beta t}) \qquad t \geq 0 \qquad (2)$$

$$H(t) = H_0(e^{-\alpha t} - e^{-\beta t}) \qquad t \geq 0 \qquad (3)$$

$E_0 = 5.2 \times 10^4$ V/m (1.6×10^4 V/ft), $H_0 = 1.4 \times 10^2$ A/m (4.2×10 A/ft), $\alpha = 4.0 \times 10^6$ s^{-1}, and $\beta = 4.76 \times 10^8$ s^{-1}. (In air, E and H are related by the impedance of free space, 377 ohms.) If the observer is directly below the detonation (ground zero), the polarization of both fields is predominantly horizontal; if the observer is at the horizon, the fields can have both horizontal and vertical components. The magnitude of these field components varies according to the latitude and longitude of the observer and the direction of the burst. *See* POLARIZATION OF WAVES.

Surface-burst EMP. Should the nuclear detonation occur closer to the Earth, the EMP generation process becomes far more complex and the electric and magnetic fields become very complicated. The most dramatic change occurs with a surface burst (**Fig. 3**).

When the observer is somewhere within the source region of a near-surface burst, where the air conductivity varies between 10^{-4} and 10^{-2} siemen/m (3×10^{-5} and 3×10^{-3} siemen/ft), the resultant electric field is predominantly vertical, and the resultant magnetic field is polarized perpendicular to the plane of the figure. The electric field polarization is due to the Compton-electron and ion-charge-separation fields which tend to become perpendicular to the conducting earth, leaving a resultant vertical electric field near the ground. As these Compton electrons move radially away from the detonation, they curve earthward and return to the detonation point through the conducting earth. This Compton current loop then gives rise to a resultant magnetic field. When the observer is far from the detonation (outside the source region), the electric and magnetic fields begin to approximate fields from a vertical dipole, decaying with distance r as $1/r$.

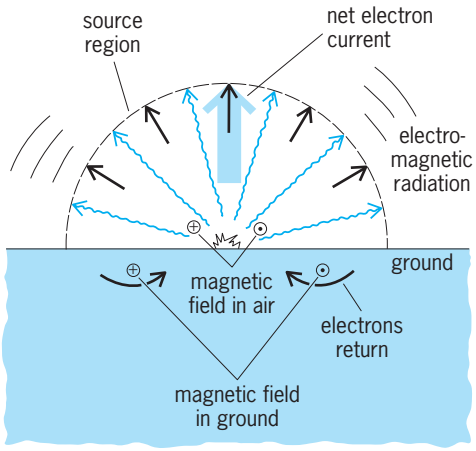


Fig. 3. Schematic representation of the EMP in a surface burst. (*After S. Glasstone and P. J. Dolan, eds., The Effects of Nuclear Weapons, U.S. Department of Defense and the Energy Research and Development Administration, 3d ed., 1977*)

Internal EMP. It is also possible for INR (both x-rays and gamma rays) to directly interact with systems, causing EMP signals internal to structures. This phenomenon has been called internal or system-generated EMP and is potentially a serious problem for satellites in orbit and for electronics in metallic enclosures on or near the ground. These forms of EMP are generated by x-rays interacting with satellites and gamma rays impinging on ground-based enclosures, producing currents of Compton electrons internally that then produce internal electromagnetic waves. They are very dependent upon the nuclear detonation, the system topology, and the relative position of one to the other.

Coupling. The coupling of these rather wide-band (10^4 to 10^8 Hz) signals to systems of different topologies can be significant. For example, it is not unusual to predict voltage and current levels of hundreds of thousands of volts and a few thousand amperes coupled by high-altitude EMP onto extended systems. The exact level of coupling depends upon the size of the system and its orientation with respect to the incident field, and upon whether or not it is near an earth ground. For ground-based systems the incident field components can actually be enhanced or degraded, depending upon the polarization. Solving Maxwell's equations at the interface between an ideal conductor and a dielectric shows that the net horizontal electric field is zero at the interface and the net vertical electric field doubles. (For the magnetic field components the opposite is true.) Since the Earth ground is not a perfect conductor, the field components do not completely cancel or quite double in amplitude. *See* MAXWELL'S EQUATIONS.

An estimate of about 1 joule (0.7 ft-lbf) of EMP-coupled energy is considered reasonable for many systems. Even if the coupling onto circuits is inefficient, as little as 10^{-13} J can upset some semiconductor devices and 10^{-6} J can cause damage. The potential for such upset and damage in critical electronic circuits has led to the incorporation of EMP protection in many system designs. This protection

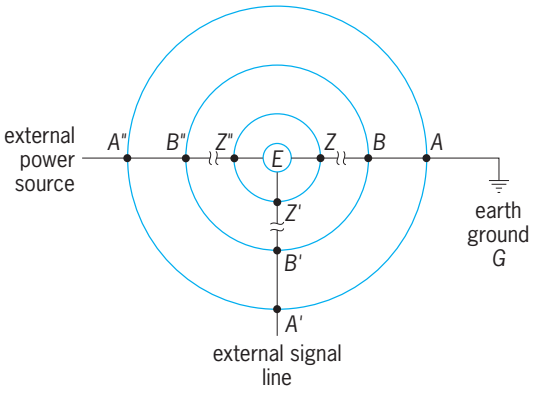


Fig. 4. Typical system protection scheme.

is most prevalent in communications systems whose disruption by EMP is considered an important civil and military vulnerability. *See* ELECTROMAGNETIC RADIATION; NUCLEAR EXPLOSION.

Protection. The most common form of protection incorporated in system designs is a combination of shielding and penetration control. **Figure 4** illustrates a protection scheme in which a system's electronics *E* is isolated from the external environment by one or more nested, shielded enclosures (often called Faraday cages). Penetration control is then maintained by minimizing the number of shield penetrations (in this case, a power line, a signal line, and a ground wire connecting *E* to earth ground *G*) and by applying terminal protection devices, such as spark gaps, high-power Zener diodes, or metal oxide varistors, at selected shield penetration points (A, A', A'', B, B', and B''; or Z, Z', and Z''; or both). In this way, system protection can be designed not only for EMP but also for other electromagnetic transients (such as near-strike lightning and electromagnetic interference). Furthermore, cost-effective, field maintainable protection can be achieved by properly selecting off-the-shelf shielding techniques and terminal protection devices and applying them to systems in accordance with such military standards as MIL-STD 461. *See* ELECTRIC PROTECTIVE DEVICES; ELECTRICAL INTERFERENCE; ELECTRICAL SHIELDING; ELECTROMAGNETIC RADIATION; LIGHTNING AND SURGE PROTECTION; NUCLEAR EXPLOSION. Robert A. Pfeffer

Bibliography. S. Glasstone and P. J. Dolan (eds.), *The Effects of Nuclear Weapons*, U.S. Department of Defense and the Energy Research and Development Administration, 3d ed., 1977, reprint 1983; K. S. Lee (ed.), *EMP Interaction: Principles, Techniques and Reference Data*, rev. ed, 1986; J. R. Pierce et al., *Evaluation of Methodologies for Estimating Vulnerability to Electromagnetic Pulse Effects,* National Research Council Rep., 1984.

Electromagnetic pump

A pump that operates on the principle that a force is exerted on a current-carrying conductor in a magnetic field. The high electrical conductivity of the liquid metals pumped (liquid metals are used as the heat-transfer media in some nuclear reactors and magnetohydrodynamic systems) allows a pumping force to be developed within the metals when they are confined in a duct or channel and subjected to a magnetic field and to an electric current. These pumps are designed principally for use in liquid-metal-cooled reactor plants where liquid lithium, sodium, potassium, or sodium-potassium alloys are pumped. Other metallic and nonmetallic liquids of sufficiently high electrical conductivity, such as mercury or molten aluminum, lead, and bismuth, may also be pumped in nonnuclear applications. The absence of moving parts within the pumped liquid eliminates the need for seals and bearings that are found in conventional mechanical pumps, thus minimizing leaks, maintenance, and repairs, and improving reliability. In liquid-metal-cooled nuclear reactor plants, electromagnetic pumps with a capacity of up to several thousand gallons per minute have operated without maintenance for decades. *See* MAGNETOHYDRODYNAMICS; NUCLEAR POWER; NUCLEAR REACTOR.

The major classifications of electromagnetic pumps, either conduction or induction, is based on the method employed to cause the electric current to flow in the pumped liquid.

Direct-current conduction pumps. In this simplest form of electromagnetic pump the liquid-containing pump duct has two heavy electrodes to provide a path for direct current to be conducted at a right angle to the direction of pumping force in the duct. A magnetic field generated by either a permanent magnet or electromagnet is impressed across the duct at the location of the electrodes in a direction at right angles to both the current flow and the pumping direction. This orthogonal arrangement of current, magnetic field, and pump duct produces a force in the liquid which is dependent on the magnitude of the current, magnetic field, and the length of the current path in the liquid. It is also possible to arrange the duct so that the current flowing in the liquid also produces the required magnetic field (that is, the self-induced pump). The physical arrangement of the conduction pump allows only a single passage of current between electrodes, resulting in a relatively short length of current path in the liquid. To compensate, the current must be very high (thousands of amperes), thus requiring special dc power supplies such as homopolar generators. However, the very low voltage allows special electrical insulation to be used, which is suitable for high-temperature (1500°F or 800°C) operation.

In addition to the externally energized dc conduction pump described above, there are two other unique pump design concepts of the dc conduction pump family in use or under development. In the thermoelectric electromagnetic pump, the current in the duct is generated by thermoelectric elements attached to the duct electrodes which are driven by the temperature difference between the duct and a heat sink. The major application of thermoelectric electromagnetic pumps is in the

liquid-metal loops of space reactors. *See* THERMOELECTRICITY.

The second design concept is the flow coupler. In this design a dc conduction pump duct in a secondary (driven) loop is directly coupled to a magnetohydrodynamic generator located in an adjacent parallel duct of a primary (driving) loop to which it is connected by appropriate buswork. Both ducts lie in a magnetic field produced by a permanent magnet or electromagnet. The initial use of this concept was to drive small auxiliary loops in experimental liquid-metal facilities. However, with the advent of liquid-metal pool reactors the concept is being studied as a means of circulating the primary coolant in the pool.

Alternating-current conduction pump. This pump uses a pump duct, magnetic field, and electrode arrangement similar to the dc conduction pump, except that the magnetic field is produced by an ac (single-phase) electromagnet and the current is produced by the transformer action of the same or separate single-phase electromagnet with a secondary (step-down) winding of usually one turn. The heavy secondary winding connects to the electrodes. With the sinusoidal magnetic field and duct current in phase with each other, a pumping force will be developed. This design can be operated directly from normal single-phase ac power equipment since it has an integral transformer to develop the necessary high current. It too can have a high-temperature application. *See* ELECTROMAGNET; TRANSFORMER.

Helical induction pump. This pump closely resembles a polyphase squirrel-cage ac induction motor in principle and operating characteristics. The fluid channel is formed by helical vanes encased in a concentric cylinder and is analogous to the rotor of a conventional motor. Arranged around the periphery of the outer cylinder is a polyphase multipole winding which is similar to the stator of an induction motor. The polyphase winding creates a rotating magnetic field which induces a current in the liquid metal, and this current in turn reacts with the magnetic field to produce a force on the fluid tending to rotate it in the duct. However, the fluid is channeled by the helical vanes, forcing it to produce an axial pressure and resultant axial flow in the duct. The helical induction pump is used in applications requiring higher pressures and lower flows than the linear induction pumps described below. *See* INDUCTION MOTOR.

Linear induction pumps. These pumps use a stationary polyphase ac winding to produce a traveling magnetic field that induces a current in the liquid. The reaction between the induced current and the traveling magnetic field causes the liquid to follow the traveling magnetic field. The force generated in the liquid is confined in the duct, producing a linear flow.

Two principal designs are the flat linear and the annular linear induction pumps. The flat linear induction pump has a rectangular duct connected to inlet and outlet nozzles. Several liquid-metal-cooled reactors use such pumps in their main heat transport systems. The annular induction pump design has a cylindrical duct formed by two concentric tubes.

These pumps are designed for loop application. They are temperature-limited with conventional electrical insulation and in most cases require a separate cooling system for the electrical windings. However, for application to pool- or tank-type liquid-metal-cooled reactors, there is considerable interest in the design and development of large submersible self-cooled (windings cooled by the liquid being pumped) linear induction pumps. Leslie R. Dahl

Bibliography. R. S. Baker and M. J. Tessier, *Handbook of Electromagnetic Pump Technology*, 1987; J. Collett et al., *Thermoelectric Electromagnetic Pump Design for Sp-100, Symposium on Space Nuclear Power Systems*, 1988; *Fast Breeder Reactors—Experience and Trends: Proceedings of a Symposium*, Lyons, 1986; *International Conference on Design and Safety of Advanced Nuclear Power Plants*, Tokyo, 1992; *International Conference on Fast Reactors and Related Fuel Cycles*, Kyoto, 1991; A. E. Walter and A. B. Reynolds, *Fast Breeder Reactors*, 1981.

Electromagnetic radiation

Energy transmitted through space or through a material medium in the form of electromagnetic waves. The term can also refer to the emission and propagation of such energy. Whenever an electric charge oscillates or is accelerated, a disturbance characterized by the existence of electric and magnetic fields propagates outward from it. This disturbance is called an electromagnetic wave. The frequency range of such waves is tremendous, as is shown by the electromagnetic spectrum in the **table**. The sources given are typical, but not mutually exclusive, as is shown by the fact that the atomic interstellar hydrogen radiation whose wavelength is 0.210614 m falls in the radar region. The other monochromatic radiation listed is that from positron-electron annihilation whose wavelength is 2.42626×10^{-12} m.

Detection of radiation. In theory, any electromagnetic radiation can be detected by its heating effect. This method has actually been used over the range from x-rays to radio. Ionization effects measured by cloud chambers, photographic emulsions, ionization chambers, and Geiger counters have been used in the gamma- and x-ray regions. Direct photography can be used from the gamma-ray to the infrared region. Fluorescence is effective in the x-ray and ultraviolet ranges. Bolometers, thermocouplers, and other heat-measuring devices are used chiefly in the infrared and microwave regions. Crystal detectors, vacuum tubes, and transistors cover the microwave and radio frequency ranges.

Free-space waves. A charge in simple harmonic (linear sinusoidal) motion in a vacuum generates a simple wave which becomes spherical at distances from the source much larger than the amplitude of the motion and so great that many oscillations have occurred before the disturbance arrives. The wave is plane when the dimensions of the area observed are very small compared with the radius of spherical

Electromagnetic spectrum

Frequency, Hz	Wavelength, m	Nomenclature	Typical source
10^{23}	3×10^{-15}	Cosmic photons	Astronomical
10^{22}	3×10^{-14}	γ-rays	Radioactive nuclei
10^{21}	3×10^{-13}	γ-rays, x-rays	
10^{20}	3×10^{-12}	X-rays	Atomic inner shell, positron-electron annihilation
10^{19}	3×10^{-11}	Soft x-rays	Electron impact on a solid
10^{18}	3×10^{-10}	Ultraviolet, x-rays	Atoms in sparks
10^{17}	3×10^{-9}	Ultraviolet	Atoms in sparks and arcs
10^{16}	3×10^{-8}	Ultraviolet	Atoms in sparks and arcs
10^{15}	3×10^{-7}	Visible spectrum	Atoms, hot bodies, molecules
10^{14}	3×10^{-6}	Infrared	Hot bodies, molecules
10^{13}	3×10^{-5}	Infrared	Hot bodies, molecules
10^{12}	3×10^{-4}	Far-infrared	Hot bodies, molecules
10^{11}	3×10^{-3}	Microwaves	Electronic devices
10^{10}	3×10^{-2}	Microwaves, radar	Electronic devices
10^{9}	3×10^{-1}	Radar	Electronic devices, interstellar hydrogen
10^{8}	3	Television, FM radio	Electronic devices
10^{7}	30	Short-wave radio	Electronic devices
10^{6}	300	AM radio	Electronic devices
10^{5}	3000	Long-wave radio	Electronic devices
10^{4}	3×10^{4}	Induction heating	Electronic devices
10^{3}	3×10^{5}		Electronic devices
100	3×10^{6}	Power	Rotating machinery
10	3×10^{7}	Power	Rotating machinery
1	3×10^{8}		Commutated direct current
0	Infinity	Direct current	Batteries

curvature. In this case the choice of the rectangular coordinates x and z as the directions of the oscillation and of the observation or field point, respectively, permits the electric intensity **E** and the magnetic flux density **B** to be written as Eq. (1). The field

$$E_x = vB_y = E_0 \cos\,[\omega(t - v^{-1}z)] \qquad (1)$$

amplitude E_0 is constant over the specified area and not dependent on z if the z range is small compared with the source distance, as in stellar radiation. The angular frequency of the source is ω radians per second, which is the frequency ν in hertz multiplied by 2π. The velocity of the wave is v, the direction of propagation z, and the time t. The wavelength λ is $2\pi v/\omega$. If t is in seconds and z is in meters, then v is in meters per second and λ is in meters. It is found that in a lossless, isotropic, homogeneous medium Eq. (2) holds; here μ is the permeability, and ϵ the

$$v = (\mu\epsilon)^{-1/2} \qquad (2)$$

capacitivity, or dielectric constant. This wave is transverse because **E** and **B** are normal to z. It is plane-polarized because E_x and B_y are parallel to fixed axes. The plane of polarization is taken as that defined by the electric vector and the direction of propagation.

Plane waves. An electromagnetic disturbance is a plane wave when the instantaneous values of any field element such as **E** and **B** are constant in phase over any plane parallel to a fixed plane. These planes are called wavefronts. In empty unbounded space, **E** and **B** lie in the wavefront normal to each other; if the wave is unpolarized, their direction fluctuates in this plane in random fashion. If the plane waves are bounded, as on transmission lines and in waveguides, the amplitudes may vary over the wavefront, and in the case of waveguides and crystals some of the elements will not in general lie in the wavefront. The equation for an undamped plane wave whose front is normal to z is Eq. (3), where F is one of the

$$F = \Phi_1(x, y)f_1(z - vt) + \Phi_2(x, y)f_2(z + vt) \qquad (3)$$

field elements such as **E** or **B**. Note that if an observer sees a certain value of $\Phi_1(x,y)$ at z and then jumps instantaneously in the z direction to a point $z + \Delta z$, the observer will, after waiting a time $\Delta z/v$, see the same value $\Phi_1(x,y)$ because Eq. (4) is valid.

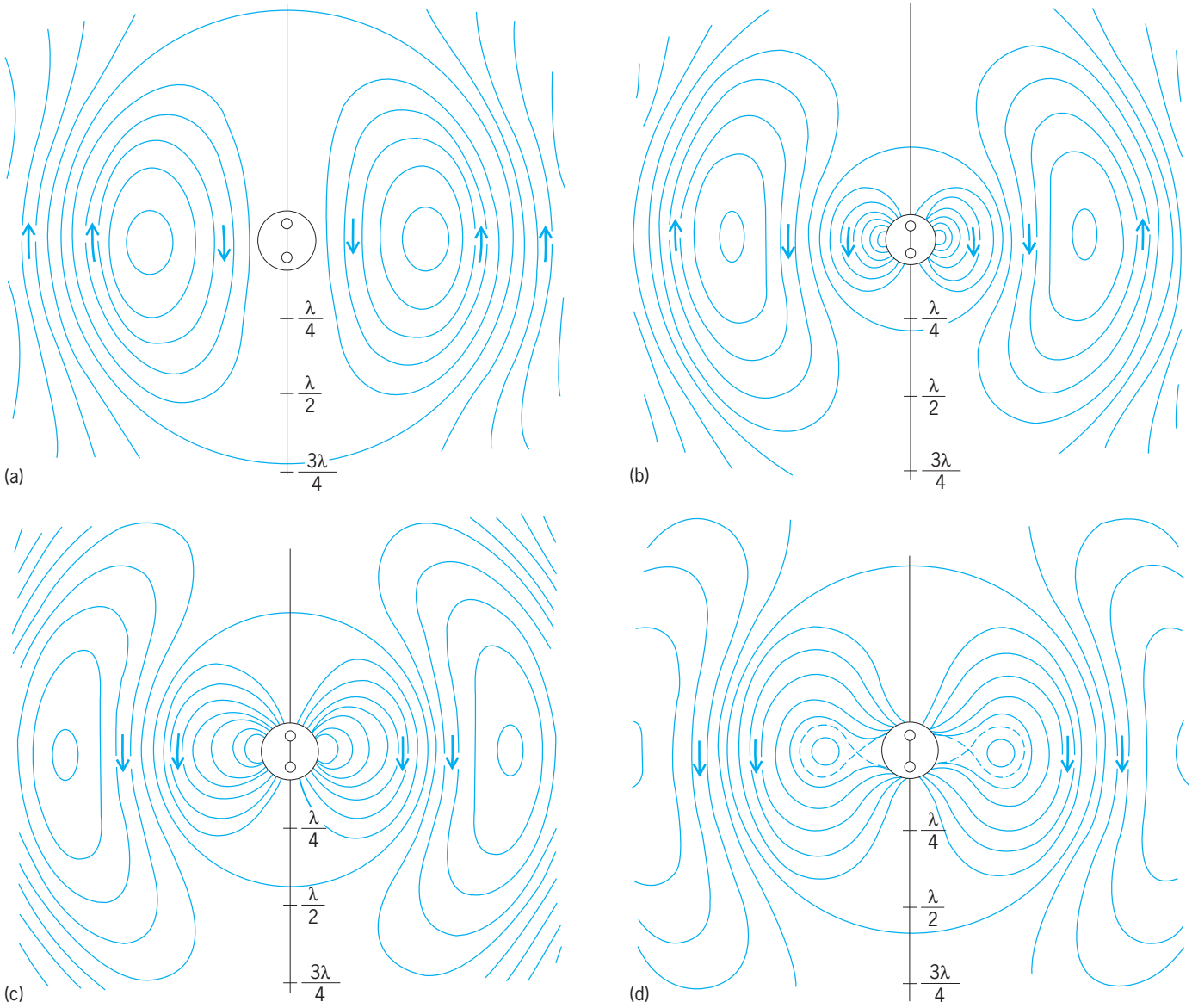
$$f(z - vt) = f[z + \Delta z - v(t + \Delta z/v)] \qquad (4)$$

Thus, the first term represents a wave moving in the z direction with a velocity v. The second term represents a wave in the negative z direction. The form of $\Phi_1(x,y)$ and $\Phi_2(x,y)$ depends on the boundary conditions. *See* WAVE EQUATION.

Spherical waves. A wave is spherical when the instantaneous value of any field element such as **E** or **B** is constant in phase over a sphere. The radiation from any source of finite dimensions becomes spherical at great distances in an unbounded, isotropic, homogeneous medium. The equation for an undamped spherical wave is Eq. (5). The first term represents a

$$F = r^{-1}\Phi_1(\theta, \varphi)f\,(r - vt) + r^{-1}\Phi_2(\theta, \varphi)f\,(r + vt) \qquad (5)$$

diverging and the second a converging wave. Again, the form of $\Phi_1(\theta,\varphi)$ and $\Phi_2(\theta,\varphi)$ depends on the nature of the source and other boundary conditions.



Diagrams of electric dipole. The outward moving electric field lines generated by the hertzian oscillator are shown at successive eighth-period intervals. (*a*) $t = 0$; (*b*) $t = T/8$; (*c*) $t = T/4$; (*d*) $t = 3T/8$.

Damped waves. If there are energy losses which are proportional to the square of the amplitude, as in the case of a medium of conductivity γ which obeys Ohm's law, then the wave is exponentially damped, and Eq. (1) becomes Eq. (6). The symbol α is called

$$E_x = E_0 e^{-\alpha z} \cos(\omega t - \beta z) \tag{6}$$

the attenuation constant, and β the wave number or phase constant which equals ω/v', where v' is the damped-wave velocity. The electric wave amplitude at the origin has been taken as E_0. The ratio of E_0 to B_0, as well as that of α to β, depends on the permeability μ, the capacitivity ϵ, and the conductivity γ of the medium. In terms of the phasor $\check{E}_x$, Eq. (6) may be written as the real part of Eq. (7).

$$E_x = \check{E}_x e^{j\omega t} = E_0 e^{-(\alpha + j\beta)z} e^{j\omega t} \tag{7}$$

This is exactly the form for the current on a transmission line. (Phasors are complex numbers of form such that, when multiplied by $e^{j\omega t}$, the real part of the product gives the amplitude, phase, and time dependence.)

Wave impedance. Those trained in transmission line theory find it useful to apply the same techniques to wave theory. Consider an isolated tubular section of the wave in Eq. (1) bounded by $x = 0$, $x = 1$, and $y = 0$, $y = 1$ as a transmission line. The potential across the line between $x = 0$ and $x = 1$ is **E**. The line integral of **B** around the $x = 0$ boundary from $y = 0$ to $y = 1$ is μI by Ampère's law and equals **B** because **B** is zero on the negative side. Thus, the impedance of the line is, making use of Eqs. (1) and (2), given by Eq. (8). This depends only on the properties of

$$\check{Z}_k = \frac{V}{I} = \frac{\mu E}{B} = \frac{E}{H} = \left(\frac{\mu}{\epsilon}\right)^{1/2} = \eta \tag{8}$$

the medium and is known as the wave impedance.

In transmission line theory the ratio μ/ϵ would be replaced by the ratio of the series impedance $\check{Z}_L = j\omega L$ to the shunt admittance $\check{Y} = j\omega C$, where L is the inductance per unit length, and C the capacitance per unit length across the line. If there is a resistance R per unit length across the line, then $1/R$ must be added to Y. This resistance is $1/\gamma$ for the tubular section. Thus, for a conducting medium, Eq. (8) becomes Eq. (9). The last term is a common

$$\check{Z}_k = \left(\frac{j\omega\mu}{\gamma + j\omega\epsilon}\right)^{1/2} = \frac{j\omega\mu}{\alpha + j\beta} \tag{9}$$

transmission line form. The reflection and refraction of plane waves at plane boundaries separating different mediums may be calculated by transmission line formulas with the aid of Eqs. (8) and (9).

Electric dipole. A charge undergoing simple harmonic motion in free space is a dipole source when the amplitude of the motion is small compared with the wavelength. The term is loosely applied to the hertzian oscillator, usually pictured as a dumbbell-shaped conductor in which the electrons oscillate from one end to the other, leaving the opposite end periodically positive. An electric dipole of moment M is defined as the product qa when two large, equal and opposite charges, $+q$ and $-q$, are placed a small distance a apart. A dipole is oscillating when M is periodic in time and is the simplest source of spherical waves. Much can be learned by a study of H. Hertz's picture of the outward moving electric field lines at successive time intervals of one-eighth period in a plane which passes through the hertzian oscillator axis, shown in the **illustration**. The most striking feature of the pictures is that, after breaking loose from the dipole, all electric field lines are closed, which means that the divergence of **E** is zero. This is true of all unbounded waves. It is also noteworthy that the waves become truly spherical with a fixed wavelength λ only in a direction perpendicular to the dipole and at a distance which greatly exceeds the dipole dimensions. This distance is beyond the edges of the picture. Lengths $\lambda/4$, $\lambda/2$, and $3\lambda/4$ are marked off on the axis for comparison. The magnetic field lines are circles coaxial with the oscillator, so they intersect the plane of the diagram normally. They are most dense where the electric lines are closely spaced. The radiant energy emitted by atoms and molecules is essentially radiation of the dipole type. *See* ABSORPTION OF ELECTROMAGNETIC RADIATION; ANTENNA (ELECTROMAGNETISM); DIFFRACTION; ELECTROMAGNETIC WAVE TRANSMISSION; GAMMA RAYS; HEAT RADIATION; INFRARED RADIATION; INTERFERENCE OF WAVES; LIGHT; MAXWELL'S EQUATIONS; MICROWAVE; POLARIZATION OF WAVES; RADIATION; RADIO-WAVE PROPAGATION; REFLECTION OF ELECTROMAGNETIC RADIATION; REFRACTION OF WAVES; SCATTERING OF ELECTROMAGNETIC RADIATION; TRANSMISSION LINES; ULTRAVIOLET RADIATION; WAVEGUIDE; WAVE MOTION; X-RAYS.

William R. Smythe

Bibliography. M. Born and E. Wolf, *Principles of Optics*, 7th ed., 1999; N. A. Dyson, *X-Rays in Atomic and Nuclear Physics*, 2d ed., 1990; B. D. Guenther, *Modern Optics*, 1990; A. Ishimaru, *Electromagnetic Wave Propagation, Radiation, and Scattering*, 1996; M. F. Iskander, *Electromagnetic Fields and Waves*, 2000; J. A. Kong, *Electromagnetic Wave Theory*, 3d ed., 2000; S. Ramo, J. R. Whinnery, and T. Van Duzer, *Fields and Waves in Communications Electronics*, 3d ed., 1993; V. V. Sarwate, *Electromagnetic Fields and Waves*, 1993; W. R. Smythe, *Static and Dynamic Electricity*, 3d ed., 1968, reprint 1989.

Electromagnetic wave

A disturbance, produced by the acceleration or oscillation of an electric charge, which has the characteristic time and spatial relations associated with progressive wave motion. A system of electric and magnetic fields moves outward from a region where electric charges are accelerated, such as an oscillating circuit or the target of an x-ray tube. The wide wavelength range over which such waves are observed is shown by the electromagnetic spectrum. The term electric wave, or hertzian wave, is often applied to electromagnetic waves in the radar and radio range. Electromagnetic waves may be confined in tubes, such as wave guides, or guided by transmission lines. They were predicted by J. C. Maxwell in 1864 and verified experimentally by H. Hertz in 1887. *See* ELECTROMAGNETIC RADIATION; MAXWELL'S EQUATIONS.

William R. Smythe

Electromagnetic wave transmission

The transmission of electrical energy by wires, the broadcasting of radio signals, and the phenomenon of visible light are all examples of the propagation of electromagnetic energy. Electromagnetic energy travels in the form of a wave. Its speed of travel is approximately 186,000 mi/s (3×10^8 m/s) in a vacuum and is somewhat slower than this in liquid and solid insulators. An electromagnetic wave does not penetrate far into an electrical conductor, and a wave that is incident on the surface of a good conductor is largely reflected.

Electromagnetic waves originate from accelerated electric charges. For example, a radio wave originates from the oscillatory acceleration of electrons in the transmitting antenna. The light that is produced within a laser originates when electrons fall from a higher energy level to a lower one. *See* LASER.

The waves emitted from a source are oscillatory in character and are described in terms of their frequency of oscillation. Local telephone lines (not using carrier systems) carry electromagnetic waves with frequencies of about 200–4000 Hz. Medium-wave radio uses frequencies of the order of 10^6 Hz, radar uses frequencies of the order of 10^{10} Hz, and a ruby laser emits light with a frequency of

4.32×10^{14} Hz. The method of generating an electromagnetic wave depends on the frequency used, as do the techniques of transmitting the energy to another location and of utilizing it when it has been received. *See* RADAR.

The communication of information to a distant point is generally accomplished through the use of electromagnetic energy as a carrier. A familiar example is the telephone, in which sound waves in the range of frequencies from a few hundred to a few thousand hertz are converted into corresponding electromagnetic waves, which are then guided to their destination by a pair of wires. Another familiar example is radio, in which the signals are caused to modify an identifiable characteristic, such as the amplitude or frequency, of an electromagnetic carrier wave. The electromagnetic wave, thus modified, or modulated, is radiated from an antenna and can be received over a considerable region. *See* MODULATION; RADIO BROADCASTING.

Features of electromagnetic waves. **Figure 1** illustrates schematically some of the essential features of an electromagnetic wave. Shown in Fig. 1 are vectors that represent the electric field intensity E and the magnetic field intensity H at various points along a straight line taken in the direction of propagation of the wave. The electric field is in a vertical plane and the wave is said to be vertically polarized. The magnitude of the field, at a given instant, varies as a sinusoidal function of distance along the direction of propagation. The magnetic field intensity H lies in a plane normal to that of E and, at each point, is proportional in magnitude to E, as shown in Eq. (1),

$$\frac{E}{H} = \sqrt{\frac{\mu}{\epsilon}} \tag{1}$$

where H is the magnetic field intensity in amperes/m, E is the electric field intensity in volts/m, ϵ is the permittivity, or absolute dielectric constant, of the medium, and μ is the absolute permeability of the medium. For a vacuum, $\epsilon = 8.854 \times 10^{-12}$ farad/m and $\mu = 4\pi \times 10^{-7}$ henry/m; therefore for a vacuum the ratio E/H is approximately 377 ohms. This ratio is termed the wave impedance of the medium. *See* MAXWELL'S EQUATIONS; WAVE EQUATION.

The E and H waves travel along a straight line, as suggested in Fig. 1. Of the two possible directions along this line, the actual direction of travel can be determined by imagining a screw with a right-hand thread placed along the axis and turned from E toward H; then the longitudinal direction of travel of the screw is the direction of propagation of the energy.

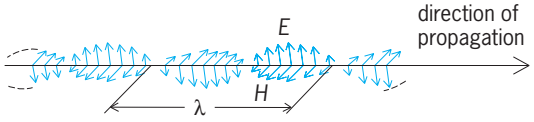


Fig. 1. Representation of an electromagnetic wave at a particular instant of time.

The velocity of travel of the wave is shown in Eq. (2). In a vacuum this is approximately $3 \times$

$$v = 1\sqrt{\mu\epsilon} \tag{2}$$

10^8 m/s. The velocity in air is only slightly smaller.

The wavelength is the distance between two successive similar points on the wave, measured along the direction of propagation. The wavelength is denoted by λ in Fig. 1.

As the wave travels past a stationary point, the values of E and H at the point vary sinusoidally with time. The time required for one cycle of this variation is termed the period, T seconds. The number of hertz is the frequency f; and $f = 1/T$. In one cycle the wave, traveling at the velocity v, moves one wavelength along the axis of propagation. Therefore, $\lambda = vT$, or may be calculated by Eq. (3). Assuming a velocity

$$\lambda = \frac{v}{f} \tag{3}$$

of 3×10^8 m/s, an electromagnetic wave having a frequency of 60 Hz has a wavelength of 5×10^6 m, or approximately 3100 mi. At a frequency of 3 MHz (3×10^6 Hz), λ is 100 m, and at 3000 MHz, λ is 10 cm. Visible light has a frequency of the order of 5×10^{14} Hz and a wavelength of approximately 6×10^{-5} cm.

The density of energy in an electric field is $\epsilon E^2/2$ joules/m^3, and that in a magnetic field is $\mu H^2/2$ J/m^3. With the aid of Eq. (1), the relationships become those shown in Eq. (4). Therefore the

$$\frac{\mu H^2}{2} = \frac{\mu(\sqrt{\epsilon/\mu}E)^2}{2} = \frac{\epsilon E^2}{2} \tag{4}$$

electric and magnetic fields carry equal energies in the electromagnetic wave. The total energy density at any point is equal to ϵE^2 J/m^3. Since this is transported with a velocity equal to $1/\sqrt{\mu\epsilon}$, the rate of flow of energy per square meter normal to the direction of propagation is $\epsilon E^2 v$ or $E^2\sqrt{\epsilon/\mu}$ watts/m^2. In radio broadcasting a field of 50 millivolts/m is considered to be strong. An electromagnetic wave with this intensity has an average energy density of 2.2×10^{-14} J/m^3, and the average rate of energy flow is 6.6×10^{-6} W/m^2.

Radiation from an antenna. **Figure 2** illustrates the configuration of the electric and magnetic fields about a short vertical antenna in which flows a sinusoidal current of the form $i = I_{max} \sin 2\pi ft$ amperes. The picture applies either to an antenna in free space (in which case the illustration shows only the upper half of the fields), or to an antenna projecting above the surface of a highly conducting plane surface. In the latter case the conducting plane represents to a first approximation the surface of the Earth. The fields have symmetry about the axis through the antenna. For pictorial simplicity only selected portions of the fields are shown in Fig. 2. The magnetic field is circular about the antenna, is perpendicular at every point to the direction of the electric field, and is proportional in intensity to the magnitude of the electric field, as indicated by Eq. (1). All parts of the wave

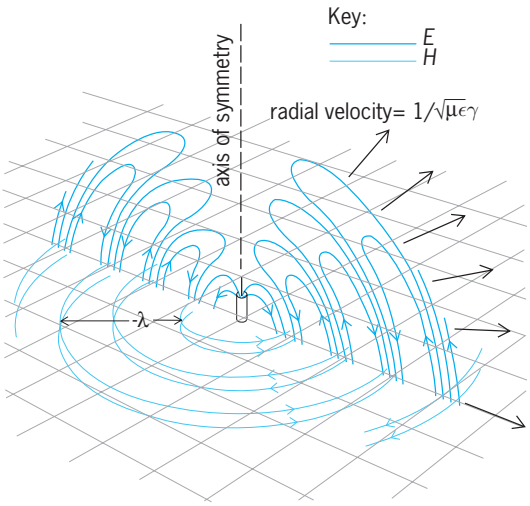


Fig. 2. Configuration of electric and magnetic fields about a short vertical antenna.

travel radially outward from the antenna with the velocity given by Eq. (2); the wave is described as spherical, with the antenna located at the center of the wave. The wavelength of the radiation is given by Eq. (3).

If a short antenna that projects above a highly conducting plane surface carries a current of $i = I_{\max} \sin 2\pi ft$ that is uniform throughout the length of the antenna, the intensity of the electric field in the radiated wave is that shown in Eq. (5). where l is the

$$E_{\max} = \sqrt{\frac{\mu}{\epsilon}} \frac{I_{\max}}{r} \frac{l}{\lambda} \cos\theta \qquad \text{V/m} \qquad (5)$$

length of the antenna, r is the radial distance from the antenna, and θ is 0, and varies inversely with the distance from the antenna.

If the rate of flow of energy per unit area $E^2\sqrt{\epsilon/\mu}$ is integrated over an imaginary spherical surface about the antenna, the average power radiated is that shown by Eq. (6). The factor $(l/\lambda)^2$ is of par-

$$P_{\text{av}} = \frac{2\pi}{3} \sqrt{\frac{\mu}{\epsilon}} I_{\max}^2 \left(\frac{l}{\lambda}\right)^2 \qquad \text{W} \qquad (6)$$

ticular importance, for it indicates that a longer antenna is required at the longer wavelengths (lower frequencies). The radiation of appreciable energy at a very low frequency requires an impracticably long antenna.

The foregoing relations assume a uniform current throughout the length of the antenna. An approximation to this can be achieved in practice by connecting a long horizontal conductor to the top of the antenna. Where some such construction is not utilized, the current will not be uniform in the antenna and will, in fact, be zero at the tip. The results given by Eqs. (4) and (5) must be modified, but the qualitative features of the radiation remain as shown in Fig. 2. *See* ANTENNA (ELECTROMAGNETISM).

Often it is desired to concentrate the radiated energy into a narrow beam. This can be done either by the addition of more antenna elements or by placing a large reflector, generally parabolic in shape, behind the antenna. The production of a narrow beam requires an antenna array, or alternatively a reflector, that is large in width and height compared with a wavelength. The very narrow and concentrated beam that can be achieved by a laser is made possible by the extremely short wavelength of the radiation as compared with the cross-sectional dimensions of the radiating system.

Propagation over the Earth. The foregoing discussion shows some of the important features of the radiation of electromagnetic energy from an antenna, but is oversimplified insofar as communication to and from positions on or near the Earth is concerned. The ground is a reasonably good, but not perfect, conductor; hence, the actual propagation over the surface of the Earth will show a more rapid decrease of field strength than that indicated by the factor of $1/r$ in Eq. (5). Irregularities and obstructions may interfere. In long-range transmission the spherical shape of the Earth is important. Inhomogeneities in the atmosphere refract the wave somewhat. For long-range transmission, the ionized region high in the atmosphere known as the Kennelly-Heaviside layer, or ionosphere, can act as a reflector. The electric field of the wave produces oscillation of the charged particles of the region, and this causes the refractive index of the layer to be smaller than that of the atmosphere below. The result is that, if the angle of incidence is not too near the normal and if the frequency of the wave is not too high, the wave may be refracted back toward the Earth. Successive reflections between ionosphere and Earth can provide communication for long distances around the periphery of the Earth. *See* RADIO-WAVE PROPAGATION.

Hollow waveguides. When an electromagnetic wave is introduced into the interior of a hollow metallic pipe of suitably large cross-sectional dimensions, the energy is guided along the interior of the pipe with comparatively little loss. The most common cross-sectional shapes are the rectangle and the circle. The cross-sectional dimensions of the tube must be greater than a certain fraction of the wavelength; otherwise the wave will not propagate in the tube. For this reason hollow waveguides are commonly used only at wavelengths of 10 cm or less (frequencies of 3000 MHz or higher).

A single wave of the type in Fig. 1 cannot propagate longitudinally inside a tubular conductor since, at some portions of the inner surface of the conducting tube, the E vector of the wave necessarily would have a component tangential to the surface. This is impossible because an electric field cannot be established along a good conductor, such as the wall of the tube. An electromagnetic wave can propagate along the interior of the tube only by reflecting back and forth between the walls of the tube. This reflection is a comparatively simple one between the plane surfaces of a rectangular tube, but is a complex reflection in tubes of other cross-sectional shapes.

A dielectric rod can also be used as a waveguide. Such a rod, if of insufficient cross-sectional

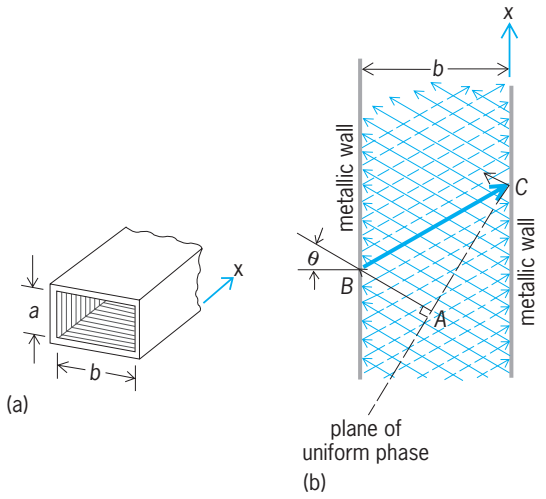


Fig. 3. Hollow metallic waveguide of rectangular cross section. (*a*) Guide. (*b*) Paths of electromagnetic energy in the simplest mode of propagation.

dimensions, can contain the electromagnetic wave by the phenomenon of total reflection at the surface.

A hollow metallic waveguide of rectangular cross section is shown in **Fig. 3***a*. The simplest mode of propagation is indicated in Fig. 3*b*. The entire space is filled with a plane electromagnetic wave which moves obliquely to the left in the direction shown by the solid arrows. This wave has its E vector normal to the paper and its H vector in the plane of the paper. Any plane normal to the direction of propagation is a plane of uniform phase (thus the name plane wave), and one such plane is indicated in the illustration by a broken line. The wave strikes the wall at an angle θ from the normal and is reflected at an equal angle. As the wave is reflected, the direction of its E vector reverses so as to make the tangential component of the electric field equal to zero at the conducting wall. The wave incident on the left wall thus is reflected to the right, where it is again reflected and moves to the left. By successive reflections the energy propagates longitudinally along the interior of the guide. As the wave incident upon the wall reflects and reverses the direction of its E vector, electric currents are caused to flow in the conducting wall. Since the wall is not a perfect conductor, some of the energy of the wave is transformed into heat. Consequently the amplitude of the wave diminishes exponentially as it passes down the guide; this phenomenon is termed attenuation. For an electromagnetic wave with a frequency of 3000 MHz (wavelength of 10 cm) propagating down the interior of a rectangular copper waveguide with cross-sectional dimensions of 1.4 by 3.2 in (or 4 by 8 cm) half the power is lost in a distance of approximately 450 ft (150 m). Hollow waveguides are used chiefly for short-distance transmission, as from a transmitter to an antenna. *See* WAVEGUIDE.

The requirements on the reflection of the wave, as outlined above, restrict the wavelength that can be propagated in a hollow guide. Consider the ray ABC in Fig. 3. The wave propagates from A to B, where it is reflected with reversal of the E vector; thereupon it propagates from B to C, where it is again reflected with another reversal of the E vector. But AC is a line of equal phase, and so the wave emerging from C must have the same phase as that at A. Thus the distance ABC must be an integral multiple of a wavelength, or $n\lambda$, where n is a positive integer. The distance ABC is $2b \cos \theta$ where b is the breadth of the guide; hence, $n\lambda = 2b \cos \theta$. The condition for propagation down the axis of the guide is that $\theta > 0$; hence $\cos \theta < 1$, and the restriction on wavelength is $\lambda < 2b/n$. The greatest ratio of wavelength to breadth of guide is obtained when $n = 1$, whence $\lambda/b < 2$. Therefore, the breadth of the waveguide must be somewhat greater than $\lambda/2$.

In the simple mode of propagation described above, the fields are independent of distance in the direction of the dimension a, and this dimension has no influence on the propagation. The net electric vector caused by the sum of the two waves is everywhere transverse to the longitudinal axis of the guide, and so the mode is described as transverse electric (TE).

If the wavelength of the radiation is small enough in comparison with the cross-sectional dimensions of the guide, more complex modes of propagation are possible, in which the wave reflects obliquely against a side wall, proceeds to the top of the guide, and reflects from there to the other side wall, then to the bottom wall, and so on. With this type of reflection it is possible to have both transverse-electric and transverse-magnetic (TM) modes. In the latter the net H vector is everywhere transverse to the axis of the guide.

When the dimensions of the guide are such that complex modes are possible, so also are the simple ones. The transmission of energy by a combination of modes introduces complications in abstracting the energy from the guide at the receiving end. Propagation in only the simplest mode is ensured by selecting the dimension b to be greater than $\lambda/2$ but not as large as λ, and also by restricting the dimension a so as to render complex modes impossible.

Waveguides of circular cross section are sometimes used. Analysis of these shows that the first TE mode is propagated if the diameter of the guide is greater than 0.586λ, and that the first TM mode is propagated if the radius is greater than 0.766λ.

Two-conductor transmission lines. Electromagnetic energy can be propagated in a simple mode along two parallel conductors. Such a waveguiding system is termed a transmission line. Three common forms are shown in **Fig. 4**. If the spacing between conductors is a small fraction of the wavelength of the transmitted energy, only one mode of propagation is possible. This corresponds to the wave of Fig. 1, with the direction of propagation taken longitudinally along the line. The E and H vectors are in the plane of the cross section, and the mode is termed transverse electromagnetic (TEM). The E vector must be at right angles to a highly conducting surface, and the oscillating H vector must be parallel with such a surface. With two separated conductors, there is for each

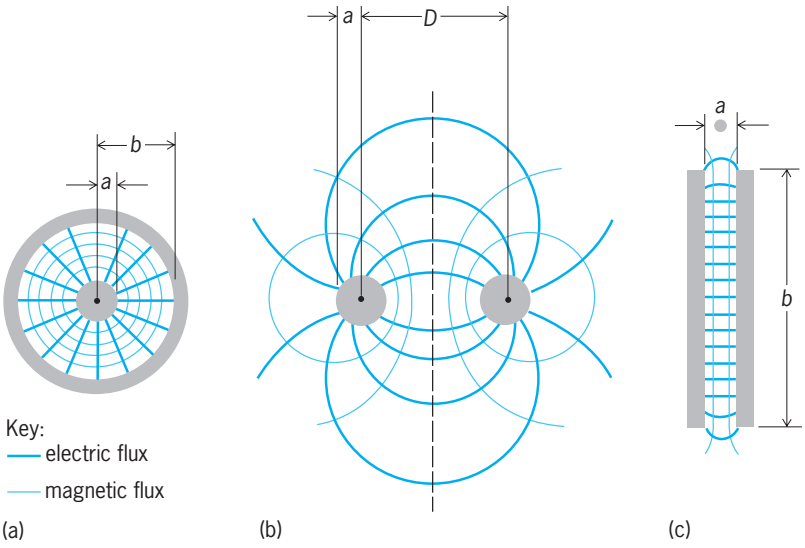


Fig. 4. Cross sections of common two-conductor transmission lines. (*a*) Coaxial cable. (*b*) Two-wire line. (*c*) Parallel-strip line.

geometrical arrangement of conductors one and only one cross-sectional field configuration which will satisfy the boundary conditions at the metal surface. The field configurations for coaxial and two-wire lines are shown in Fig. 4. At each point the ratio of E to H is as given by Eq. (1), and the velocity of propagation of the wave is as given by Eq. (2). Half of the propagated energy is contained in the electric field and half in the magnetic field. This mode of propagation is in contrast with the more complex modes required in a hollow metal pipe, where the conditions required at the boundaries can be satisfied only by means of reflections at the metal walls. As a result, the two-conductor transmission line does not have the upper limit on wavelength that was imposed on the hollow waveguide by the requirement of reflections; in fact, the two-conductor line operates completely normally at zero frequency (direct current).

At wavelengths that are small enough to be comparable with the cross-sectional dimensions of the line, more complex modes, involving reflections from the surfaces of the conductors, become possible. High-frequency energy can thus be propagated in several modes simultaneously. In a coaxial cable a rough criterion for the elimination of higher modes is that the wavelength should be greater than the average of the circumferences of the inner and outer conductors.

As the wave propagates along the line, it is accompanied by currents which flow longitudinally in the conductors. These currents can be regarded as satisfying the boundary condition for the tangential H field at the surface of the conductor. The conductors have a finite conductivity, and so these currents cause a transformation of electrical energy into heat. The energy lost comes from the stored energy of the wave, and so the wave, as it progresses, diminishes in amplitude. The conductors are necessarily supported by insulators which are imperfect and cause additional attenuation of the wave. In a typical open-wire telephone line operating at voice frequencies, half the energy is lost in a distance of perhaps 60 mi (95 km). The losses increase with frequency, and for a typical air-insulated coaxial line operating at 5 MHz, half the energy is lost in a distance of less than 1 mi (1.6 km). At a frequency of 3000 MHz ($\lambda = 10$ cm), typical distances in which half the energy is lost are, for air-insulated coaxial cable, 82 ft (25 m); for flexible coaxial cable insulated with polyethylene, 33 ft (10 m).

Noise. In a transmission line intended for the transmission of large amounts of power, such as the cross-country lines joining electrical generating stations to centers of population, the loss of an appreciable proportion of the power enroute is a serious matter. In a communication system, however, the average rate of flow of energy is rather small, and the intrinsic value of the energy itself is not of prime importance. The important characteristic of such a system is the accurate transmission of information, and the limiting factor is noise. Noise is always present in a transmission channel. Two common causes are thermal agitation and nearby electrical discharges. In a transmission system conveying information by an electromagnetic wave, the loss of energy in transmission becomes a serious matter if the wave is attenuated to the point where it is not large enough to override the noise. Amplifiers must be inserted in the transmission system at sufficiently close intervals, so that the signal never falls into the noise level, from which it could not be recovered and interpreted accurately.

Circuit analysis of transmission lines. Because the conductors of a transmission line are almost always spaced much closer together than a quarter wavelength of the electromagnetic energy that they are guiding, it is possible to analyze their performance quantitatively by circuit theory. It is then possible to deal with the voltages between the conductors and the currents flowing along the conductors, instead of with the electric and magnetic fields that exist in the insulating medium.

The waveguiding properties of the transmission line can be examined most conveniently if losses of energy are ignored. If L is defined as the inductance of the pair of conductors per unit length and C the capacitance between the conductors per unit length, field theory shows that L is μF_g and C is ϵ/F_g, where F_g is a geometrical factor that depends on the cross-sectional configuration of the conductors. For a coaxial line (Fig. 4*a*), $F_g = (1/2\pi) \log_\epsilon(b/a)$. For a two-wire line (Fig. 4*b*), $F_g = (1/\pi) \log_\epsilon(D/a)$. For a parallel-strip line (Fig. 4*c*), neglecting edge effects, $F_g = a/b$.

In the circuit analysis of a transmission line, the line can be visualized as being composed of a cascaded set of sections, each of short length Δx, as shown in **Fig. 5***b*. The partial differential equations which describe the voltage e and the current i are shown by Eqs. (7) and (8).

$$\frac{\partial e}{\partial x} = -L \frac{\partial i}{\partial t} \tag{7}$$

$$\frac{\partial i}{\partial x} = -C \frac{\partial e}{\partial t} \tag{8}$$

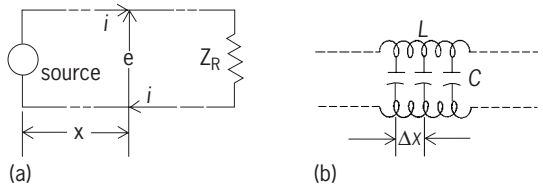


Fig. 5. Schematic representation of a transmission line. (*a*) Circuit diagram. (*b*) Visualization of *L* and *C*.

The solution of these equations is shown by Eqs. (9) and (10), where f_1 and f_2 are any finite, single-

$$e = f_1\frac{x-t}{\sqrt{LC}} + f_2\frac{x+t}{\sqrt{LC}} \tag{9}$$

$$i = \frac{1}{\sqrt{L/C}}\left(f_1\frac{x-t}{\sqrt{LC}} - f_2\frac{x+t}{\sqrt{LC}}\right) \tag{10}$$

valued functions of the arguments $x - t/\sqrt{LC}$ and $x + t/\sqrt{LC}$, respectively. These are interpreted physically as traveling waves, the first travelling in the positive x direction with the speed $1/\sqrt{LC}$ and the second traveling in the negative x direction at the same speed. Substitution of the values for L and C for any configuration of conductors yields the velocity $1/\sqrt{LC}$, which equals $1/\sqrt{\mu\epsilon}$.

The quantity $\sqrt{L/C}$ has the dimensions of ohms, termed the characteristic impedance Z_0 of the line, as shown by Eq. (11). Thus, Z_0 is a real quantity

$$Z_0 = \sqrt{L/C} = \sqrt{\mu/\epsilon}\,F_g \tag{11}$$

(a resistance) and is equal to the wave impedance of the insulating medium, $\sqrt{\mu/\epsilon}$, multiplied by the geometrical factor, F_g, characteristic of the particular configuration of conductors. For the traveling waves of voltage and current, Eqs. (9) and (10), the ratio of voltage to current of the forward-traveling wave is Z_0; that of the backward-traveling wave is $-Z_0$.

In Fig. 5*a*, a source of electrical energy is connected at one end of a transmission line and an electrical load is connected to the other. Electromagnetic energy is propagated from the sending end to the receiving end, and a portion of the energy is reflected back toward the sending end if the load impedance Z_R is different from the characteristic impedance Z_0 of the line. If Z_R equals Z_0, there is no reflection of energy at the load, and in Eqs. (9) and (10), the function f_2, representing leftward-traveling energy, is absent. This is the condition desired when the purpose of the line is to deliver energy from the source of the load. The sending-end impedance of the line is then equal to Z_0.

In addition to impedance matching to reduce reflections (echoes) along a transmission line, it is also necessary to minimize signal distortion, which consists of amplitude and phase (delay) distortion. If the line attenuation is frequency-dependent, then a signal consisting of a group of different-frequency components will undergo amplitude distortion due to the unequal attenuation of each component of the signal. Similarly, if the velocity of propagation along the line is frequency-dependent, then a delay in phase of each component will result in associated phase distortion of the signal.

Signal distortion can be minimized by the use of line loading, which is the addition of series impedances along the line and which is used to adjust the line parameters to obtain the so-called distortionless condition. Under distortionless operation the attenuation and velocity of propagation are independent of frequency. For a discussion of the distortionless line *See* TRANSMISSION LINES.

Instead of loading a line, one may employ equalizing circuits to compensate for the phase distortion along the line.

Short sections of transmission line are sometimes used to provide low-loss reactive impedances and resonant circuits at high frequencies. This is done by open-circuiting or short-circuiting the receiving end of the line to provide complete reflection of the incident energy. A short-circuited low-loss line provides the sending-end impedance shown by Eq. (12),

$$Z_s = jZ_0 \tan(2\pi f l/v) \tag{12}$$

measured in ohms. When l equals $v/4f$, the line is a quarter wavelength long, the argument $(2\pi fl/v)$ of the tangent function in Eq. (12) is $\pi/2$, and Z_s approaches an infinite value. In actual practice, losses keep Z_s to a finite value. However, at high frequencies the quarter wavelength is short and the losses are small, and such a short-circuited quarter-wave section can be used successfully as a low-loss insulator. Such a section is a resonant one and can be used as a substitute for a parallel-resonant LC circuit, for example, a tank circuit for a high-frequency oscillator. At low frequencies the required quarter wavelength is so large that losses impair the performance; also, the length becomes inconveniently great. At a frequency lower than $v/4l$, the sending-end impedance of the short-circuited line is inductive, and at frequencies between $v/4l$ and $v/2l$ the impedance is capacitive. This provides the possibility of using sections of short-circuited line as reactive elements in circuits.

Walter C. Johnson

Bibliography. F. A. Benson and T. M. Benson, *Fields, Waves, and Transmission Lines*, 1991; C. W. Davidson, *Transmission Lines for Communications,* 2d ed., 1989; C. T. A. Johnk, *Engineering Electromagnetic Fields and Waves*, 2d ed., 1987; J. A. Kong, *Electromagnetic Wave Theory*, 3d ed., 2000; P. C. Magnusson et al., *Transmission Lines and Wave Propagation*, 2000; S. Ramo, J. R. Whinnery, and T. Van Duzer, *Fields and Waves in Communication Electronics*, 3d ed., 1993; W. Sinnema, *Electronic Transmission Technology: Lines, Waves, and Antennas,* 2d ed., 1988.

Electromagnetism

The branch of science dealing with the observations and laws relating electricity to magnetism. Electromagnetism is based upon the fundamental observations that a moving electric charge produces a

magnetic field and that a charge moving in a magnetic field will experience a force.

The magnetic field produced by a current is related to the current, the shape of the conductor, and the magnetic properties of the medium around it by Ampère's law. *See* AMPÈRE'S LAW.

The magnetic field at any point is described in terms of the force it exerts on a moving charge at that point. The electrical and magnetic units are defined in terms of the ampere, which in turn is defined from the force of one current upon another.

The association of electricity and magnetism is also shown by electromagnetic induction, in which a changing magnetic field sets up an electric field within a conductor and causes the charges to move in the conductor. *See* EDDY CURRENT; ELECTRICITY; ELECTROMAGNET; ELECTROMAGNETIC INDUCTION; FARADAY'S LAW OF INDUCTION; HALL EFFECT; INDUCTANCE; LENZ'S LAW; MAGNETISM; MAXWELL'S EQUATIONS; RELUCTANCE. Kenneth V. Manning

Electrometallurgy

The branch of process metallurgy dealing with the use of electricity for smelting or refining of metals. The electrochemical effect of an electric current brings about the reduction of metallic compounds, and thereby the extraction of metals from their ores (electrowinning) or the purification of the metals (electrorefining).

In other metallurgical processes, electrically produced heat is utilized in smelting, refining, or alloy manufacturing. For a discussion of electrothermics, that is, the theory and applications of electric heating to metallurgy, *see* ELECTRIC FURNACE; ELECTRIC HEATING; ELECTROCHEMICAL PROCESS; STEEL MANUFACTURE.

Electrowinning. This metallurgical process involves the recovery of a metal, usually from its ore, by dissolving a metallic compound in a suitable electrolyte and reducing it electrochemically through passage of a direct electric current.

Following beneficiation and, sometimes, chemical pretreatment, the metal-bearing constituent of the ore is dissolved in an aqueous solution or in a molten salt. The electrolysis of the purified electrolyte with direct current yields the reduced metal at the cathode (negative electrode of an electrolytic cell), and the nonmetal is the oxidation product at the anode (positive electrode). Since the process is very selective, the purity of the electrowon metal is high, and no further refining is needed.

Water is a suitable solvent for electrolysis of metals less active than zinc and manganese, and sulfuric acid is the preferred leaching agent. The high mobility of the solvated proton provides high-conductivity solutions, and the sulfate ion is electrochemically inert; acid is regenerated by anodic oxidation and recycled for leaching.

Zinc. The process of electrowinning of zinc was developed in 1915 for treating complex ores not amenable to thermal processing, and it is presently the process of choice for most new plants. Changes in the chemistry have been minor, but steady progress has been made in the areas of process instrumentation and control, automation, and pollution-free operation.

The zinc sulfide ores are concentrated, roasted, and leached in several stages with sulfuric acid, and the electrolyte is purified. Lead and silver are insoluble through the acid leach; iron precipitates by oxidation in the neutral leach, and it coprecipitates impurities such as arsenic, antimony, and germanium. Other impurities more noble than zinc, such as copper, cadmium, cobalt, and nickel, are removed by galvanic precipitation on zinc dust.

The reduction potential of zinc is far more cathodic (−0.763 V) than that of the hydrogen ion, and its efficient electrodeposition depends upon maintaining a high hydrogen overvoltage. Metals more noble than zinc deposit on the cathode, and most promote hydrogen evolution and zinc corrosion. This reduces the yield of zinc, which is expressed by the current efficiency—the ratio of the amount of electricity (coulombs) theoretically required to yield a given quantity of metal to the amount actually consumed. Current efficiencies equal to 90% are achieved by bringing the concentration of the most harmful impurities such as germanium, arsenic, and antimony down to 0.01 mg/liter (1.3×10^{-6} oz/gal). *See* OVERVOLTAGE.

The zinc electrolyte, containing 120–160 g/liter (16–21 oz/gal) of zinc as sulfate, is circulated in lead-lined or plastic-lined [for example, poly(vinyl chloride)] concrete or wooden tanks in which the electrolysis proceeds between vertically suspended electrodes. The insoluble anodes are lead that contains small amounts of silver, and the cathodes are aluminum sheets from which the zinc deposits are stripped every 24–28 h. The submerged cathode area varies from 1.3 to 2.6 m^2. The current density varies from 300 to 600 A/m^2. From the Gibbs free-energy change for chemical reaction (1), a reversible voltage

$$ZnSO_4 + H_2O = Zn + {}^1\!/_2O_2 + H_2SO_4 \qquad (1)$$

of 2 V can be calculated. The operating cell voltage is about 3.5 V; the difference is due to the anodic and cathodic overvoltages and the ohmic voltage drop in the electrolyte and at the electrode contacts. The power consumption averages 3.2 kWh/kg zinc for modern plants. This relatively high energy cost is acceptable since special high-grade zinc is produced (99.995%). *See* ZINC METALLURGY.

Copper. Copper electrowinning is similar to zinc electrowinning, but the pyrometallurgical reduction is preferred for sulfide ores. Most of the electrowon copper comes from the direct leaching of low-grade sulfide ores, or from oxide ores. The recent development of liquid ion-exchange reagents specially for copper makes possible an upgrading, through solvent extraction of dilute solutions, from dump or in-place leaching in order to obtain a concentrated solution suitable for electrolysis. The electrowinning of copper is carried out in sulfate solutions containing

from 20 to 40 g/liter (2.7–5.4 oz/gal) of copper and about 100 g/liter (13 oz/gal) of acid at current densities from 160 to 200 A/m^2 (15–19 A/ft^2). The anodes are made of lead, alloyed with antimony, and copper is deposited on copper starting sheets.

Aluminum. Electrowinning has been the only process used for the commercial production of aluminum since it was developed independently by C. M. Hall and by P. L. T. Héroult in 1885. Alumina extracted from bauxite ores and purified by digestion in a caustic soda solution is dissolved in a mixture of cryolite ($AlF_3 \cdot 3NaF$) and other fluorides (for example, CaF_2 and LiF). The molten-salt electrolyte floats on the molten aluminum, which is the cathode, and carbon anodes dip into the bath from above.

The carbon reacts with the oxygen produced by anodic oxidation. The overall cell reaction (2) corresponds to a theoretical reversible voltage of 1.2 V, but the cell operates under a total voltage of about 4.2 V. The heat generated by irreversible electrochemical and ohmic losses maintains the cell temperature at about 960°C (1760°F).

$$Al_2O_3 + {}^3\!/_2C = 2Al + {}^3\!/_2CO_2 \qquad (2)$$

Current densities are about 1 A/cm^2, with total currents of 100,000 A for a modern cell. Alumina is periodically added to the melt to maintain its concentration between 2 and 5%. The current efficiency averages 89%; the main losses are attributed to the recombination of aluminum by reduction of carbon dioxide to carbon monoxide. The energy consumption is about 14 kWh/kg aluminum. *See* ALUMINUM.

Magnesium. The electrowinning of magnesium, another well-established process, accounts for about 80% of the metal output. The electrolyte is a mixture of chlorides of potassium, sodium, calcium, and magnesium. Since the reduced metal has a lower specific gravity than the electrolyte has, it floats to the surface; the cells are designed so that it does not come in contact with air or with the chlorine gas produced at the anode.

Other metals. Cadmium, and in smaller quantities, thallium and indium, are recovered by electrolysis in sulfate solutions. Cobalt is also electrowon by a similar process, and some of the nickel production is achieved by electrolysis in sulfate-chloride solutions. Pure metallic chromium and manganese are produced by electrowinning; these metals are rather difficult to electrodeposit, and the acid produced at the anode must be kept away from the cathode by a diaphragm. Antimony is electrowon from a sodium sulfide electrolyte, and gallium is produced by electrolysis of a caustic solution.

Metals more active than manganese cannot be reduced cathodically from aqueous solutions, but most can be electrowon from mixtures of molten salts. Usually, a compound of the metal is dissolved in salts of still more active metals in order to improve the conductivity and lower the melting point of the melt. Solid metallic deposits are difficult to recover, and all the commercially successful processes yield the liquid metal. The solubility and reactivity of the reduced metal in the melt is often important, and the cells must be designed to protect the reduced metal from contact with the anodic reaction product or with air. The cell electrolyte cannot be easily recycled, and the feed must be carefully purified and dehydrated.

All the sodium metal produced today comes from electrolysis of a molten chloride mixture between a steel cathode and a graphite anode separated by a perforated steel diaphragm. Lithium is produced by a similar process. The other alkali and alkaline-earth metals, as well as the rare earths, can be electrowon from molted chlorides. Processes have been developed for electrowinning titanium, tantalum, and niobium in molten salts, and uranium, thorium, and zirconium have been obtained in the laboratory. Beryllium and boron are also produced by electrolysis.

Electrorefining. This is a purification process in which an impure metal anode is dissolved electrochemically in a solution of a salt of the metal being refined; the pure metal is recovered by electrodeposition at the cathode.

Electrorefining is the most economical method for securing the high purity required for many uses of nonferrous metals. Since the same electrochemical reaction proceeds in opposite directions at the anode and the cathode, the overall chemical change is a small change in the activity of the metal at the two electrodes, and the reversible cell voltage is practically zero. The anodic and cathodic overvoltages, required for an economical reaction rate, are small, and the operating cell voltage consists mainly of the ohmic voltage drop through the electrolyte and at the electrode contacts. The power consumption is moderate, and electrorefining is less sensitive to the cost of electric power than other electrometallurgical processes are.

Electochemical refining is a more efficient purification process than other chemical methods are because it is very selective. In particular, for metals such as copper, silver, gold, and lead, the operating electrode potential is close to the reversible potential, and a sharp separation is accomplished, both at the anode, where more noble metals do not dissolve, and at the cathode, where more active metals do not deposit.

The first commercial electrolytic copper refinery was established in 1871, and today most of the copper is electrorefined. Minor quantities of some impurities lower the electrical conductivity of copper markedly, and the refining process ensures that the metal will meet the specifications of the electrical industry. Furthermore, silver, gold, and other precious metals are removed by the electrolytic refining, and their recovery is thus an economic asset of the process.

The impure copper is cast in the shape of anodes approximately 3.0 × 3.3 ft (0.9 × 1.0 m) and 1.4–1.8 in. (3.5–4.5 cm) thick. The starting cathodes are pure copper sheets, prepared by electrodeposition. The electrodes are placed vertically in cells which are lead-lined or plastic-lined wooden or concrete tanks. The anodes are replaced at regular intervals of

20–28 days, and two successive cathodes are produced during the same period. The solution contains about 45 g/liter (6 oz/gal) of copper as copper sulfate and approximately 200 g/liter (26 oz/gal) of sulfuric acid. The temperature is maintained at 55–60°C (131–140°F) to lower the resistance of the electrolyte which is circulated through the cell. The current density is usually 200–250 A/m^2, and the current efficiency averages 95%. The cell voltage is 0.2–0.3 V, and the power consumption varies from 0.18 to 0.25 kWh/kg copper.

Metals less noble than copper, such as iron, nickel, cobalt, zinc, and manganese, dissolve from the anode. Lead is oxidized, and it precipitates as lead sulfate. Other impurities such as arsenic, antimony, and bismuth partly remain as insoluble compounds in the slimes and partly dissolve as complexes in the electrolyte. Precious metals do not dissolve at the potential of the anode, and they remain as metals in the slimes, along with insoluble selenium and tellurium compounds and some metallic copper particles that fall from the anode. Metals less noble than copper do not deposit at the cathode. In particular, nickel and arsenic accumulate in the electrolyte, and their concentration is controlled by partial withdrawal and purification of the solution.

Nickel is electrorefined in a sulfate-chloride electrolyte. Its anodic dissolution requires a significant overvoltage, and both copper and nickel dissolve at the anode potential. To prevent plating of the dissolved copper at the cathode, a diaphragm cell is used, and the anodic solution is circulated through a purification circuit before entering the cathode compartment. Cobalt is electrorefined by a similar process.

Very high-purity lead is produced by electrorefining in which a fluosilicate electrolyte is used, and a sulfate bath is used for purifying tin. Silver is electrorefined in a copper and silver nitrate electrolyte, and gold is refined by electrolysis in a chloride bath.

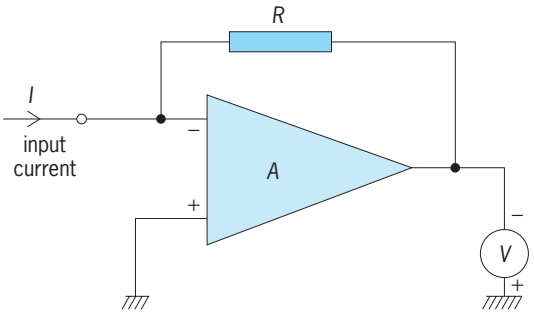
The electrorefining of many metals can be carried out with molten-salt electrolytes, but the processes are usually expensive, and the only industrial application is the electrorefining of aluminum by the "three-layer" process. The density of the electrolyte is adjusted so that a pure molten aluminum cathode will float on the molten salt, which in turn will float on an anode consisting of a molten mixture of impure aluminum and copper. *See* PYROMETALLURGY.

Paul Duby

Bibliography. A. T. Kuhn, *Industrial Electrochemical Processes,* 1971; A. T. Kuhn (ed.), *Techniques in Electrochemistry, Corrosion, and Metal Finishing: A Handbook*, 1987; D. Pletcher and F. C. Walsh, *Industrial Electrochemistry*, 2d ed., 1993.

Electrometer

A highly sensitive instrument which measures all or some of the following variables: current, charge, voltage, and resistance. There are two classes of electrometers, mechanical and electronic. The mechanical instruments have been largely replaced by electronic types.



A common arrangement for an electronic electrometer.

Mechanical. These instruments, with origins in the nineteenth century and earlier, rely for their operation on the mechanical forces associated with electrostatic fields. The negligible current drawn by many electrometers contributes to their high sensitivity, particularly in the context of charge and current measurement.

In common with other electrostatic instruments, electrometers are classified as either attracted-disk or symmetrical types. Attracted-disk instruments, in which the attractive force between two plates, with a potential difference between them, is measured in terms of the fundamental units of mass and length, are sometimes termed absolute electrometers. Attracted-disk instruments are widely used as electrostatic voltmeters for measuring potentials greater than 1 kilovolt. The quadrant electrometer was a particularly sensitive instrument based on a horizontal vane suspended by a torsion fiber so that it could rotate between fixed electrodes. *See* ELECTROSCOPE; ELECTROSTATICS; VOLTMETER.

Electronic. Electronic types utilize some form of electronic amplifier, typically an operational amplifier with a field-effect-transistor input stage to minimize the input current. In the **illustration** the amplifier A has a current measuring resistor R providing negative feedback. The unknown input current I can be found from the voltmeter reading V, as $I = V/R$. Charge Q can be measured if R is replaced by a capacitor C. For this case, $Q = VC$. *See* AMPLIFIER; TRANSISTOR.

In the most sensitive applications, problems arise due to drift in the amplifier characteristics and electrical noise present in the circuit components. To obviate these effects, electrometers employing a vibrating capacitor or varactor diodes are used. The signal to be measured is converted to an alternating-current (ac) signal and subsequently amplified by an ac amplifier which is less susceptible to drift and noise. The amplified signal is finally reconverted to direct current (dc). *See* ELECTRICAL NOISE; VARACTOR.

The most sensitive electrometers commercially available can resolve currents of 0.2 femtoampere under favorable circumstances. *See* CURRENT MEASUREMENT; ELECTRICAL MEASUREMENTS; VOLTAGE MEASUREMENT.

R. W. J. Barker

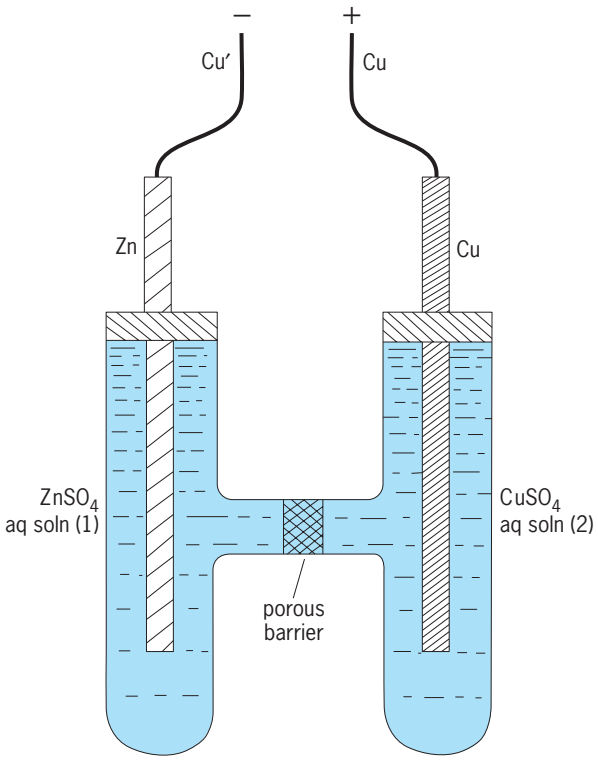
Electromotive force (cells)

The voltage or electric potential difference across the terminals of a cell when no current is drawn from it. The electromotive force (emf) is the sum of the electric potential differences produced by a separation of charges (electrons or ions) that can occur at each phase boundary (or interface) in the cell. The magnitude of each potential difference depends on the chemical nature of the two contacting phases. Thus, at the interface between two different metals, some electrons will have moved from the metal with a higher free energy of electrons to the metal with a lower free energy of electrons. The resultant charge separation will produce a potential difference, just as charge separation produces a voltage across a capacitor; at equilibrium this exactly opposes further electron flow. Similarly, potential differences can be produced when electrons partition across a metal|solution interface or metal|solid interface, and when ions partition across a solution|membrane|solution interface.

Origin. How a cell emf is composed of the sum of interfacial potential differences is shown by the Daniell cell (see **illus.**) and scheme (1);

$$\begin{aligned} &\mathrm{Cu'}\ \underset{a}{|}\ \mathrm{Zn}\ \underset{b}{|}\mathrm{aq\ soln\ (1)}: \\ &\qquad \mathrm{ZnSO_4}\ \underset{c}{\vdots}\ \mathrm{aq\ soln\ (2)}:\ \mathrm{CuSO_4}\ \underset{d}{|}\ \mathrm{Cu} \end{aligned} \tag{1}$$

aq soln denotes an aqueous solution, a solid line indicates a phase boundary, and the broken line a porous barrier permeable to all the ions in the adjacent solutions. The copper connector to the zinc electrode is denoted Cu′. The barrier prevents physical mixing of the zinc sulfate ($ZnSO_4$) and copper sulfate ($CuSO_4$) solutions.



Cross-section diagram of Daniell cell. (*After W. J. Moore, Physical Chemistry, 5th ed., Longman, 1974*)

The cell emf is the open-circuit (that is, zero current) potential difference measured between the two Cu leads (any potential-measuring device must ultimately measure the potential difference between two chemically identical phases, in this example the Cu and Cu′ phases). An expression can be written for the cell emf in terms of the electrochemical potentials of the species involved in the interfacial equilibria. The electrochemical potential $\bar{\mu}_j^{\alpha}$ of species j in phase α is defined by Eq. (2),

$$\bar{\mu}_j^{\alpha} = \mu_j^{\alpha} + z_j F \phi^{\alpha} \tag{2}$$

where μ_j^{α} is the chemical potential of species j in phase α, z_j is the number of electronic charges (with sign) on the species, F is the Faraday constant (96485.3 coulombs for Avogadro's number of electrons), and ϕ^{α} is the inner potential of phase α; $z_j F \phi^{\alpha}$ is the electrical work required to move $z_j F$ coulombs of charge into phase α from infinite distance in a vacuum. For a species in solution, Eq. (2) can be written as Eq. (3),

$$\bar{\mu}_j^{\alpha} = \mu_j^{\circ\alpha} + RT\ \ln\ a_j^{\alpha} + z_j F \phi^{\alpha} \tag{3}$$

where $\mu_j^{\circ\alpha}$ is the standard chemical potential of species j in phase α, R is the gas constant (8.3145 joules/mole/degree Kelvin), T is the temperature in Kelvin, and a_j^{α} is the activity of species j in phase α. For a neutral species (that is, when $z_j = 0$), Eq. (4) applies.

$$\bar{\mu}_j^{\alpha} = \mu_j^{\alpha} = \mu_j^{\circ\alpha} + RT \ln a_j^{\alpha} \tag{4}$$

For a fully ionized solution species at low concentrations, a_j^{α} is well approximated by the concentration c_j^{α}; the activity of a pure phase (for example, zinc and copper in the Daniell cell) is defined as unity.

The value of $\bar{\mu}_j$ must be identical for any species j equilibrated between two different phases. Thus for equilibrium of species j between phases α and β, Eq. (5) applies.

$$\bar{\mu}_j^{\alpha} = \bar{\mu}_j^{\beta} \tag{5}$$

More generally, for any interfacial equilibrium between phases α and β, Eq. (6) applies; ν_j is the stoichiometric number of species j (ν_j is positive for products and negative for reactants), and $\bar{\mu}_j^{\text{phase}}$ is the electrochemical potential of species j in the phase in which it is located. Thus, for the interfacial equilibrium $Cu^{2+} + 2e^- \rightleftarrows Cu$ at interface d in scheme (1), Eq. (6) becomes Eq. (7).

$$\sum_i \nu_j \bar{\mu}_j^{\text{phase}} = 0 \tag{6}$$

$$\bar{\mu}_{\mathrm{Cu}^{2+}}^{\mathrm{soln(2)}} + 2\bar{\mu}_{e^-}^{\mathrm{Cu}} = \bar{\mu}_{\mathrm{Cu}}^{\mathrm{Cu}} \tag{7}$$

Similar equations can be used to quantify the potential differences across each interface in

scheme (1), assuming that the potential difference across interface c is zero (a good approximation when the separator is equally permeable to all the ions in the two adjacent phases). For interfaces a, b, and d of scheme (1), Eqs. (8) can be written.

Interface a (partitioning of electrons between Cu′ and Zn phases):

$$\bar{\mu}_{e^-}^{\mathrm{Cu}'} = \bar{\mu}_{e^-}^{\mathrm{Zn}} \tag{8a}$$

Interface b (equilibrium: $Zn^{2+} + 2e^- \rightleftarrows Zn$):

$$\bar{\mu}_{\mathrm{Zn}^{2+}}^{\mathrm{soln}(1)} + 2\bar{\mu}_{e^-}^{\mathrm{Zn}} = \bar{\mu}_{\mathrm{Zn}}^{\mathrm{Zn}} \tag{8b}$$

Interface d (equilibrium: $Cu^{2+} + 2e^- \rightleftarrows Cu$)

$$\bar{\mu}_{\mathrm{Cu}^{2+}}^{\mathrm{soln}(2)} + 2\bar{\mu}_{e^-}^{\mathrm{Cu}} = \bar{\mu}_{\mathrm{Cu}}^{\mathrm{Cu}} \tag{8c}$$

Rearranging Eqs. (8) gives Eq. (9).

$$\bar{\mu}_{e^-}^{\mathrm{Cu}} - \bar{\mu}_{e^-}^{\mathrm{Cu}'} = \bar{\mu}_{e^-}^{\mathrm{Cu}} - \bar{\mu}_{e^-}^{\mathrm{Zn}} = \frac{1}{2}\left[\bar{\mu}_{\mathrm{Cu}}^{\mathrm{Cu}} - \bar{\mu}_{\mathrm{Cu}^{2+}}^{\mathrm{soln}(2)}\right] - \frac{1}{2}\left[\bar{\mu}_{\mathrm{Zn}}^{\mathrm{Zn}} - \bar{\mu}_{\mathrm{Zn}^{2+}}^{\mathrm{soln}(1)}\right] \tag{9}$$

Since $\mu_{e^-}^{\circ\mathrm{Cu}} = \mu_{e^-}^{\circ\mathrm{Cu}'}$ and for an electron $z_{e^-} = -1$, Eq. (3) can be used to write Eq. (10).

$$\bar{\mu}_{e^-}^{\mathrm{Cu}} - \bar{\mu}_{e'}^{\mathrm{Cu}'} = -F(\phi^{\mathrm{Cu}} - \phi^{\mathrm{Cu}'}) \tag{10}$$

The cell potential, E_{cell}, is simply the sum of all the equilibrium interfacial potential differences, as in Eq. (11), where it has been assumed that $\phi^{\mathrm{soln}(2)} - \phi^{\mathrm{soln}(1)} - 0$. Combining Eqs. (3), (10) and (11) (and

$$E_{\mathrm{cell}} = \left(\phi^{\mathrm{Cu}} - \phi^{\mathrm{soln}(2)}\right) + \left(\phi^{\mathrm{soln}(1)} - \phi^{\mathrm{Zn}}\right) + (\phi^{\mathrm{Zn}} - \phi^{\mathrm{Cu}'}) = (\phi^{\mathrm{Cu}} - \phi^{\mathrm{Cu}'}) \tag{11}$$

noting that the copper and zinc ions have a charge of 2+ and that the activity of the metal phases is unity) gives Eq. (12),

$$E_{\mathrm{cell}} = (\phi^{\mathrm{Cu}} - \phi^{\mathrm{Cu}'}) = -\frac{1}{F}\left[\bar{\mu}_{e^-}^{\mathrm{Cu}} - \bar{\mu}_{e^-}^{\mathrm{Cu}'}\right] = \frac{1}{2F}\left[\mu_{\mathrm{Cu}^{2+}}^{\circ\,\mathrm{soln}(2)} - \mu_{\mathrm{Cu}}^{\circ} + \mu_{\mathrm{Zn}}^{\circ} - \mu_{\mathrm{Zn}^{2+}}^{\circ\,\mathrm{soln}(1)}\right] + \frac{RT}{2F}\ln a_{\mathrm{Cu}^{2+}}^{\mathrm{soln}(2)} - \frac{RT}{2F}\ln a_{\mathrm{Zn}^{2+}}^{\mathrm{soln}(1)} \tag{12}$$

where E_{cell} is the potential difference between the two copper leads [scheme (1)] measured at open circuit (that is, when no current passes through the cell) and is therefore the emf of the cell. Dropping the superscripts denoting the solutions, Eq. (12) can be rewritten as Eq. (13),

$$E_{\mathrm{cell}} = \left[E_{\mathrm{Cu}^{2+}|\mathrm{Cu}}^{\circ} + \frac{RT}{2F}\ln a_{\mathrm{Cu}^{2+}}\right] - \left[E_{\mathrm{Zn}^{2+}|\mathrm{Zn}}^{\circ} + \frac{RT}{2F}\ln a_{\mathrm{Zn}^{2+}}\right] \tag{13}$$

where $E_{\mathrm{Cu}^{2+}|\mathrm{Cu}}^{\circ}$ and $E_{\mathrm{Zn}^{2+}|\mathrm{Zn}}^{\circ}$ are termed the standard electrode potentials of the copper and zinc half cells. Equation (13) is an example of a Nernst equation, which relates cell emf to the activities of the cell constituents. *See* CHEMICAL EQUILIBRIUM; CHEMICAL THERMODYNAMICS; ELECTROMOTIVE FORCE (EMF); THERMODYNAMIC PRINCIPLES.

Half cells. It is convenient to describe any electrochemical cell in terms of half cells. A half cell consists of an oxidant (Ox) and reductant (Red) such that $Ox + ne^- \rightleftarrows Red$; species Ox and Red are commonly referred to as a redox couple. The Daniell cell, for example, comprises the two half cells $Cu^{2+} + 2e^- \rightleftarrows Cu$ (redox couple is $Cu^{2+}|Cu$) and $Zn^{2+} + 2e^- \rightleftarrows Zn$ (redox couple is $Zn^{2+}|Zn$) with the half-cell potentials $E_{\mathrm{Cu}^{2+}|\mathrm{Cu}}$ and $E_{\mathrm{Zn}^{2+}|\mathrm{Zn}}$ given by Eqs. (14) and (15). Such expressions are useful only if the E°

$$E_{\mathrm{Cu}^{2+}|\mathrm{Cu}} = E_{\mathrm{Cu}^{2+}|\mathrm{Cu}}^{\circ} + \frac{RT}{2F}\ln a_{\mathrm{Cu}^{2+}} \tag{14}$$

$$E_{\mathrm{Zn}^{2+}|\mathrm{Zn}} = E_{\mathrm{Zn}^{2+}|\mathrm{Zn}}^{\circ} + \frac{RT}{2F}\ln a_{\mathrm{Zn}^{2+}} \tag{15}$$

value for each half cell is known. Values of E° can be assigned to any given half cell by arbitrarily specifying that $E^{\circ}_{\mathrm{H}^+|\mathrm{H}_2}$ [the E° for the standard hydrogen electrode (SHE) half cell, $H^+(aq, a = 1) + e^- \rightleftarrows {}^1/_2H_2(g, 1\ atm)$] is zero. The temperature dependence $dE^{\circ}_{\mathrm{H}^+|\mathrm{H}_2}/dT$ is also specified as zero. E° values versus the SHE for selected half cells are given in the **Table 1**. In principle, any E° (and its temperature dependence) can be measured directly versus the SHE, or another half cell whose electrode potential and temperature dependence have been determined. Such an electrode is termed a reference electrode. The reference electrode is a half cell designed so that its potential is stable and reproducible, and it neither contaminates nor is contaminated by the medium in which it is immersed. Two convenient reference electrodes commonly used in aqueous systems are the saturated calomel and the silver|silver chloride electrodes (see table). *See* ELECTROCHEMISTRY; ELECTRODE POTENTIAL; OXIDATION-REDUCTION; REFERENCE ELECTRODE.

E° values and thermodynamics. The E° values associated with the two half cells that make up a cell provide fundamental thermodynamic information about the chemical reaction between the redox couples of the two half cells. Returning to the example of the Daniell cell and combining Eqs. (12) and (13) yields

Selected standard electrode potentials at 25°C

Electrode reaction	E°/V
$Li^+(aq) + e^- \rightleftarrows Li(s)$	−3.045
$Zn^{2+}(aq) + 2e^- \rightleftarrows Zn(s)$	−0.763
$2H^+(aq) + 2e^- \rightleftarrows H_2(g)$	0
$Hg_2Cl_2(s) + 2e^- \rightleftarrows 2Hg(s) + 2Cl^-$ (sat aq KCl)	0.246
$AgCl(s) + e^- \rightleftarrows Ag(s) + Cl^-(aq)$	0.222
$Cu^{2+}(aq) + 2e^- \rightleftarrows Cu(s)$	0.337
$Fe(CN)_6^{3-}(aq) + e^- \rightleftarrows Fe(CN)_6^{4-}(aq)$	0.69
$O_2(g) + 4H^+(aq) + 4e^- \rightleftarrows 2H_2O$	1.223
$F_2(g) + 2H^+ + 2e^- \rightleftarrows 2HF(aq)$	3.06

Eq. (16). The term on the left-hand side is the stan-

$$E^\circ_{Cu^{2+}|Cu} - E^\circ_{Zn^{2+}|Zn} = \frac{1}{2F}\left[\mu^{\circ\,soln(2)}_{Cu^{2+}} - \mu^\circ_{Cu} + \mu^\circ_{Zn} - \mu^{\circ\,soln(1)}_{Zn^{2+}}\right] \quad (16)$$

dard cell emf E°_{cell} (that is, the value of E_{cell} when all the cell constituents are in their standard states of unit activity), and the bracketed term on the right-hand side is $-\Delta G^\circ$, where ΔG° is the standard Gibbs energy change of the cell reaction $Cu^{2+} + Zn \rightarrow Cu + Zn^{2+}$. Hence Eq. (16) can be written as Eq. (17).

$$E^\circ_{cell} = -\frac{\Delta G^\circ}{2F} \quad (17)$$

More generally, for the two half-cell reactions $Ox_1 + n_1e^- \rightleftarrows Red_1$ and $Ox_2 + n_2e^- \rightleftarrows Red_2$, E°_{cell} is defined by Eq. (18). The cell reaction (19) involves the trans-

$$E^\circ_{cell} = E^\circ_{Ox_1|Red_1} - E^\circ_{Ox_2|Red_2} \quad (18)$$

$$n_2Ox_1 + n_1Red_2 \rightleftarrows n_2Red_1 + n_1Ox_2 \quad (19)$$

fer of n_1n_2 electrons, and Eq. (17) takes the general form (20), where $n = n_1n_2$. If the cell is not in its standard state, the relation becomes Eq. (21).

$$E^\circ_{cell} = -\frac{\Delta G^\circ}{n_1n_2F} = -\frac{\Delta G^\circ}{nF} \quad (20)$$

$$E_{cell} = -\frac{\Delta G}{nF} \quad (21)$$

Using the general thermodynamic relation $\Delta G^\circ = -RT \ln K_{eq}$, it is also possible to find the equilibrium constant K_{eq} of the cell reaction from the relation (22).

$$K_{eq} = \frac{[Ox_2]^{n_1}[Red_1]^{n_2}}{[Ox_1]^{n_2}[Red_2]^{n_1}} = \exp\left[-\frac{\Delta G^\circ}{RT}\right] = \exp\left[\frac{n_1n_2FE^\circ_{cell}}{RT}\right] \quad (22)$$

See FREE ENERGY.

Membrane potentials. When two ionic solutions of different composition are separated by a membrane, a potential difference can develop across the membrane. The complete cell requires electrical contacts with the solutions on each side of the membrane, accomplished with reference electrodes, as in scheme (23),

re(lhs) | soln(lhs) | membrane | soln(rhs) | re(rhs) (23)

where re(lhs) and re(rhs) are the reference electrodes in the left and right solutions, respectively. The cell emf will be as in Eq. (24).

$$E_{cell} = E_{re(rhs)} + \phi^{soln(rhs)} - \phi^{soln(lhs)} - E_{re(lhs)} \quad (24)$$

The potential difference across the membrane, $\phi^{soln(rhs)} - \phi^{soln(lhs)}$, is a function of the membrane properties as well as the composition of the adjacent solutions. The simplest example occurs when only one particular ion (such as the hydrogen ion, H^+) is transported across the membrane. The transported ion will tend to move from the side with high concentration to the side with low concentration; however, in the process of doing so, charge is separated across the membrane, producing a potential difference that exactly counters the ion transport. The membrane potential difference is given by Eq. (25),

$$\phi^{soln(rhs)} - \phi^{soln(lhs)} = \frac{RT}{z_jF}\ln\frac{a_j^{soln(lhs)}}{a_j^{soln(rhs)}} \approx \frac{RT}{z_jF}\ln\frac{c_j^{soln(lhs)}}{c_j^{soln(rhs)}} \quad (25)$$

where a_j and c_j are the activity and concentration of the transported ion. The cell emf is given by Eq. (26).

$$E_{cell} = E_{re(rhs)} + \frac{RT}{z_jF}\ln\frac{c_j^{soln(lhs)}}{c_j^{soln(rhs)}} - E_{re(lhs)} \quad (26)$$

If the concentration on one side, for example, $c_j^{soln(lhs)}$ is kept constant, E_{cell} will reflect any variation of the other concentration, $c_j^{soln(rhs)}$, and the system becomes a sensor for that ion.

As long as the membrane transports only a single ionic species, the potential difference is thermodynamically determined. As soon as the membrane transports more than one ionic species, irreversible thermodynamics come into play; the membrane potential difference will exhibit a complicated dependence on the concentrations and mobilities of the ions within the membrane as well as on their concentrations in the adjacent solutions. In certain circumstances, the response is still selective for one ion and the membrane electrode has analytical utility. The glass electrode used for measurement of pH, in which a thin glass membrane responds selectively to the H^+ ion, is an electrode of this type.

Potential differences known as liquid junction potentials arise where two different ionic solutions make contact through a permeable separator such as a glass frit. These interfere with thermodynamic measurements and can be minimized by interposing a salt bridge, that is, a concentrated immobilized solution of potassium chloride (KCl), or other salt in which the cation and anion have nearly identical mobilities, between the two solutions. Numerous theories based on irreversible thermodynamics have evolved to estimate the magnitude of liquid junction potentials. *See* BIOPOTENTIALS AND IONIC CURRENTS; ION-SELECTIVE MEMBRANES AND ELECTRODES.

Uses. The emf produced by a single cell or a set of cells in series (a battery) is used as a dc power source for a wide array of applications, ranging from powering wristwatches to emergency power supplies. The emf of the saturated Weston cell (scheme 27;

Hg(liquid) | Hg_2SO_4(solid) | $CdSO_4$ (sat aq) | Cd(Hg) (27)

sat aq = saturated aqueous solution) still serves as a high-level voltage reference for the National Institute of Standards and Technology. However, Josephson

arrays are now considered the most precise voltage references, and Zener diodes are used to produce reference voltages for many laboratory and field voltage measurement devices.

There are also a few photoelectrochemical cells with possible application in solar energy conversion, such as the Grätzel cell (scheme 28; nonaq = non-

$$TiO_2(\text{solid}) \mid \text{adsorbed } Ru^{2G} \text{ dye} \mid I_2,\ I^-\ (\text{nonaq}) \mid C \quad (28)$$

aqueous solution), where the cell emf is produced by absorption of visible light.

The emf of a cell can also be used as an indicator of chemical composition. Devices that depend on the measurement of an open circuit-cell potential work well when the device is sensitive to a single analyte, but they are notoriously sensitive to interferences. An example is the so-called alkaline error that occurs when glass electrodes are used to measure very high pH values. For complex systems containing several species that can undergo electrochemical reaction, individual E° values can be determined using electroanalytical techniques that involve controlling the potential applied to a cell with measurement of the resultant current. Techniques such as polarography and cyclic voltammetry, for example, involve changing the potential of an indicator electrode and observing a wave or peak in the current at the redox potential of the species in solution; the height of the wave or peak indicates the concentration of that species. *See* BATTERY; ELECTROCHEMICAL TECHNIQUES; FUEL CELL; POLAROGRAPHIC ANALYSIS.

Cell voltages when current passes. E_{cell} values, which are thermodynamic quantities, do not give any information about the kinetics of the electrode reactions. When current passes through a cell, the cell voltage (that is, the potential difference between the terminals) will differ from the emf because of resistive losses within the cell (a function of cell design) and because the kinetics of the electrode reactions are not fast enough to sustain thermodynamic (nernstian) behavior at high rates of reaction. These kinetics are very dependent on temperature and the nature of the electrode material. When a cell is discharged through an external load, the cell voltage will be less than the emf; if the direction of current flow is reversed (using an external power supply), the cell voltage will be greater than the emf. In both cases the greater the current, the greater the deviation of the cell voltage from the emf. In some cases the electrode reactions are so slow that even the open-circuit cell voltage may not be a reliable measure of the emf.

For example, a cell comprising the two half cells $2H^+(aq) + 2e^- \rightleftarrows H_2(g)$ and $O_2(g) + 4H^+(aq) + 4e^- \rightleftarrows 2H_2O$ (where g = gas), both under standard conditions, has a cell emf of 1.223 V at 25°C (see table). A highly sophisticated version of this cell, the hydrogen/oxygen fuel cell, is designed to minimize internal cell resistance, and has electrodes tailored to maximize the rates of the electrode reactions; nevertheless, the voltage under load is usually less than 0.8 V, with most of the voltage loss occurring at the oxygen electrode. When an external power supply is used to electrolyze water, the cell voltage required to produce H_2 and O_2 at a reasonable rate is $\gtrsim$1.6 V; and this again depends critically on the cell design and the choice of electrodes.

Mary D. Archer; Stephen W. Feldberg

Bibliography. R. G. Compton and G. H. W. Sanders, *Electrode Potentials*, Oxford University Press, 1996; D. R. Crow, *Principles and Application of Electrochemistry*, 4th ed., Blackie Academic & Professional, 1994; A. R. Denaro, *Elementary Electrochemistry*, 2d ed., 1971; D. J. G. Ives and G. J. Janz, *Reference Electrodes: Theory and Practice*, Academic Press, 1961; J. Koryta, *Ion-Selective Electrodes*, Cambridge University Press, 1975; W. J. Moore, *Physical Chemistry*, 5th ed., Longman, 1974.

Electromotive force (emf)

A measure of the strength of a source of electrical energy. The term is often shortened to emf. It is not, of course, a force in the usual mechanical sense (and for this reason has sometimes been called electromotance), but it is a conveniently descriptive term for the agency which drives current through an electric circuit. In the simple case of a direct current I (measured in amperes) flowing through a resistor R (in ohms), Ohm's law states that there will be a voltage drop (or potential difference) of $V = IR$ (in volts) across the resistor. To cause this current to flow requires a source with emf (also measured in volts) $E = V$. More generally, Kirchhoff's voltage law states that the sum of the source emf's taken around any closed path in an electric circuit is equal to the sum of the voltage drops. This is equivalent to the statement that the total emf in a closed circuit is equal to the line integral of the electric field strength around the circuit. *See* ELECTRIC CURRENT; ELECTRIC FIELD; ELECTRICAL RESISTANCE; OHM'S LAW.

A source of emf E delivering a current I supplies electrical energy at a rate of EI (in watts). Therefore, the emf in a circuit is numerically equal to the work done in carrying a charge of 1 coulomb around the circuit.

Open-circuit emf. A source such as a battery has an internal resistance which causes an internal voltage drop when a current is drawn from it. Consequently the emf at the external terminals varies with the current. By extrapolating to zero current, the open-circuit emf may be derived. It is the open-circuit emf which is usually said to be the emf of a source; together with the value of the internal resistance, it enables the performance of the source in any circuit to be calculated.

Types of emf. An emf may be steady (direct), as for a battery, or time-varying, as for a charged capacitor discharging through a resistor. A class of time-varying sources of very great practical importance is that of generators of alternating (sinusoidal) emf. Emf's may be generated by a variety of physical, chemical, and

biological processes. Some of the more important are:

1. Electrochemical reactions, as used in direct-current (dc) batteries, in which the emf results from the reactions between electrolyte and electrodes. *See* BATTERY; ELECTROCHEMISTRY; ELECTROMOTIVE FORCE (CELLS).

2. Electromagnetic induction, in which the emf results from a change in the magnetic flux linking the circuit. This finds application in alternating-current rotary generators and transformers, providing the basis for the electricity supply industry. *See* ALTERNATING-CURRENT GENERATOR; ELECTROMAGNETIC INDUCTION; FARADAY'S LAW OF INDUCTION; TRANSFORMER.

3. Thermoelectric effects, in which a temperature difference between different parts of a circuit produces an emf. The main use is for the measurement of temperature by means of thermocouples; there are some applications to electric power generation. *See* THERMOCOUPLE; THERMOELECTRIC POWER GENERATOR; THERMOELECTRICITY.

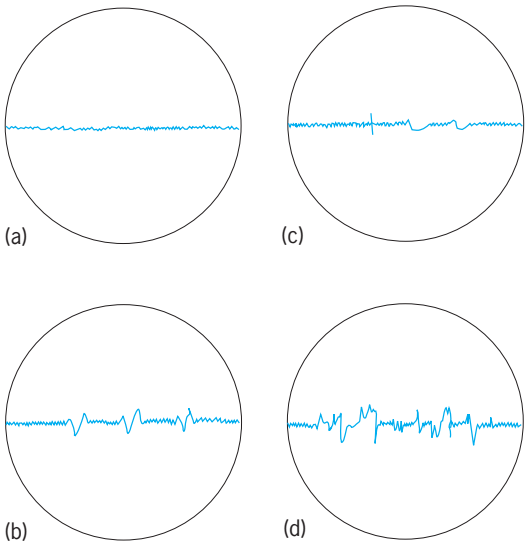
4. The photovoltaic effect, in which the absorption of light (or, more generally, electromagnetic radiation) in a semiconductor produces an emf. This is widely used for scientific purposes in radiation detectors and also, increasingly, for the generation of electric power from the Sun's radiation. *See* PHOTOVOLTAIC EFFECT; RADIOMETRY; SOLAR CELL.

5. The piezoelectric effect, in which the application of mechanical stress to certain types of crystal generates an emf. There are applications in sound recording, in ultrasonics, and in various types of measurement transducer. *See* MICROPHONE; PIEZOELECTRICITY; TRANSDUCER; ULTRASONICS. A. Earle Bailey

Electromyography

The detection and recording of electrical activity generated by muscle fibers. The basic elements of motor control in the body are the motor units which comprise motor neurons in the brainstem or spinal cord, their axons, and from ten to several hundred muscle fibers supplied by each motor neuron. Motor units vary in the size and properties of their motoneurons, the sizes and conduction velocities of their axons, the morphology of their nerve muscle junctions, and the structure and physiological properties of the muscle fibers supplied by each motor neuron. Voluntary activity recruits motor units in an orderly manner, beginning with smaller, twitch units with more slowly conducting axons recruited in weak contractions and progressing to much-harder-to-recruit, larger and faster-twitch motor units with more rapidly conducting axons.

Impulses originating in single motoneurons in response to various command signals from the central nervous system conduct to the periphery of the unit, normally causing all the muscle fibers in the unit to discharge. The electrical activity generated by the more or less synchronous discharges of all the muscle fibers in the unit may be detected by recording electrodes on the skin surface or by needles inserted into the muscle (see **illus.**). Such potentials are called motor-unit action potentials and reflect the electrical activity generated by the whole motor unit.



Record of synchronous and asynchronous muscle discharge. (*a*) Normal resting muscle. (*b*) Partial contraction. (*c*) Degenerating muscle. (*d*) Dystrophic muscle.

Diseases affecting motor neurons are sometimes accompanied by spontaneous discharges of the axons (fasciculations) which, if near the skin, may be visible as twitches in the muscle. Additionally, degeneration of motor axons may leave some muscle fibers deprived of their normal innervation, some of which spontaneously fire. Such single muscle-fiber discharges are called fibrillations and are readily detected for diagnostic purposes by needle electrodes inserted into the muscle.

Electromyography may also be used to study primary muscle diseases such as the muscular dystrophies, and a wide variety of other metabolic inflammatory and congenital myopathies affecting the muscle fibers rather than motor neurons or their axons. *See* BIOPOTENTIALS AND IONIC CURRENTS; ELECTRODIAGNOSIS. William F. Brown

Electron

The negatively charged constituent of ordinary matter, responsible for the chemical properties of matter, and the carrier of electricity. The electron is the lightest known particle that possesses an electric charge. Its rest mass m_e is 9.1×10^{-31} kg (2.0×10^{-30} lb), about 1/1836 of the mass of the proton or neutron, which are, respectively, the positively charged and neutral constituents of ordinary matter. The rest mass of an electron can also be expressed as 0.511 MeV/c^2, where c is the speed of light. Electrons were discovered in 1895 by J. J. Thomson

in the form of cathode rays. The electron was the first elementary particle to be identified. *See* ATOMIC STRUCTURE AND SPECTRA; CATHODE RAYS; ELECTRIC CHARGE; ELECTRONVOLT; ELEMENTARY PARTICLE; NUCLEAR STRUCTURE.

Charge. The charge of the electron is $e = 1.602 \times 10^{-19}$ coulomb. The sign of the electron's charge is negative by convention, and that of the equally charged proton is positive. This is a somewhat unfortunate convention, because the flow of electrons in a metallic conductor is thus opposite to the conventional direction of the current. *See* CONDUCTION (ELECTRICITY).

The historical method of measuring the charge of the electron is the celebrated oil drop experiment of R. A. Millikan in 1909, in which the charges of droplets of oil in air are measured by finding the electric field which balances each drop against its weight. The weight of each drop is determined by observing its rate of free fall through the air, and using Stokes' formula for the viscous drag on a slowly moving sphere. The charges thus measured are integral multiples of e. For more precise values of e and m_e, *see* FUNDAMENTAL CONSTANTS.

Electrons and matter. Electrons are emitted in radioactivity (as beta rays) and in many other decay processes; for instance, the ultimate decay products of all mesons are electrons, neutrinos, and photons, the meson's charge being carried away by the electrons. The electron itself is completely stable, according to all available evidence. Electrons contribute the volume to ordinary matter; the volume of an atom is nearly all occupied by the cloud of electrons surrounding the nucleus, which occupies only about 10^{-13} of the atom's volume. Virtually all the mass of an atom, however, is in the nucleus. The chemical properties of ordinary matter are determined by the electron cloud. *See* BETA PARTICLES; CHEMICAL BONDING; QUANTUM CHEMISTRY; RADIOACTIVITY; VALENCE.

The electron obeys the Fermi-Dirac statistics, and for this reason is often called a fermion. One of the primary attributes of matter, impenetrability, results from the fact that the electron, being a fermion, obeys the Pauli exclusion principle; the world would be completely different if the lightest charged particle were a boson, that is, a particle that obeys Bose-Einstein statistics. *See* BOSE-EINSTEIN STATISTICS; EXCLUSION PRINCIPLE; FERMI-DIRAC STATISTICS.

Spin. Every elementary particle possesses an intrinsic angular momentum called its spin. The spin of the electron has the magnitude $^1/_2\hbar$, where $\hbar$ is Planck's constant h divided by 2π. An electron thus has two spin states: spin up and spin down. To describe this, the nonrelativistic wave function is a two-component function, that is, a vector in a two-dimensional spin space; the two linearly independent vectors represent the two possible spin states. In 1928 P. A. M. Dirac derived the corresponding relativistic wave equation (Dirac equation). Here, the electron wave function must have four components; correspondingly, for a wave of given momentum, there are four internal states. In addition to the two-valued spin coordinate, there is an energy coordinate; that is, for a momentum p, the energy can be $\pm\sqrt{(m_e c^2)^2 + (pc)^2}$, where c is the velocity of light. *See* ANGULAR MOMENTUM; ELECTRON SPIN; NONRELATIVISTIC QUANTUM THEORY; RELATIVISTIC QUANTUM THEORY; SPIN (QUANTUM MECHANICS).

Positron. The negative energy states were at first an embarrassment, for they extend downward indefinitely; an electron would cascade indefinitely downward in energy, radiating photons. Electrons obey the exclusion principle, however, and this conclusion can be avoided by assuming that in empty space all the negative energy states are already occupied, and so exclude any more electrons. A new process is possible now; such a negative energy electron, by absorbing energy, can go to a positive energy state. This leaves behind a hole in the sea of negative-energy electrons; the hole has a positive energy, because it represents a missing negative energy. In fact, such a hole has all the properties of an electron except that it appears to have a positive charge (because it represents a missing negative charge). This particle is the positron, discovered in 1932 by C. D. Anderson in a cloud-chamber study of cosmic radiation. *See* POSITRON.

The electron and the positron are on an equal footing; if one started with a Dirac wave equation for the positron, identifying electrons with holes in the negative-energy positron states, one would get an equivalent theory. The apparent dissymmetry inherent in the construction of the hole theory disappears from the results when the total charge, energy, and momentum of empty space is defined to be zero; actually, the dissymmetry can be avoided at all stages in the formalism of quantum field theory. *See* ANTIMATTER; QUANTUM FIELD THEORY; SYMMETRY LAWS (PHYSICS).

Magnetic moment. The electron has magnetic properties by virtue of its orbital motion about the nucleus of its atom and its rotation (spin) about its own axis. The intrinsic magnetic moment of the electron is predicted by the Dirac equation to be the value (Bohr magneton) shown in the equation below. The actual moment μ differs from μ_D

$$\mu_D = \frac{e\hbar}{2m_e}$$

by a small amount (anomalous magnetic moment) due to electromagnetic radiative corrections: $\mu = 1.0011\ \mu_D$. This theoretical value, calculated using renormalized quantum field theory, agrees with the experimental value. *See* MAGNETON; QUANTUM ELECTRODYNAMICS.

Other leptons. The electron is the lightest of a family of elementary particles, the leptons. The other known charged leptons are the muon and the tau. These three particles differ only (as far as known) in mass; they have the same spin, charge, strong interactions (namely, none), and weak interactions. In a weak interaction a charged lepton is either unchanged (neutral weak current reaction) or changed

(charged weak current reaction) into an uncharged lepton, that is, a neutrino. In the latter case, each charged lepton (electron, muon, or tau) is seen to change only into the corresponding neutrino (ν_e, ν_μ, or ν_τ). *See* LEPTON; NEUTRINO; WEAK NUCLEAR INTERACTIONS.

Electrons are responsible for many phenomena of importance. Electrons moving through a conducting material constitute electricity. There is a magnetic field surrounding a current-carrying conductor, which makes possible electric motors and generators. Moving electrons are generally responsible for the production of light and electromagnetic radiation; examples include radio and television waves emitted by electrons moving in an antenna, microwaves generated in a magnetron in a microwave oven, and x-rays emitted by electrons passing through matter or electrons deflected in a magnetic field. Electron spin is also responsible for the behavior of magnetic materials. *See* ANTENNA (ELECTROMAGNETISM); BREMSSTRAHLUNG; ELECTRIC CURRENT; ELECTRICITY; ELECTROMAGNETIC RADIATION; GENERATOR; LIGHT; MAGNETIC MATERIALS; MAGNETISM; MAGNETRON; MOTOR; SYNCHROTRON RADIATION.

Electrons are responsible for chemical bonding in molecules. Additionally, the number of protons or electrons in an atom determines the elemental species of the atom, and thus the ranking of elements by atomic number (the number of electrons or protons in the atom) in the periodic table. *See* ATOMIC NUMBER; CHEMICAL BONDING; ELEMENT (CHEMISTRY); PERIODIC TABLE.

Electrons are a valuable probe of matter, as in an electron microscope. Electrons at high energies have a very short wavelength, and microscopes with atomic resolution can be constructed. *See* ELECTRON DIFFRACTION; ELECTRON MICROSCOPE.

Other information. It would be difficult to list all the articles wherein the electron forms an integral part of the discussion. The following articles and those referenced in the preceding discussion are intended to be merely a representative sample. *See* BAND THEORY OF SOLIDS; CHARGED PARTICLE OPTICS; COMPTON EFFECT; ELECTRON CAPTURE; ELECTRON CONFIGURATION; ELECTRON EMISSION; ELECTRON MOTION IN VACUUM; ELECTRONICS; EXCHANGE INTERACTION; FREE-ELECTRON THEORY OF METALS; MAGNETISM; PARTICLE ACCELERATOR; QUANTUM MECHANICS; RELATIVISTIC ELECTRODYNAMICS.

Charles J. Goebel; Alfred S. Schlachter

Electron affinity

The amount of energy release when an electron e at rest is captured by a species M, producing the negative ion M^-. The electron affinity of a species M can also be thought of as the ionization potential of the negative ion M^-. In these definitions, it is assumed that both M and M^- are in the lowest electronic, vibrational, and rotational state.

If the electron affinity of M is negative, the M^- ion is unstable with respect to decomposition into $M + e$. Most atoms have positive electron affinities, even though there is no net Coulomb attraction between the electron and the atom until the electron is close enough to be "a part of the atom." The simple rules of chemical valence provide a qualitative guide to the magnitude of electron affinities. Thus the noble gases, which have a filled outer electronic shell and are chemically inert, are not capable of binding an additional electron to form a negative ion. The largest electron affinities are possessed by the halogens, atoms which require only one additional electron to fill the valence shell. *See* HALOGEN ELEMENTS.

The major exception to this concept is that multiply charged negative ions—for example, O^{2-}, one of many such ions which are stable in solution—have not been observed in the gas phase. The ability to place more than one additional electron in the valence shell of a neutral atom or molecule arises from many particle electrostatic interactions with the ion's environment: either the solvent shell surrounding the ion in liquid solutions or the amorphous or crystalline region surrounding the ion in solids. *See* VALENCE.

Experimental methods. Accurate ionization potentials of the elements were known for a number of years before comparable data for electron affinities of the elements became available. In order to determine the ionization potential of an element, the element is vaporized, placed in an optical spectrometer, and observed for the onset of photoabsorption corresponding to the photoionization process, $h\nu + M \rightarrow M^+ + e$. The photon energy corresponding to the threshold wavelength for this process is the ionization energy of the species M. The analogous method for determination of an electron affinity is through observation of the threshold of the very similar photodetachment reaction, $h\nu + M^- \rightarrow M + e$. Again, the threshold wavelength for this process corresponds to the electron affinity of the species M. Unfortunately, it has not proved possible to produce the sufficiently large densities of negative ions required to utilize photoabsorption spectrometers to observe directly the threshold for the photodetachment process. Consequently, determination of accurate electron affinities lagged far behind the determination of accurate ionization potentials. However, an ion velocity modulation scheme has been devised that permits the direct observation of vibrational photoabsorption in ions; it is not generally applicable to the electron detachment process required to determine electron affinities.

The two major experimental advances that have enabled accurate electron affinity determinations have been the development of ion-beam techniques affording the detection of photoabsorption by the monitoring of photodetachment products and the availability of intense light sources in the form of lasers.

In experiments to determine electron affinities, negative ions are formed in an electrical

1 H 0.754201							2 He <0
3 Li 0.6180	4 Be <0	5 B 0.277	6 C 1.2629	7 N ≈0	8 O 1.4611215	9 F 3.399	10 Ne <0
11 Na 0.54793	12 Mg <0	13 Al 0.441	14 Si 1.385	15 P 0.7465	16 S 2.077120	17 Cl 3.61272	18 Ar <0
19 K 0.50147	20 Ca 0.043	31 Ga 0.30	32 Ge 1.233	33 As 0.81	34 Se 2.02069	35 Br 3.365	36 Kr <0
37 Rb 0.48592	38 Sr 0.10	49 In 0.30	50 Sn 1.112	51 Sb 1.07	52 Te 1.9708	53 I 3.0591	54 Xe <0
55 Cs 0.47163	56 Ba 0.18	81 Tl 0.2	82 Pb 0.364	83 Bi 0.946	84 Po 1.9	85 At 2.8	86 Rn <0

21 Sc 0.188	22 Ti 0.079	23 V 0.525	24 Cr 0.677	25 Mn <0	26 Fe 0.151	27 Co 0.662	28 Ni 1.156	29 Cu 1.235	30 Zn <0
39 Y 0.307	40 Zr 0.426	41 Nb 0.893	42 Mo 0.746	43 Tc 0.55	44 Ru 1.05	45 Rh 1.137	46 Pd 0.557	47 Ag 1.302	48 Cd <0
57 La 0.5	72 Hf ≈0	73 Ta 0.322	74 W 0.815	75 Re 0.15	76 Os 1.1	77 Ir 1.565	78 Pt 2.128	79 Au 2.30863	80 Hg <0

Periodic table showing the best values for the electron affinities of the elements. All values are reported in electronvolts. The value <0 implies that the negative ion is unstable with respect to decomposition to an electron and a neutral atom. In general, the last decimal place in each value is uncertain. (*After H. Hotop and W. C. Lineberger, Binding energies in atomic negative ions: II, J. Phys. Chem. Ref. Data, 14:731–750, 1985*)

discharge, extracted through an aperture into a high-vacuum region, formed into a negative-ion beam, mass-analyzed, and intersected by an intense laser beam. The laser-beam–negative-ion-beam intersection takes place in a high-vacuum region where no collisions are likely. The occurrence of a photodetachment event is determined by detection of the photodetached electron.

Two experimental methods evolved which produce accurate electron affinities. In the first method the laser is tunable, and a search is made for the wavelength corresponding to the threshold for the photodetachment process. In this case the electron affinity is given directly by the threshold wavelength, and experimental methods have been developed that provide accuracies of 10^{-8} eV. In the second type of experiment, called photoelectron spectroscopy, a fixed-frequency laser (of known photon energy) is employed, and the electrostatic fields are used to determine the kinetic energy of the ejected electron. From simple energy conservation arguments, the electron affinity is then given by the photon energy less the kinetic energy of the ejected electron. This latter technique is quite general, but is limited in accuracy by the resolution of the electron energy analyzer (approximately 3×10^{-3} eV for state-of-the-art analyzers). *See* LASER; LASER SPECTROSCOPY; MASS SPECTROMETRY.

Periodic trends. These laser photodetachment studies dramatically improved knowledge of the electron affinities of the elements. The **illustration** is a periodic table showing values of the electron affinities of the elements. Most of the data shown here were obtained by using laser photodetachment methods. The periodic trends in electron affinities and the qualitative effects described earlier are immediately apparent. In addition, a number of more subtle trends are observable. For example, while it might be expected that filled-shell species such as the rare gases would not be capable of binding an additional electron, the illustration shows that half-filled shells (for example, N and P) also exhibit small or negative electron affinities. Again, this effect is the result of the fact that a half-filled valence shell is spherically symmetric and behaves somewhat as though it were a filled shell. A similar situation is also seen for half-filled *d* shells, as is the case for Mn, Tc, and Re in the transition metals.

These same techniques have provided a number of accurate electron affinity determinations for molecules and free radicals, and new insight into the structural and chemical properties of ions in the gas phase. *See* IONIZATION POTENTIAL; PERIODIC TABLE.

W. C. Lineberger

Bibliography. M. T. Bowers (ed.), *Gas Phase Ion Chemistry 3*, 1984; H. Hotop and W. C. Lineberger, Binding energies in atomic negative ions: II, *J. Phys. Chem. Ref. Data*, 14:731–750, 1985; I. N. Levine, *Physical Chemistry*, 4th ed., 1994; R. J. Saykally, Infrared laser spectroscopy of molecular ions, *Science*, 239:157–161, 1988.

Electron capture

A term that refers to two processes, one atomic and the other nuclear, in both of which an electron is captured: by a charged particle (an ion) passing through matter, when an electron is captured into an atomic orbital (in the atomic case); or by a nucleus (in the nuclear case), which results in radioactive decay of the nucleus.

Atomic process. The capture and loss of electrons is an important mechanism for slowing down fast positive ions (atoms with fewer electrons than their nuclear charge) passing through matter. Electron capture is significant generally when the velocity of the ion is similar to that of the target electrons to be captured. Fast ions passing through matter thus generally lose much of their energy near the end of their range passing through matter, as their velocity becomes similar to that of the bound target electrons.

An example of electron capture is shown in **Fig. 1**, which shows the cross section (the likelihood) for electron capture by a proton passing through lithium vapor. The reaction is $H^+ + Li \rightarrow H^0 + Li^+$ (or possibly $Li^{2+} + e^-$). The cross section (the likelihood of capturing an electron) becomes very small when the proton energy (which is proportional to the square

of the velocity) becomes large. *See* CHARGED PARTICLE BEAMS.

Nuclear process. The nuclear capture of electrons (K capture) occurs by a process quite different from atomic capture and leads to a nuclear reaction. An electron is captured by the nucleus; the electron is taken from the K shell (the electron shell nearest the nucleus) of the atom. The probability of electron capture by the nucleus depends on the amount of time the electrons spend near the nucleus, that is, on the size of the electron wave function at the nuclear center. Since to a very good approximation only electrons with zero orbital angular momentum have a wave function that is finite at the center, K capture generally occurs for capture of a K-shell electron. However, L capture is also possible. *See* RADIOACTIVITY.

Electron capture (K capture) generally occurs for a nucleus with too many protons (or too few neutrons). A K-shell electron from the atom is captured by the nucleus, creating a neutron and a neutrino, resulting in a more stable nucleus. The result of essentially transforming a proton into a neutron in the nucleus is to reduce the atomic number by one, resulting in a different element (since the number of electrons in an atom is equal to the number of protons in the nucleus, and the number of electrons or protons determines the element). Atomic mass is essentially unchanged by K capture, as the mass of a proton and neutron are essentially the same. The atom is missing a K-shell electron as a result of K capture; it stabilizes by emission of an x-ray or an Auger electron: an electron from an outer shell fills the vacancy in the K shell of the atom. Thus, K capture is a simultaneous atomic and nuclear event. *See* AUGER EFFECT.

K capture is favored when there is not enough energy for the nucleus to emit a positron, which requires 1.022 MeV. An example of a nucleus without sufficient energy to emit a positron is rubidium-83, which decays to krypton-83 by K capture. *See* POSITRON.

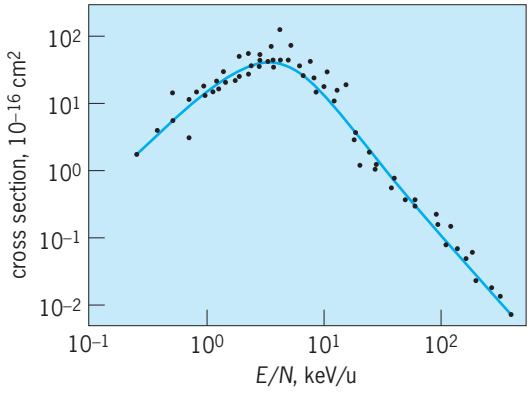


Fig. 1. Cross section (units of 10^{-16} cm^2) for capture of an electron by a proton passing through lithium vapor. Energy per nucleon (E/N) is in units of kiloelectronvolts per nucleon (keV/u). Data points from experiments and a curve fitted to the data are shown. (*After T. J. Morgan et al., Charge transfer of hydrogen ions and atoms in metal vapors, J. Phys. Chem. Ref. Data, 14:971–1040, 1985*).

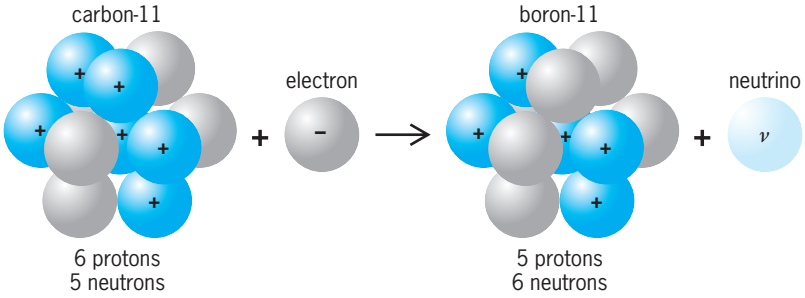


Fig. 2. Example of K capture: The reaction carbon-11 (6 protons and 5 neutrons) + electron → boron-11 (5 protons and 6 neutrons) + neutrino. (*Jefferson Lab, http://education.jlab.org/glossary/electroncapture.html*)

An example of K capture is shown in **Fig. 2**. Both carbon-11 and boron-11 have the same atomic mass (11); however, boron (atomic number 5) is one element before carbon (atomic number 6) in the periodic table.

Alfred S. Schlachter

Electron configuration

The orbital arrangement of an atom's electrons. Negatively charged electrons are attracted to a positively charged nucleus to form an atom or ion. Although such bound electrons exhibit a high degree of quantum-mechanical wavelike behavior, there still remain particle aspects to their motion. Bound electrons occupy orbitals that are somewhat concentrated in spatial shells lying at different distances from the nucleus. As the set of electron energies allowed by quantum mechanics is discrete, so is the set of mean shell radii. Both these quantized physical quantities are primarily specified by integral values of the principal, or total, quantum number n. The full electron configuration of an atom is correlated with a set of values for all the quantum numbers of each and every electron. In addition to n, another important quantum number is l, an integer representing the orbital angular momentum of an electron in units of $h/2\pi$ where h is Planck's constant. The values 1, 2, 3, 4, 5, 6, 7 for n and 0, 1, 2, 3, for l together suffice to describe the electron configurations of all known normal atoms and ions, that is, those that have their lowest possible values of total electronic energy. The first seven shells are also given the letter designations K, L, M, N, O, P, and Q, respectively. Electrons with l equal to 0, 1, 2, and 3 are designated s, p, d, and f, respectively. *See* QUANTUM MECHANICS; QUANTUM NUMBERS.

In any configuration the number of equivalent electrons (same n and l) is indicated by an integral exponent (not a quantum number) attached to the letters s, p, d, and f. According to the Pauli exclusion principle the maximum is s^2, p^6, d^{10}, and f^{14}. For example, the configuration $1s^2 2s^2 2p^6 3s$ of the

Distribution of electrons in the atoms*

Element and atomic number		K	L		M			N				O				Ground term	Ionization potential, eV
		1, 0 1*s*	2, 0 2*s*	2, 1 2*p*	3, 0 3*s*	3, 1 3*p*	3, 2 3*d*	4, 0 4*s*	4, 1 4*p*	4, 2 4*d*	4, 3 4*f*	5, 0 5*s*	5, 1 5*p*	5, 2 5*d*	5, 3 5*f*		
H	1	1	—	—	—	—	—	—	—	—	—	—	—	—	—	$^2S_{1/2}$	13.5984
He	2	2	—	—	—	—	—	—	—	—	—	—	—	—	—	$^1S_0$	24.5874
Li	3	2	1	—	—	—	—	—	—	—	—	—	—	—	—	$^2S_{1/2}$	5.3917
Be	4	2	2	—	—	—	—	—	—	—	—	—	—	—	—	$^1S_0$	9.3227
B	5	2	2	1	—	—	—	—	—	—	—	—	—	—	—	$^2P^o_{1/2}$	8.2980
C	6	2	2	2	—	—	—	—	—	—	—	—	—	—	—	$^3P_0$	11.2603
N	7	2	2	3	—	—	—	—	—	—	—	—	—	—	—	$^4S^o_{3/2}$	14.5341
O	8	2	2	4	—	—	—	—	—	—	—	—	—	—	—	$^3P_2$	13.6181
F	9	2	2	5	—	—	—	—	—	—	—	—	—	—	—	$^2P^o_{3/2}$	17.4228
Ne	10	2	2	6	—	—	—	—	—	—	—	—	—	—	—	$^1S_0$	21.5645
Na	11	Neon configuration			1	—	—	—	—	—	—	—	—	—	—	$^2S_{1/2}$	5.1391
Mg	12				2	—	—	—	—	—	—	—	—	—	—	$^1S_0$	7.6462
Al	13				2	1	—	—	—	—	—	—	—	—	—	$^2P^o_{1/2}$	5.9858
Si	14				2	2	—	—	—	—	—	—	—	—	—	$^3P_0$	8.1517
P	15				2	3	—	—	—	—	—	—	—	—	—	$^4S^o_{3/2}$	10.4867
S	16				2	4	—	—	—	—	—	—	—	—	—	$^3P_2$	10.3600
Cl	17				2	5	—	—	—	—	—	—	—	—	—	$^2P^o_{3/2}$	12.9676
Ar	18				2	6	—	—	—	—	—	—	—	—	—	$^1S_0$	15.7596
K	19	Argon configuration					—	1	—	—	—	—	—	—	—	$^2S_{1/2}$	4.3407
Ca	20						—	2	—	—	—	—	—	—	—	$^1S_0$	6.1132
Sc	21						1	2	—	—	—	—	—	—	—	$^2D_{3/2}$	6.5615
Ti	22						2	2	—	—	—	—	—	—	—	$^3F_2$	6.8281
V	23						3	2	—	—	—	—	—	—	—	$^4F_{3/2}$	6.7462
Cr	24						5	1	—	—	—	—	—	—	—	$^7S_3$	6.7665
Mn	25						5	2	—	—	—	—	—	—	—	$^6S_{5/2}$	7.4340
Fe	26						6	2	—	—	—	—	—	—	—	$^5D_4$	7.9024
Co	27						7	2	—	—	—	—	—	—	—	$^4F_{9/2}$	7.8810
Ni	28						8	2	—	—	—	—	—	—	—	$^3F_4$	7.6398
Cu	29						10	1	—	—	—	—	—	—	—	$^2S_{1/2}$	7.7264
Zn	30						10	2	—	—	—	—	—	—	—	$^1S_0$	9.3942
Ga	31						10	2	1	—	—	—	—	—	—	$^2P^o_{1/2}$	5.9993
Ge	32						10	2	2	—	—	—	—	—	—	$^3P_0$	7.8994
As	33						10	2	3	—	—	—	—	—	—	$^4S^o_{3/2}$	9.7886
Se	34						10	2	4	—	—	—	—	—	—	$^3P_2$	9.7524
Br	35						10	2	5	—	—	—	—	—	—	$^2P^o_{3/2}$	11.8138
Kr	36						10	2	6	—	—	—	—	—	—	$^1S_0$	13.9996
Rb	37	Krypton configuration								—	—	1	—	—	—	$^2S_{1/2}$	4.1771
Sr	38									—	—	2	—	—	—	$^1S_0$	5.6949
Y	39									1	—	2	—	—	—	$^2D_{3/2}$	6.2173
Zr	40									2	—	2	—	—	—	$^3F_2$	6.6339
Nb	41									4	—	1	—	—	—	$^6D_{1/2}$	6.7589
Mo	42									5	—	1	—	—	—	$^7S_3$	7.0924
Tc	43									5	—	2	—	—	—	$^6S_{5/2}$	7.28
Ru	44									7	—	1	—	—	—	$^5F_5$	7.3605
Rh	45									8	—	1	—	—	—	$^4F_{9/2}$	7.4589
Pd	46									10	—	—	—	—	—	$^1S_0$	8.3369

ground or lowest energy state of sodium means that there are two electrons with $n = 1$, $l = 0$; two with $n = 2, l = 0$; six with $n = 2, l = 1$; and one with $n = 3$, $l = 0$. Higher values of n and l can be achieved by excitation of normal atoms, either through photon absorption or by particle impact, temporarily moving one or more electrons to unfilled shells of larger radii and increasing the atom's electronic energy. Such excited electronic configurations are inherently unstable. Excited electrons ultimately fall back down to their normal locations closer to the nucleus, a process accompanied by the emission of one or more quanta of electromagnetic radiation (photons). While the atom is excited, the partially depleted shells are said to possess vacancies. *See* EXCLUSION PRINCIPLE.

An electron configuration is categorized as having even or odd parity, according to whether the sum of p and f electrons is even or odd. Strong spectral lines result only from transitions between configurations of unlike parity. *See* PARITY (QUANTUM MECHANICS).

Insofar as they are known from spectroscopic investigations, the electron configurations characteristic of the normal or ground states of the first 102 chemical elements are shown in the **table**. In the next-to-last column of the table, the spectral term of the energy level with lowest total electronic energy is shown. For all elements except protactinium (Pa), uranium (U), and neptunium (Np), an *LS*-coupling term is given. Here, the main part of the term symbol is a capital letter, *S*, *P*, *D*, *F*, and so on, that

Distribution of electrons in the atoms* (cont.)

Element and atomic number		Configuration of inner shells	N	O				P			Q	Ground term	Ionization potential, eV
			4, 3 4*f*	5, 0 5*s*	5, 1 5*p*	5, 2 5*d*	5, 3 5*f*	6, 0 6*s*	6, 1 6*p*	6, 2 6*d*	7, 0 7*s*		
Ag	47	Palladium configuration	—	1	—	—	—	—	—	—	—	$^2S_{1/2}$	7.5762
Cd	48		—	2	—	—	—	—	—	—	—	$^1S_0$	8.9938
In	49		—	2	1	—	—	—	—	—	—	$^2P^\circ_{1/2}$	5.7864
Sn	50		—	2	2	—	—	—	—	—	—	$^3P_0$	7.3439
Sb	51			2	3	—	—	—	—	—	—	$^4S^\circ_{3/2}$	8.6084
Te	52		—	2	4	—	—	—	—	—	—	$^3P_2$	9.0096
I	53		—	2	5	—	—	—	—	—	—	$^2P^\circ_{3/2}$	10.4513
Xe	54		—	2	6	—	—	—	—	—	—	$^1S_0$	12.1298
Cs	55	The shells 1*s* to 4*d* contain 46 electrons	—	The shells 5*s* to 5*p* contain 8 electrons		—	—	1	—	—	—	$^2S_{1/2}$	3.8939
Ba	56		—			—	—	2	—	—	—	$^1S_0$	5.2117
La	57		—			1	—	2	—	—	—	$^2D_{3/2}$	5.5769
Ce	58		1			1	—	2	—	—	—	$^1G^\circ_4$	5.5387
Pr	59		3			—	—	2	—	—	—	$^4I^\circ_{9/2}$	5.473
Nd	60		4			—	—	2	—	—	—	$^5I_4$	5.5250
Pm	61		5			—	—	2	—	—	—	$^6H^\circ_{5/2}$	5.582
Sm	62		6			—	—	2	—	—	—	$^7F_0$	5.6437
Eu	63		7			—	—	2	—	—	—	$^8S^\circ_{7/2}$	5.6704
Gd	64		7			1	—	2	—	—	—	$^9D^\circ$	6.1498
Tb	65		9			—	—	2	—	—	—	$^6H^\circ_{15/2}$	5.8638
Dy	66		10			—	—	2	—	—	—	$^5I_8$	5.9389
Ho	67		11			—	—	2	—	—	—	$^4I^\circ_{15/2}$	6.0215
Er	68		12			—	—	2	—	—	—	$^3H_6$	6.1077
Tm	69		13			—	—	2	—	—	—	$^2F^\circ_{7/2}$	6.1843
Yb	70		14			—	—	2	—	—	—	$^1S_0$	6.2542
Lu	71		14			1	—	2	—	—	—	$^2D_{3/2}$	5.4259
Hf	72	The shells 1*s* to 5*p* contain 68 electrons				2	—	2	—	—	—	$^3F_2$	6.8251
Ta	73					3	—	2	—	—	—	$^4F_{3/2}$	7.5496
W	74					4	—	2	—	—	—	$^5D_0$	7.8640
Re	75					5	—	2	—	—	—	$^6S_{5/2}$	7.8335
Os	76					6	—	2	—	—	—	$^5D_4$	8.28
Ir	77					7	—	2	—	—	—	$^4F_{9/2}$	9.02
Pt	78					9	—	1	—	—	—	$^3D_3$	8.9588
Au	79	The shells 1*s* to 5*d* contain 78 electrons					—	1	—	—	—	$^2S_{1/2}$	9.2255
Hg	80						—	2	—	—	—	$^1S_0$	10.4375
Tl	81						—	2	1	—	—	$^2P^\circ_{1/2}$	6.1082
Pb	82						—	2	2	—	—	$^3P_0$	7.4167
Bi	83						—	2	3	—	—	$^4S^\circ_{3/2}$	7.2855
Po	84						—	2	4	—	—	$^3P_2$	8.414
At	85						—	2	5	—	—	$^2P^\circ_{3/2}$	
Rn	86						—	2	6	—	—	$^1S_0$	10.7485
Fr	87						—	2	6	—	1	$^2S_{1/2}$	4.0727
Ra	88						—	2	6	—	2	$^1S_0$	5.2784
Ac	89						—	2	6	1	2	$^2D_{3/2}$	5.17
Th	90						—	2	6	2	2	$^3F_2$	6.3067
Pa	91						$2(^3H_4)$	2	6	1	2	$(4, 3/2)_{11/2}$	5.89
U	92						$3(^4I^\circ_{9/2})$	2	6	1	2	$(9/2, 3/2)^\circ_6$	6.1941
Np	93						$4(^5I_4)$	2	6	1	2	$(4, 3/2)_{11/2}$	6.2657
Pu	94						6	2	6	—	2	$^7F_0$	6.0260
Am	95						7	2	6	—	2	$^8S^\circ_{7/2}$	5.9738
Cm	96						7	2	6	1	2	$^9D^\circ_2$	5.9914
Bk	97						9	2	6	0	2	$^6H^\circ_{5/2}$	6.1979
Cf	98						10	2	6	0	2	$^5I_8$	6.2817
Es	99						11	2	6	0	2	$^4I_{15/2}$	6.42
Fm	100						12	2	6	0	2	$^3H_6$	6.50
Md	101						13	2	6	0	2	$^2F^\circ_{7/2}$	6.58
No	102						14	2	6	0	2	$^1S_0$	6.65

*Updated and revised with data from G. W. F. Drake (ed.), *Springer Handbook of Atomic, Molecular, and Optical Physics*, pp. 182–183, Springer, 2006.

represents the total electronic orbital angular momentum. Attached to this is a superior prefix, 1, 2, 3, 4, and so on, that indicates the multiplicity, and an anterior suffix, 0, $^1/_2$, 1, $^3/_2$, 2, $^5/_2$, and so on, that shows the total angular momentum, or *J* value, of the atom in the given state. The sign $^\circ$ above the *J* value signifies that the spectral term and electron configuration have odd parity. (For the elements Pa, U, and Np, a *jj* coupling term is given. Here, parentheses enclose two numbers, the first of which is the total angular momentum of the electrons in the 5*f* shell, while the second is the total angular momentum of the electron in the 6*d* shell. The suffixes following the parentheses are the same as in *LS* coupling. For these elements the *LS* coupling of the electrons in the 5*f* shell is also given in the column that lists the number of such electrons.)

The last column of the table presents the first

ionization potential of the atom when this has been derived from spectroscopic observations. In any atomic spectrum, two or more spectral lines with certain similar properties may form a series such that the reciprocal wavelengths $1/\lambda$ (number of waves per centimeter $= \sigma$) can be closely represented by a formula of the Rydberg type, $\sigma = L - R/(n + \mu)^2$, in which L is the limit of the series. R is called the Rydberg constant, and the principal quantum number n has successive integral values to which a constant fractional part μ is added. The second term vanishes when n approaches infinity, and the series limit is thus evaluated. This limit is usually coincident with the ground state of the ion, and is thus a measure (in wave-number units) of the energy required to remove from an atom its least firmly bound electron and transform a neutral atom into a singly charged ion. The energy required to ionize an atom is usually expressed in electronvolts (1 eV = 8065.48 wave numbers) and is called its first ionization potential. *See* ATOMIC STRUCTURE AND SPECTRA; IONIZATION POTENTIAL; RYDBERG ATOM; RYDBERG CONSTANT.

Molecules consist of two or more atoms that at least partially share one or more electrons with one another. Sets of quantum numbers specify molecular electron configurations in much the same way as for atoms. However, the physical meaning of some molecular quantum numbers is different than that for the atomic case, as is the term notation. *See* MOLECULAR STRUCTURE AND SPECTRA.

James E. Bayfield

Electron diffraction

The phenomenon associated with interference processes that occur when electrons are scattered by atoms to form diffraction patterns. The wave character of electrons is shown most strikingly, and doubtless most conclusively, by the phenomena of interference. For this reason, the diffraction of electrons presents the most obvious confirmation of quantum mechanics. Because of the dependence of the diffraction pattern on the distances between the atoms, electron diffraction is also an important tool for the study of the structure of crystals and of free molecules, analogous to the use of x-rays for these purposes. *See* X-RAY CRYSTALLOGRAPHY; X-RAY DIFFRACTION.

According to quantum theory, any particle moving with momentum $m\upsilon$ has a wavelength $\lambda = h/m\upsilon$, where h is Planck's constant. If the particle is an electron, and its velocity is the result of acceleration by a the potential difference V, this formula becomes Eq. (1), where m_0 and e are the rest mass

$$\lambda = \frac{h}{(2m_0eV)^{1/2}(1 + eV/2m_0c^2)^{1/2}} \qquad (1)$$

and charge of the electron, and c is the velocity of light. The last factor in the denominator represents the relativity correction, which is negligible at low voltages and amounts to only 5% at 100,000 V. For V in volts and λ in nanometers, Eq. (2) is a good

$$\lambda = \left(\frac{1.5}{V}\right)^{1/2} \qquad (2)$$

approximation to Eq. (1) at nonrelativistic energies. According to energy $E = eV$, two major techniques of structure analysis with electron beams are distinguished: low-energy electron diffraction (LEED) [$E \simeq$ 5–500 eV] and high-energy electron diffraction (HEED) [$E \simeq$ 5–500 keV]; medium-energy electron diffraction (MEED) [$E \simeq$ 500 eV–5 keV] is of little importance. In addition, electrons generated in condensed matter by incident electrons or x-ray photons are diffracted (in Auger electron diffraction and photoelectron diffraction). Unlike neutrons and x-rays, electrons penetrate matter only for a very short distance before they lose energy (by inelastic scattering) or are scattered elastically (diffracted). Due to the energy loss, the wavelength of an electron changes according to Eq. (2) so that it can no longer interfere with the other incident electrons. This loss of coherence occurs in condensed matter typically after mean free paths for inelastic scattering ranging from several tenths of nanometers at low energies ($E \simeq$ 50 eV) to several tens of nanometers at high energies, which determines the application range of LEED and HEED. *See* COHERENCE; DE BROGLIE WAVELENGTH; DIFFRACTION; INTERFERENCE OF WAVES; MEAN FREE PATH; NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS.

Low-Energy Electron Diffraction

LEED is used mainly for the study of the structure of single-crystal surfaces and of processes on such surfaces that are associated with changes in the lateral periodicity of the surface. A monochromatic, nearly parallel electron beam, of 10^{-4} to 10^{-3} m (4×10^{-3} to 4×10^{-2} in.) in diameter, strikes the surface, usually at normal incidence. The elastically backscattered electrons are separated from all other electrons by a retarding field and detected with a suitable movable collector or, more frequently—after acceleration to about 5 keV energy—on a hemispherical fluorescent screen with the crystal in its center (**Fig. 1**). The intensity of the diffraction spots can be measured as a function of the energy $E = eV$ of the incident electrons to obtain so-called $I(V)$ curves (**Fig. 2**).

To obtain such results, the surface must be carefully cleaned and kept in ultrahigh vacuum (UHV, pressure $\simeq 10^{-10}$ torr $\simeq 10^{-8}$ pascal) during the experiment and must be exposed to well-defined amounts of gases or vapors at low pressures if adsorption, condensation, or corrosion is to be studied. This requirement, as well as the fact that many surfaces are changed by the electron beam—for example, due to dissociation in the case of ionic crystals or due to desorption of adsorbed gases—limits the applicability of LEED. Nevertheless, the number of surfaces and surface processes that can be studied

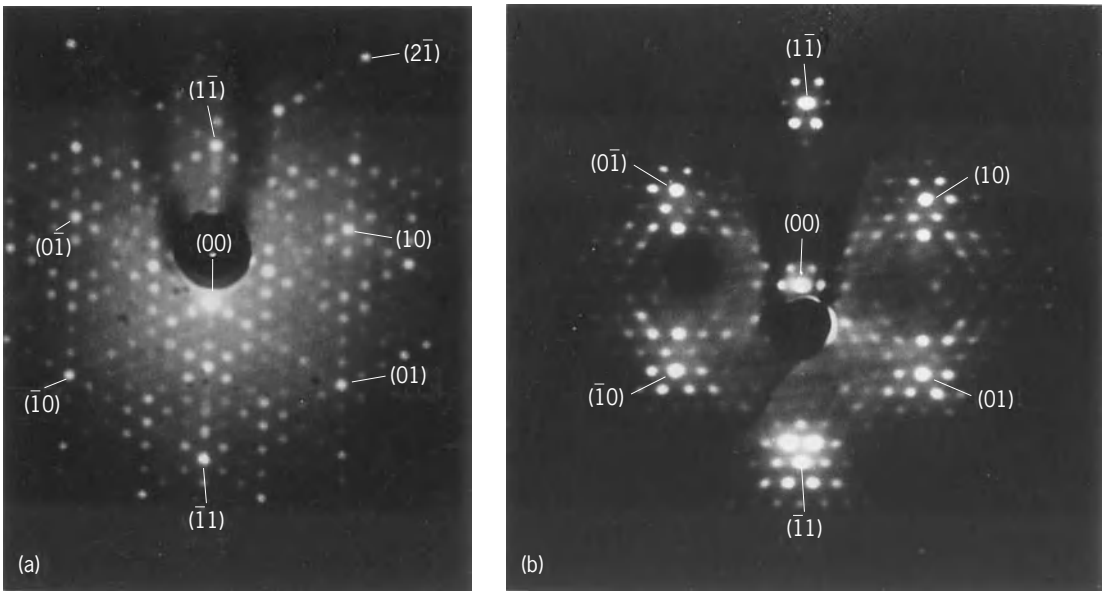


Fig. 1. LEED patterns from single-crystal surfaces. Numbers are indices (*h*, *k*) labeling diffraction spots of the surfaces. (*a*) Clean silicon (111) surface (7 × 7 structure). *E* = 85 eV. (*b*) Ordered carbon monoxide adsorption layer on a tungsten (110) surface. *E* = 120 eV.

is very large, and the main limitation of LEED is not caused by difficulties in obtaining LEED patterns but rather in evaluating them.

Interpretation of patterns. The difficulties in interpreting LEED patterns are caused by the strong elastic scattering of slow electrons by atoms. As a consequence of this, the electrons observed in the LEED pattern have been scattered several times on the average, in contrast to x-ray and neutron diffraction, in which single scattering is a good approximation. Thus evaluation of the spot intensities of LEED patterns, with the additional complications arising from the elastic scattering at the surface barrier and from inelastic scattering, requires large-scale computers. As a consequence, only the structures of surfaces whose unit meshes (two-dimensional unit cells) are small have been determined by LEED. For this reason, alternative evaluation methods have been developed in which multiple-scattering effects are reduced by proper averaging procedures.

The interpretation difficulties are considerably reduced if only the lateral periodicity of the surface, but not the location of the atoms in the unit mesh, is to be determined. The determination of the lateral periodicity of the surface structure and of the size and shape of domains with a given structure requires only the evaluation of the geometry of the pattern, that is, of the position, size, and shape of the diffraction spots but not of their intensity, and this evaluation is relatively easy. Thus the most important application of LEED is the study of surface phenomena associated with changes in lateral periodicity. A large amount of information on lateral disorder may be extracted from spot shapes and intensities. Special systems (spot profile analysis LEED, or SPALEED) have been developed for this purpose.

Surface structure. Clean surfaces of crystals frequently have the lateral periodicity that is expected from their bulk crystal structure. The analysis of $I(V)$ curves of such surfaces (Fig. 2) in general reveals that the distribution of the atoms normal to the surface in the topmost layers differs somewhat from that in the bulk. These relaxations, that is, displacements of atoms from their positions in the bulk, are only a few percent on densely packed surfaces. On more open surfaces they can amount to 10% or more and are frequently oscillatory. There are, however, also surfaces which show complicated superstructures (known as reconstructed surfaces), that is, structures whose lateral periodicity is larger than that of the substrate (Fig. 1*a*). Usually, such surfaces are analyzed by LEED

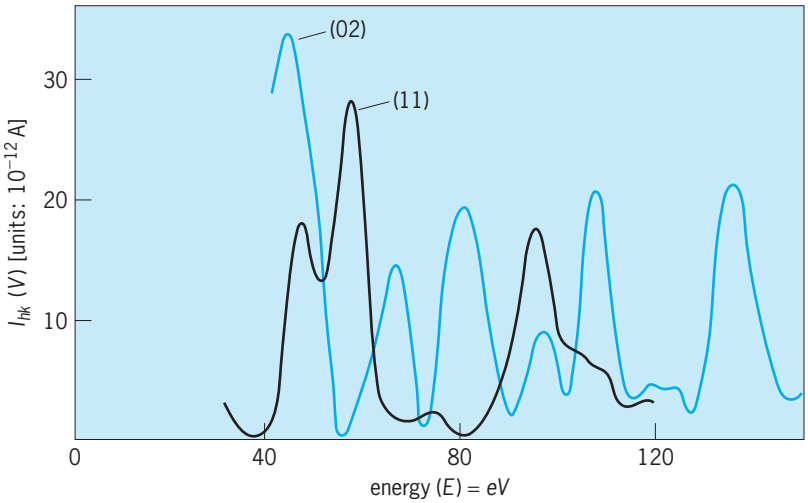


Fig. 2. *I*(*V*) curves of spots (*hk*) in the LEED pattern from a clean tungsten (110) surface. (*After G. A. Somorjai, The Structure and Chemistry of Solid Surfaces, Wiley, 1969*)

only after structural models have been proposed on the basis of other structure studies. On ideal surfaces, the decrease of the spot intensities with temperature has shown that the amplitudes of the thermal vibrations of surface atoms normal to the surface can be much larger than in the bulk. *See* CRYSTAL STRUCTURE; LATTICE VIBRATIONS.

Adsorption. Foreign atoms can interact with a clean surface in a variety of ways. The most important contribution of LEED is to the understanding of chemisorption, which precedes corrosion and, in many cases, epitaxy. Here, not only the structure of many adsorption systems, mainly of gases on metals, or metals on other metals and semiconductors, has been studied, but also the kinetics of the adsorption and desorption process as well as changes in the adsorption layer upon heating. The combination of LEED with Auger electron spectroscopy (AES) and with work-function measurements has proven particularly powerful in these studies, because such methods give the coverage and information on the location of the adsorbed atoms normal to the surface. Combining LEED with other complementary techniques such as ion scattering spectroscopy, electron energy loss spectroscopy, or photoelectron spectroscopy has become increasingly popular and can enable the elimination of ambiguities in the interpretation of many LEED results. *See* ADSORPTION; AUGER EFFECT; CORROSION; ELECTRON SPECTROSCOPY; SURFACE AND INTERFACIAL CHEMISTRY; SURFACE PHYSICS.

SPLEED. This is LEED with spin-polarized electrons. It gives information on the magnetic structure of surfaces and thin films due to the spin dependence of the diffraction of slow electrons in magnetic materials. This dependence is of particular importance in the spin-polarized version of low-energy electron microscopy (LEEM), called SPLEEM, which allows the magnetic microstructures of surfaces and thin films to be studied.

LEEM. Low-energy diffracted electrons can be used to image the surface of a specimen if it is made the cathode in an immersion objective lens of an ultrahigh-vacuum electron microscope. Because of the strong elastic backscattering of very slow electrons, the resulting LEEM images (**Fig. 3**) can be recorded at video rates. This allows the study of surface processes such as phase transitions or film growth in real time. *See* ELECTRON MICROSCOPE.

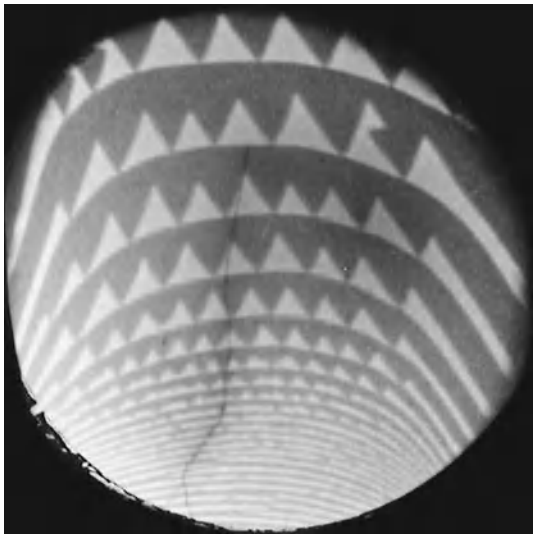


Fig. 3. LEEM image of a partially reconstructed silicon (111) surface. The bright triangles have (7 × 7) structure, the dark regions (1 × 1).

High-Energy Electron Diffraction

HEED is used mainly for the study of the structure of thin foils, films, and small particles (thickness or diameter of 10^{-9} to 10^{-6} m or 4×10^{-8} to 4×10^{-5} in.), of molecules, and also of the surfaces of crystalline materials. A monochromatic, usually nearly parallel, electron beam with a diameter of 10^{-3} to 10^{-8} m (4×10^{-2} to 4×10^{-7} in.) is incident on the target. The forward-scattered electrons (backscattering is negligible) are detected by means of a fluorescent screen, a photoplate, or some other current-sensitive detector, usually without the inelastically scattered electrons being eliminated.

Reflection HEED. The availability of UHV instruments has made it possible to study the structure of surfaces under clean conditions by reflection HEED (RHEED). Many experiments have confirmed the theoretical expectation that RHEED has a sensitivity comparable to that of LEED. Therefore, similar to LEED, RHEED can be used for the determination of the lateral arrangement of the atoms in the topmost layers of the surface (**Fig. 4**), including the structure of adsorbed layers. Although it is more convenient to deduce the periodicity of the atomic arrangement parallel to the surface from LEED patterns than from RHEED patterns, LEED frequently becomes inapplicable when the surface is rough. This usually occurs in the later stages of corrosion or in precipitation, for example, of silicon carbide on silicon, when small crystals grow on the surface. In such investigations RHEED is far superior to LEED because the fast electrons can penetrate the asperities and produce a transmission HEED (THEED) pattern. RHEED has become particularly important for thin-film growth monitoring via the specular beam intensity oscillations caused by monolayer-by-monolayer growth (**Fig. 5**). *See* CRYSTAL GROWTH.

Like LEED, RHEED gives little information on the chemical nature of the atoms that produce the diffraction pattern. Therefore, RHEED has been combined with analytical tools, such as Auger electron spectroscopy or x-ray emission spectroscopy. A second limitation inherent to both RHEED and LEED is the difficulty of determining the atomic positions normal to the surface. This is a consequence of dynamical effects and of absorption. Nevertheless, both techniques have given valuable information on corrosion, precipitation, adsorption, condensation, film growth, and other surface and interface phenomena that could not be obtained by other methods.

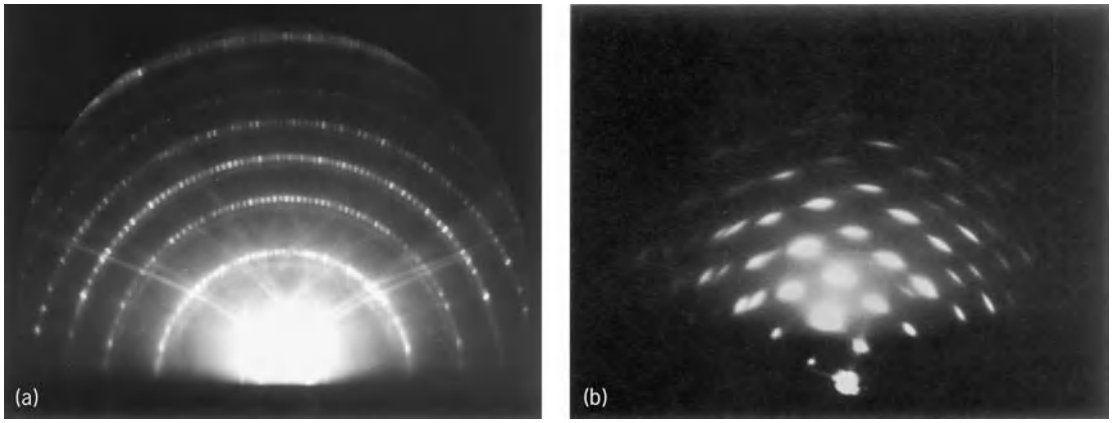


Fig. 4. RHEED pattern from (*a*) flat (111) surface of a silicon single crystal, and (*b*) rough surface of a polycrystalline evaporated film of calcium fluoride (CaF_2) with strong fiber texture, from which no LEED pattern can be obtained.

Scanning HEED. In scanning HEED (SHEED) the diffracted electrons are not recorded on photographic film but are directly measured electronically with sensitive detectors. By moving the detector across the diffraction pattern or by deflecting the diffracted electrons across a stationary detector (scanning), the intensity distribution in the diffraction pattern can be displayed quantitatively on an XY recorder. If an energy filter is put in front of the detector, the inelastically scattered electrons, which are usually not taken into account in the quantitative intensity evaluation, may be filtered out so that only the elastically scattered electrons are measured. This technique is particularly useful in transmission through polycrystalline samples, which produce a ring pattern (Debye-Scherrer diagram). It has also been used in single-crystal samples producing a spot pattern (Laue diagram) and even in reflection diffraction from surfaces. The main application of SHEED is in the study of processes which are accompanied by changes of the intensity distribution, such as the growth of thin films and annealing and corrosion processes.

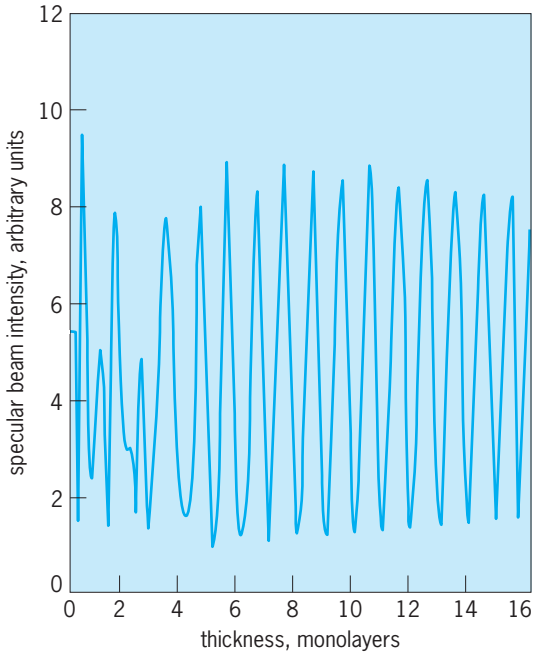


Fig. 5. RHEED specular beam intensity oscillations during molecular beam epitaxy of lead on a silicon (111) surface.

Transmission HEED. The technological importance of thin film and interface devices has led to an upsurge of thin film growth studies by conventional transmission HEED (THEED), usually combined with transmission microscopy. Information obtained this way has been mainly on the orientation of the crystallites composing the film (**Fig. 6**).

Diffraction electron microscopy. In transmission electron microscopy of crystalline specimens the main contrast mechanism is the diffraction contrast. To understand it, knowledge of the diffraction pattern and of diffraction theory is necessary. High-voltage electron microscopes (with electron energies of about 1 MeV) have become available. As a consequence, samples with thickness of the order 10^{-6} m (4×10^{-5} in.) can now be studied, as compared to the former upper thickness limit of about 10^{-7} m (4×10^{-6} in.). This has significance for metallurgy and the physics of metals and semiconductors. It is now possible to study the structure, distribution, and behavior of imperfections in metals, alloys, and semiconductors in bulk with little influence from the boundaries of the material. Other electron-microscope techniques that rely heavily on diffraction are the use of illumination systems producing tilted or conical illumination of the specimen and of objective apertures which transmit only diffracted electrons. With these techniques it has become possible to produce, by diffraction contrast, images of crystal planes 1.5×10^{-10} m (6×10^{-9} in.) apart. RHEED beams can be used for reflection electron microscopy (REM), though with considerable foreshortening due to the grazing incidence on the surface which is necessary for surface sensitivity and intensity.

Photoelectron diffraction. In photoelectron diffraction (PED), the electron wave emitted by one atom

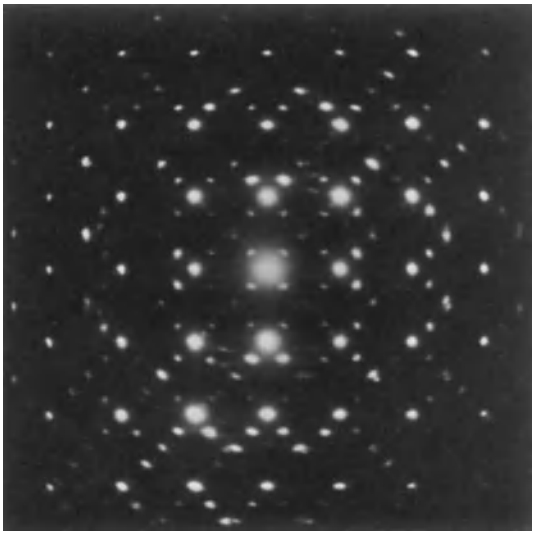


Fig. 6. THEED pattern from epitaxial gold particles on a lead sulfide (PbS) single-crystal film. (*From A. K. Green, J. Dancy, and E. Bauer, Growth of Au on PbS, PbSe, PbTe and SnTe thin film substrates, J. Vac. Sci. Technol., 10:494–502, 1973*)

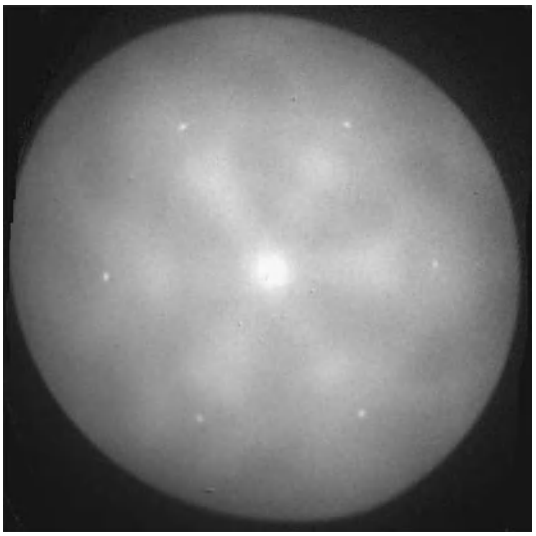


Fig. 7. PED pattern (diffuse features) with superimposed LEED pattern (sharp spots) of a lead film, five atomic layers thick, on a silicon (111) surface. Photon energy is 62 eV; electron energy, 39 eV. (*Courtesy of T. Schmidt*)

upon photoionization is diffracted by the surrounding atoms. The resulting diffraction pattern (**Fig. 7**) allows local structure analysis around the selected atom. At low energies (below several hundred electronvolts) the interpretation requires an effort comparable to that needed in LEED; at higher energies (above about 1000 eV) strong forward scattering frequently permits a simple qualitative analysis. Similar to LEEM, REM, and transmission electron microscopy (TEM), the photoemitted electron wave may also be used for imaging, a method known as x-ray photoemission electron microscopy (XPEEM) with synchrotron radiation. In this case the imaging is atom-specific, just as is PED. E. Bauer

Diffraction in Gases and Liquids

Electron diffraction in gases and liquids is similar in principle to that in solids; the differences arise from the lack in gases and liquids of any highly regular arrangement of the component atoms. In gases the low density makes it possible to study diffraction by individual atoms and molecules. The results obtained from monatomic gases represent the density of electronic charge in the atom as a function of the distance from the nucleus. The results from gaseous polyatomic molecules represent the equilibrium distances between the atomic nuclei and the average amplitudes of vibration associated with these distances. Liquids have been studied much less thoroughly, both in theory and in practice, than have gases.

It should be remembered that the structures of molecules in the gaseous state may be different from those in the liquid or solid states; for example, hydrofluoric acid in the gas forms an $(HF)_6$ ring, while in solution or liquid it forms an infinite chain (—FHFHFHFHFHF—); biphenyl in the gas is twisted with conformational equilibria, whereas in the solid state it is coplanar in one frozen conformer.

Applications. Typical questions of molecular structure studied by electron diffraction include those of configuration and size in many molecules, with special interest in the variation of chemical bond distances in related molecules (such as the 6.8-picometers or 2.7×10^{-10} in. decrease in C—F distance in the series CH_3F, CH_2F_2, CHF_3, CF_4), the variation in the angles between chemical bonds, the distinction between geometric isomers, the degree of restricted rotation around chemical bonds, and in general the relations between the geometry of molecules and their energy and chemical behavior. *See* CHEMICAL BONDING; MOLECULAR ISOMERISM; MOLECULAR STRUCTURE AND SPECTRA; STRUCTURAL CHEMISTRY.

The application of electron diffraction in liquids has been restricted because of the less precise information obtainable from molecules which lie in close contact but in irregular arrangements.

The liquids studied include certain metals (mercury, tin, tin-bismuth, and tin-aluminum alloys), paraffin oils, and thin layers of water. For monatomic molecules in the liquid state the distances between nearest neighbors are fairly uniform, whereas the separation between more distant neighbors varies considerably. Evaluation of the closest distance of approach can be made to about 10 pm (4×10^{-10} in.); major developments in the theory are still required.

The most interesting application of electron diffraction to a liquid or amorphous condensed phase has been in the examination of the surface layers on polished solids. The diffuse diffraction patterns often obtained may be the result of an amorphous arrangement in the surface or of the poor resolving power of very tiny crystalline particles. It is probable that both these states are produced in the polishing of different solids.

Theory and techniques. Structural information about gaseous molecules is obtained by having a fine beam of electrons pass through the gas and strike a photographic plate. The electrons in the beam interact with the charged particles in the atoms (electrons and nuclei) and are bent away from the original direction through varying angles, as registered in the pattern on the plate. The observed variation in the number of scattered electrons with increasing angle is interpreted as an interference effect; that is, electron scattering can be described in the language of diffracted waves for which the resultant is the sum of the component wavelets whose amplitudes and phases are influenced by the wavelength of the incident radiation and the relative positions of the scattering centers. The equivalent wavelength of the electrons, λ, is determined by their energy and is given by Eq. (1) or (2). The values of λ commonly used in diffraction by gases are in the range $7–5 \times 10^{-12}$ m ($3–2 \times 10^{-10}$ in.).

The intensity I of electrons scattered by a spherically symmetrical atom is computed by the Schrödinger wave equation; the result is given by Eq. (3), where k is a constant and f is a complex func-

$$I = k|f|^2 \tag{3}$$

tion called the atomic scattering factor. If the atomic number Z is not too large, Eq. (4) holds, where a is the Bohr radius and F is given by Eq. (5). Here $\rho(r)$ is

$$|f|^2 = [2(Z - F)/as^2]^2 \tag{4}$$

$$F = 4\pi \int_0^\infty r^2 \rho(r)[(\sin sr)/sr]dr \tag{5}$$

density of electronic charge at distance r from the nucleus, $s = 4\pi(\sin\theta)/\lambda$ with 2θ equal to the angle between the scattered electron and the orignal beam, and λ is the electron wavelength. Experimental observations on I as a function of s in monatomic gases lead to the determination of $\rho(r)$.

The intensity of electrons scattered by a collection of independent molecules having all possible orientations is given by Eq. (6), where the summations are

$$I(s) = k\sum_i [|f_i|^2 + (2/as^2)^2 S_i] + \sum_i \sum_j f_i^* f_j \int_0^\infty P_{ij}(r)[(\sin sr)/sr]\, dr \tag{6}$$

taken over all the atoms in the molecule and $(2as^2)^2 S_i$ represents the intensity of inelastic scattering by the ith atom. The double summation has a term for each pair of atoms i and j. The relative probability of finding the distance between the atom i and j at various values is represented by $P_{ij}(r)$. In principle the use of the observed intensity to determine P_{ij} for each pair of atoms in the molecule constitutes a structure determination for the molecule. The expression is simpler when the atomic motions are nearly harmonic, as in the case of CCl_4. For this molecule the double summation has only two distinct terms. The first is given by expression (7); the second for Cl—Cl is similar.

$$8f_C f_{Cl}\ \exp(-l_{CCl}^2 s^2)(\sin sr_{CCl})/sr_{CCl} \tag{7}$$

The four parameters which describe the structure are the equilibrium distances, r_{CCl} and r_{ClCl}, and the average amplitudes of vibration, l_{CCl} and l_{ClCl}. Molecular parameters such as these can be determined with high precision when a proper treatment of the scattering factors is included, and when accurate experimental intensities are obtained.

Special equipment is required to meet the conditions assumed in the scattering theory. The electron beam is accelerated with a steady voltage between 30,000 and 70,000 V, which should be constant within about 0.01%. The beam is focused by electrostatic and magnetic lenses so that it has a diameter of no more than 0.1 mm (0.004 in.) at the photographic plate. The whole path of the beam is enclosed in a high-vacuum chamber (pressure of about 10^{-5} mm Hg or 10^{-3} Pa or 10^{-7} lbf/in.) so that no appreciable electron scattering will occur in the residual air. The gas specimen is introduced in a fine jet so that the volume in which the electrons meet the gas is no more than 1.0 mm (0.04 in.) in diameter; high-speed pumping is required to remove the gas as rapidly as possible. In front of the photographic plate a rotating sector is mounted to modify the intensity of electrons reaching the plate so that the normally rapid decrease of intensity with increasing angle of scattering is leveled off in a known way and the emulsion is able to register the incident electrons over a wide range of angle.

Interpretation of results. The pattern observed is a set of light and dark circular bands whose spacings and relative intensities depend on the composition and structure of the specimen molecules. The pattern is scanned by a recording microphotometer which yields (after calibration of the photographic emulsion) a tracing or a digital output of the experimental data of I as a function of s. These data are interpreted with the aid of a Fourier transform of the intensity expression of Eq. (6) to give the P_{ij} functions for the pairs of atoms. From the positions and widths of the peaks for CCl_4 in **Fig. 8**, it is determined that Eqs. (8) hold (1 pm = 10^{-12} m = 3.937×10^{-11} in.). *See* FOURIER SERIES AND TRANSFORMS.

$$\begin{aligned} r_{CCl} &= 176.6 \pm 0.3 \text{ pm} \\ l_{CCl} &= 6.0 \pm 0.5 \text{ pm} \\ r_{ClCl} &= 288.7 \pm 0.4 \text{ pm} \\ l_{ClCl} &= 6.8 \pm 0.3 \text{ pm} \end{aligned} \tag{8}$$

This relatively high precision in determining the distances in gas molecules has been achieved since about 1950 with the development of new instrumentation and the application of high-speed computing methods. The relative intensities of scattered electrons registered on the photographic plate can be measured to better than 0.1%; the time saved by the

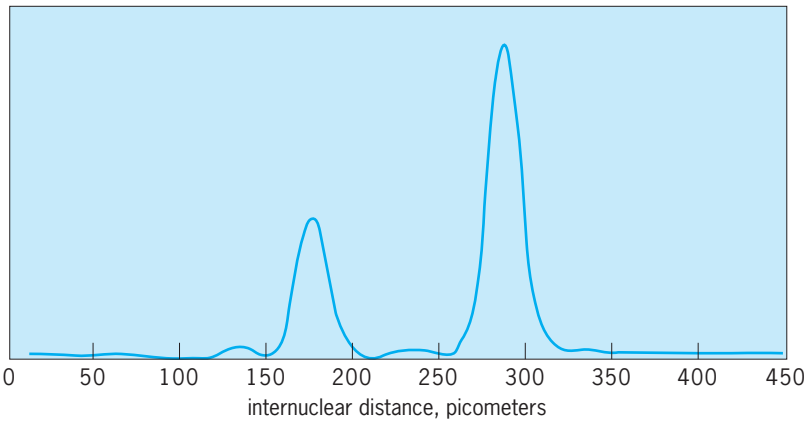


Fig. 8. Experimental distribution of internuclear distances in CCl_4. The two prominent peaks represent the C-Cl and Cl-Cl distances.

use of electronic computing methods in interpreting the data permits a more rigid application of the criteria for satisfactory agreement between experiment and theory.

Over a thousand gases have been studied by the methods giving internuclear distances within 0.4 pm (1.6×10^{-11} in.). In favorable cases precisions exceeding 0.1 pm (4×10^{-12} in.) have been obtained; when these are compared with spectroscopic results, good agreement is found if allowance is made for the difference in the nature of the distances measured. The molecular structures of some 500 gases have been reported by earlier electron diffraction procedures with uncertainties 10 times larger or more.

The applicability of the gas diffraction method alone is limited to simple molecules. Only when the three-dimensional structure can be derived uniquely from a one-dimensional spectrum of internuclear distances blurred by thermal motion is a complete structure determination possible from electron diffraction data alone; light atoms in the presence of heavy ones in the same molecule are less precisely located; distances as close as 3 pm (1.2×10^{-10} in.) can barely be resolved. The range of the method is greatly increased, however, when some structural features in the molecule can be assumed from the results of other methods of investigation. For example, if a six-membered ring of atoms is known to have trigonal symmetry, the number of structural parameters is decreased from 12 to 2. *See* NEUTRON DIFFRACTION; SCATTERING EXPERIMENTS (ATOMS AND MOLECULES); SCATTERING EXPERIMENTS (NUCLEI). Lawrence O. Brockway

Bibliography. W. Braun, *Applied RHEED: Reflection High-Energy Electron Diffraction During Crystal Growth*, Springer-Verlag, 1999; J. M. Cowley, *Diffraction Physics*, 3d ed., 1995; J. M. Cowley (ed.), *Electron Diffraction Techniques*, 2 vols. 1992, 1993; D. L. Dorset, *Structural Electron Crystallography*, 1995; C. S. Fadley, The study of structures by photoelectron diffraction and Auger electron diffraction, in R. Z. Bachrach (ed.), *Synchrotron Radiation Research: Advances in Surface and Interface Science*, 1992; A. Howie and U. Valdre (eds.), *Surface and Interface Characterization by Electron Optical Methods*, 1988; P. K. Larsen and P. J. Dobson (eds.), *Reflection High-Energy Electron Diffraction and Reflection Electron Imaging of Surfaces*, 1988; D. L. Misell and E. B. Brown, *Electron Diffraction: An Introduction for Biologists*, 1988; M. A. van Hove, W. H. Weinberg, and C. M. Chan, *Low Energy Electron Diffraction*, 1987.

Electron emission

The liberation of electrons from a substance into vacuum. Since all substances are built up of atoms and since all atoms contain electrons, any substance may emit electrons; usually, however, the term refers to emission of electrons from the surface of a solid.

The process of electron emission is analogous to that of ionization of a free atom, in which the latter parts with one or more electrons. The energy of the electrons in an atom is lower than that of an electron at rest in vacuum; consequently, in order to ionize an atom, energy must be supplied to the electrons in some way or other. By the same token, a substance does not emit electrons spontaneously, but only if some of the electrons have energies equal to, or larger than, that of an electron at rest in vacuum. This may be achieved by various means. If a substance is heated, the atoms begin to vibrate with larger amplitudes, and electrons may absorb sufficient energy from these vibrations to be emitted in the process known as thermionic emission. Electrons may also be liberated upon irradiation of the substance with light (photoemission). Electron emission from a substance may be induced by bombardment with charged particles such as electrons or ions in the phenomenon called secondary emission. Field emission, or cold emission, is the emission of electrons under influence of a strong electric field. Electrons may also be emitted from one solid into another, but this is usually referred to as electron injection. For example, a metal may inject electrons into an insulator under certain circumstances. *See* FIELD EMISSION; PHOTOEMISSION; SECONDARY EMISSION; THERMIONIC EMISSION. Adrianus J. Dekker

Bibliography. R. Gomer, *Field Emission and Field Ionization*, 1961, reprint 1993; G. Hohler, E. A. Nieksch, and J. Treusch (eds.), *Particle Induced Electron Emission*, 1991; A. Modinos, *Field, Thermionic and Secondary Electron Emission Spectroscopy*, 1984.

Electron-hole recombination

The process in which an electron, which has been excited from the valence band to the conduction band of a semiconductor, falls back into an empty state in the valence band, which is known as a hole. *See* HOLE STATES IN SOLIDS.

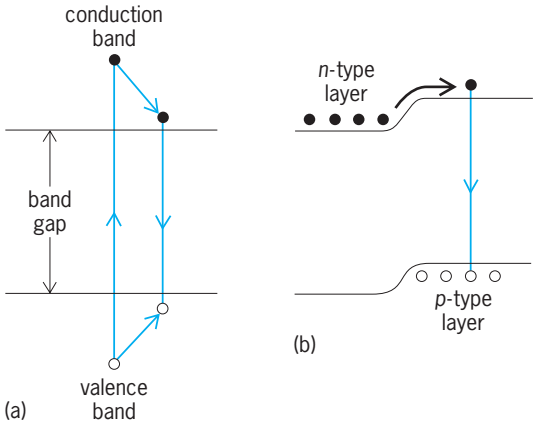


Fig. 1. Recombination of electrons and holes generated by (*a*) optical absorption and (*b*) a forward-biased *pn* junction.

When atoms come together to form a crystal, the discrete atomic energy levels broaden into a level continuum. The periodic potential of the crystal produces gaps in the continuum, separating it into a number of energy bands. In a pure semiconductor crystal, bonding electrons fill the valence band, and above the valence band in energy is an empty conduction band. The two bands are separated by the band gap, which has no levels, and the extent of this band gap is a characteristic of each semiconductor. The band gap is of order 1 eV for common semiconductor materials such as silicon (Si) or gallium arsenide (GaAs). *See* BAND THEORY OF SOLIDS.

Light with photon energies greater than the band gap can be absorbed by the crystal, exciting electrons from the filled valence band to the empty conduction band (**Fig. 1***a*). The state in which an electron is removed from the filled valence band is known as a hole. It is analogous to a bubble in a liquid. The hole can be thought of as being mobile and having positive charge. The excited electrons and holes rapidly lose energy (in about 10^{-12} s) by the excitation of lattice phonons (vibrational quanta). The excited electrons fall to near the bottom of the conduction band, and the holes rise to near the top of the valence band, and then on a much longer time scale (of 10^{-9} to 10^{-6} s) the electron drops across the energy gap into the empty state represented by the hole. This is known as electron-hole recombination. An energy approximately equal to the band gap is released in the process. Electron-hole recombination is radiative if the released energy is light and nonradiative if it is heat. *See* PHONON.

Recombination at pn junctions. Electron-hole recombination requires an excited semiconductor in which both electrons and holes occupy the same volume of the crystal. This state can be produced by purely electrical means by forward-biasing a *pn* junction. The *pn* junction is formed by adjacent *p*-type and *n*-type layers containing excess holes and excess electrons, respectively. In the *n*-type layer, donor impurities replace some of the host atoms. Each donor impurity has one more electron in its outer shell than the host atom, and these extra electrons go into the conduction band. Likewise, in the *p*-type layer, acceptor impurities replace some of the host atoms. The acceptor atoms contain one less electron in their outer shell than the host atoms. They complete their valence bonds by taking a valence electron from elsewhere in the crystal, thereby creating holes in the valence band.

The *pn* junctions forms a barrier that keeps the electrons and holes separated. By contacting these layers and applying a forward-bias voltage to the *pn* junction, the barrier height can be reduced, enabling electrons to flow into the *p*-region or holes into the *n*-region, thus allowing electron-hole recombination to take place. The current passing through a *pn* diode in electrons per second equals the rate of electron-hole recombination (Fig. 1*b*). A major application of this phenomenon is the light-emitting diode (LED), which generates light by means of radiative electron-hole recombination resulting from forward-biasing the *pn* junction. *See* LIGHT-EMITTING DIODE; LUMINESCENCE; SEMICONDUCTOR DIODE.

Radiative recombination. Semiconductor materials may be divided into two types: direct-band-gap semiconductors and indirect-band-gap semiconductors. Efficient radiative recombination between free electrons and holes takes place only in direct-band-gap semiconductors. The continuous energy levels of electrons in the conductor and valence bands are characterized by a definite value of momentum (or more precisely, wave vector). During an optical transition, momentum is conserved, and since the photon carries away negligible momentum, transitions take place only between conduction-band and valence-band states having the same momentum. This is easily satisfied in direct-band-gap semiconductors, because electrons and holes collect at the conduction band at minimum and the valence band at maximum, and both extrema have the same momentum (**Fig. 2***a*). However, for indirect-band-gap semiconductors (Fig. 2*b*), the conduction-band minimum and valence-band maximum have very

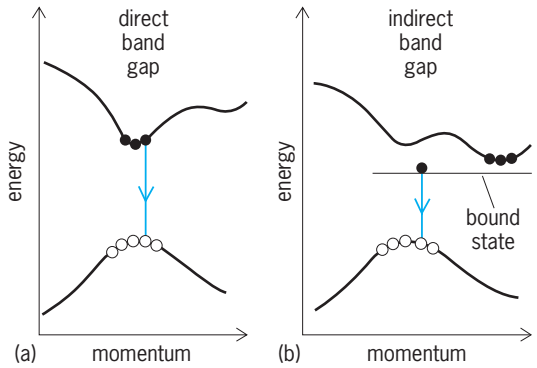


Fig. 2. Energy versus momentum for the conduction-band and valence-band states of (*a*) a direct-band-gap semiconductor and (*b*) an indirect-band-gap semiconductor. Optical transitions in indirect-band-gap semiconductors take place through bound states that have a momentum distribution as indicated.

different momenta, and consequently optical transitions between free electrons and holes are forbidden.

Radiative electron-hole recombination is possible in indirect-band-gap semiconductors when the transition is assisted by lattice phonons and impurities. Weak optical transitions can take place in which a momentum-conserving phonon is emitted along with the photon. Electrons can be trapped at impurities having energy levels within the energy gap. Unlike the free electron, the bound electron has a spread of momentum as a consequence of its localization (the uncertainty principle), with some components of momentum at the valence-band maxima (Fig. 2*b*). The isoelectronic impurity nitrogen substituted for phosphorus in gallium phosphide (GaP) and gallium arsenide phosphide (GaAsP) is used for this purpose. When heavily doped with nitrogen, these indirect semiconductors have moderate luminescence efficiency and are used as light-emitting diodes to generate green light. Bismuth impurities can be used in the same way to produce yellow light. It would be more desirable to use direct-band-gap semiconductors for this purpose, but currently there are no direct-band-gap semiconductors capable of forming *pn* junctions which emit these colors at room temperature. *See* CRYSTAL DEFECTS.

Apart from its application in light-emitting diodes and laser operation, radiative recombination, especially at low temperatures (approximately 2 K or −456°F), has been a very important tool for studying the interaction of electrons and holes in semiconductor crystals. At low temperatures, many weakly bound states such as excitons are able to form stably. These states interact only weakly with lattice vibrations and consequently emit much of their recombination radiation in the form of sharp (zero phonon) spectral lines. For example, at low temperatures, electrons become trapped at donor sites and holes at acceptor sites. The spectrum of donor-acceptor recombination contains hundreds of resolvable lines, one for each crystallographically inequivalent donor-acceptor pair. *See* EXCITON.

Nonradiative recombination. Competing with radiative recombination are the nonradiative recombination processes of multiphonon emission and Auger recombination. Nonradiative recombination of electrons and holes by multiphonon emission occurs in two steps. First, an electron is captured into a deep level near the middle of the energy gap, and subsequently the trapped electron state captures a hole from the valence band (**Fig. 3**). Multiphonon emission requires atoms very different from the host, or defects such as vacancy-impurity complexes, dislocations, or semiconductor-oxide interface states. Electron capture alters the positions of the surrounding lattice, in a process known as lattice relaxation that lowers the electronic energy of the bound electron (Fig. 3). The vibration of the lattice causes the electronic level to move within the energy gap and even cross into the bands. Capture takes place during the crossing.

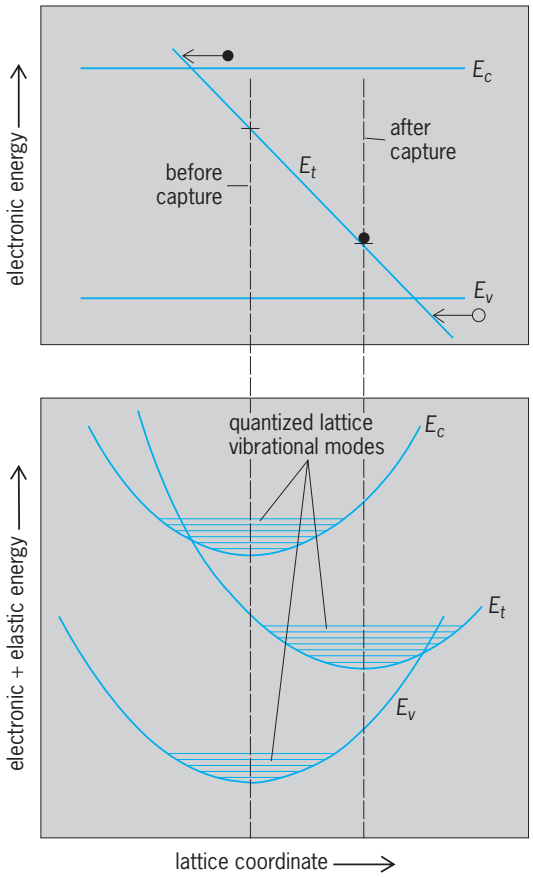


Fig. 3. Electron and hole capture into a deep level by multiphonon emission. After capture the lattice relaxes, lowering the electronic energy of the bound state E_t. E_c and E_v are the minima of the conduction band and maxima of the valence band, respectively.

In multiphonon emission, the energy lost during capture highly excites the vibrations of one or several atoms related to the defect site. This can greatly enhance the probability for an atom to jump to a different site, which also requires a highly excited atomic vibration. Such recombination-enhanced motion has been observed in detail for defects formed by ionizing radiation in GaAs and in Si. It is suspected that nonradiative recombination by multiphonon emission drives the movement of atoms at room temperature that are responsible for device degradation phenomena such as the climb of dislocations found in GaAs light-emitting diodes and lasers. *See* DIFFUSION; RADIATION DAMAGE TO MATERIALS.

Another source of nonradiative recombination is the Auger effect, in which an electron or hole takes up the energy released in the recombination of a pair of other charge carriers. Auger recombination has been observed for excitons bound to neutral donors in Si and GaP, a complex of two electrons and a hole. In these cases, the Auger recombination rate is several orders of magnitude greater than the radiative rate. Auger recombination has been shown to limit the performance of long-wavelength (1.3–1.6 micrometer) lasers and light-emitting diodes used in optical communication systems. The Auger effect reduces light-emitting diode

efficiency and greatly enhances the increase of laser threshold current with temperature. *See* AUGER EFFECT; LASER; OPTICAL COMMUNICATIONS; SEMICONDUCTOR. Charles H. Henry

Bibliography. V. N. Aba Kumov, V. I. Perel, and I. N. Yassievich, *Nonradiative Recombination in Semiconductors*, 1991; G. Gillessen and W. Schairer, *Light-Emitting Diodes: An Introduction*, 1987; P. T. Landsberg, *Recombination in Semiconductors*, 1992; J. Singh, *Semiconductor Devices An Introduction*, 1994; S. M. Sze, *Semiconductor Devices*, 1985.

Electron lens

An electric or magnetic field, or a combination thereof, which acts upon an electron beam in a manner analogous to that in which an optical lens acts upon a light beam. Electron lenses find application for the formation of sharply focused electron beams, as in cathode-ray tubes, and for the formation of electron images, as in infrared converter tubes, various types of television camera tubes, and electron microscopes.

Any electric or magnetic field which is symmetrical about an axis is capable of forming either a real or a virtual electron image of an object on the axis which either emits electrons or transmits electrons from another electron source. Hence, an axially symmetric electric or magnetic field is analogous to a spherical optical lens.

The lens action of an electric and magnetic field of appropriate symmetry can be derived from the fact that it is possible to define an index of refraction n for electron paths in such fields. This index depends on the field distribution and the velocity and direction of the electrons. It is given by the equivalencies shown below. Here e is the charge of the electron,

$$n = \sqrt{\phi + \frac{2e\phi^2}{mc^2}} - \sqrt{\frac{e}{2m}}\, A\, \cos\chi$$

$$= \sqrt{\phi + 0.978 \cdot 10^{-6}\phi^2} - 0.297A\, \cos\chi$$

m its rest mass, ϕ the potential of the point in space under consideration (so normalized that the kinetic energy of the electron vanishes for $\phi = 0$), c the velocity of light, A the magnetic vector potential, and χ the angle formed by the path of the electron with the direction of the magnetic vector potential. For an axially symmetric field the magnetic vector potential is perpendicular to the plane passing through the axis of symmetry and the reference point. Its magnitude is equal to the magnetic flux through the circle about the axis through the reference point divided by the circumference of that circle. The numerical coefficients in the final expression for n apply if ϕ is measured in volts and A in gauss-centimeters. *See* ELECTRON MICROSCOPE.

Electron lenses differ from optical lenses both in the fact that the index of refraction is continuously variable within them and that it covers an enormous range. Furthermore, in the presence of a magnetic field, n depends both on the position of the electron in space and on its direction of motion. It is not possible to shape electron lenses arbitrarily. *See* ELECTROSTATIC LENS; MAGNETIC LENS. Edward G. Ramberg

Bibliography. D. A. DeWolf, *Basics of Electron Optics*, 1990; E. M. Slayter and H. S. Slayter, *Light and Electron Microscopy*, 1993.

Electron microscope

A device for forming greatly magnified images of objects by means of electrons. Electron microscopes serve primarily two purposes: the visual examination of structures too fine to be resolved with ordinary, or light, microscopes, and the study of surfaces that emit electrons. The first function made transmission electron microscopes essential research tools in biology, chemistry, and metallurgy. Beginning in the 1960s the scanning electron microscope came to play an increasingly important role in the study of the surfaces of solid objects at more moderate magnifications. Various emission electron microscopes serve more specialized research purposes.

Instruments

A transmission electron microscope, shown in close analogy with a light microscope in **Fig. 1**, consists in its simplest form of a source supplying a beam of electrons of uniform velocity, a condenser lens for concentrating the electrons on the specimen, a specimen stage for displacing the specimen which transmits the electron beam, an objective lens, a projector lens, and a fluorescent screen on which the final image is observed. For permanent record of the

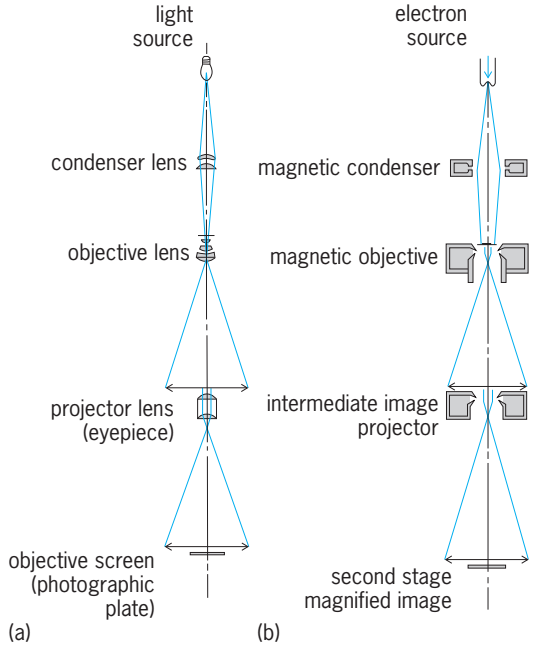


Fig. 1. Comparison between (*a*) light microscope and (*b*) magnetic electron microscope. (*After V. K. Zworykin et al., Electron Optics and the Electron Microscope, Wiley, New York, 1945*)

image, the fluorescent screen is replaced by a photographic plate or film.

Electrons are strongly scattered by all forms of matter including air. Hence the entire instrument must be evacuated to about 10^{-4} mmHg (10^{-7} atm or 10^{-2} pascal). Furthermore, the lenses cannot be material in nature. Instead, they are electric or magnetic fields, symmetrical about the axis of the instrument, that have the property of bending the electron paths toward the axis, just as converging glass lenses bend light rays toward their axis. The lenses in the instruments shown in Fig. 1 are magnetic fields formed at narrow gaps in the iron casings surrounding coils traversed by electric current. Lens strength is varied by varying the current. Most electron microscopes employ magnetic lenses of this type. These have yielded the highest resolution and magnification attained.

However, good results have also been obtained with electron microscopes employing unipotential electrostatic lenses and magnetic lenses excited by permanent magnets. *See* ELECTRON LENS; ELECTROSTATIC LENS; MAGNETIC LENS.

Resolution. A microscope can, at best, permit the discrimination of two point objects greater than $0.7\lambda/\sin\theta$ apart. Here λ is the wavelength of the illuminating radiation, and θ is the aperture angle of the cone of radiation that participates in forming the image (**Fig. 2**). For green light $\lambda = 500$ nanometers. Even for ultraviolet radiation in an immersion medium of refractive index 1.5, λ is no less than 170 nm.

Since light and ultraviolet microscope objectives can be designed to utilize practically all the radiation passing through the specimen, $\sin\theta \cong 1$ and the least resolvable distance for the ultraviolet microscope is about 100 nm. For 50- to 100-kV electrons, such as are commonly employed in electron microscopes, the wavelength range is 0.0053–0.0037 nm. Hence, even though a cone of radiation with an aperture angle less than 0.01 radian contributes to an image of optimum sharpness, object separations smaller than 0.3 nm have been resolved with the electron microscope.

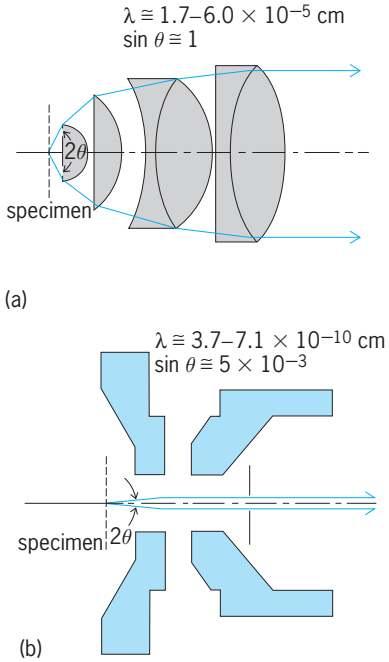


Fig. 2. Effective objective aperture in (*a*) light and (*b*) electron microscopes.

Thus the electron microscope has several hundred times the resolving power of the light microscope. Similarly, whereas the maximum useful magnification of the light microscope is about 2000, that of the electron microscope may approach 1,000,000. The maximum useful magnification is the least magnification of the image that reveals to the observer all the specimen detail that the microscope is capable of conveying.

Transmission electron microscope. Electrons are commonly emitted from the tip of a fine tungsten-wire hairpin filament or, to further reduce the size of the effective electron source, from a sharply pointed segment of wire welded to the filament tip (**Fig. 3**). The filament is maintained at a carefully stabilized negative potential of 50–100 kV with respect to the remainder of the instrument. Electrons enter the instrument through an anode aperture. The intensity and convergence of the electron beam that is falling on the specimen are adjusted by varying the coil current of the condenser lens.

Image contrasts are formed by the scattering of electrons out of the narrow cone that contributes to the formation of the image; denser or thicker portions of the specimen scatter more electrons and hence appear darker in the image. The sharpness of the image observed on the screen is adjusted by varying the objective coil current, and its magnification by varying the projector coil current. Both currents must be carefully stabilized to yield high resolution. In modern instruments the function of the projector is performed by two magnetic lenses in tandem. In the instrument shown in Fig. 3 the normal magnification range from 1400 to 200,000 is realized by varying the coil currents. The magnification range can be extended downward to 500 with the aid of a special specimen holder and projector polepiece. A binocular viewer provides an additional magnification by 10, needed for the fine adjustment of the objective current to yield images of maximum sharpness. Comparable magnifications of the final image are obtained by enlarging the fine-grain photographic plates on which the electron image is recorded.

The image intensity required for precise visual focusing and specimen exploration is considerably higher than that needed for photographic recording. This intensity, and the corresponding specimen damage by the electron beam, can be greatly reduced through the use of an image intensifier optically coupled to the image on the fluorescent screen of the electron microscope on the one hand and the target of a television camera tube on the other (**Fig. 4**). The picture signal generated by the camera tube can be employed to reproduce the microscope image instantaneously on the screen of a television receiver, can be recorded on tape by a video recorder for instant playback at a later time, or can be distributed

to a group of receivers for teaching purposes. Since the image can be photographed on the screen of the television receiver, this system makes it also possible to avoid contamination of the vacuum in the microscope column resulting from the introduction of photographic material. Furthermore, it permits arbitrary control of image contrasts.

In addition to the standard transmission microscopes operating at 50–100 kV, a number of very high-voltage instruments have been constructed. Instruments designed for operation up to 1500 kV (**Fig. 5**) are necessarily large, because clearances in the gun and the high-voltage supply are required to suppress field emission, and because the focal lengths of electron lenses increase as the accelerating voltage of the electrons is increased. The advantage of high-voltage electron microscopy does not lie in greater resolving power—the effect of a shorter electron wavelength (0.00087 nm at 1,000,000 eV) is more than outbalanced by the increased aberration of the lenses—but in increased penetration, which is particularly valuable in the direct study of metal sections prepared with a microtome.

Scanning electron microscope. A modern version of the scanning electron microscope is illustrated in **Fig. 6**. The microscope column, with the source at the top and the specimen chamber at the bottom, is at the left. The square cathode-ray tube display for visual observation with a long-persistence yellow phosphor is in the center, and the cathode-ray tube display for photographic recording with a short-persistence blue phosphor is at the right. **Figure 7** is a schematic diagram of the microscope column. The electron gun is similar to that in the transmission microscope, but operates at much lower voltages (1–20 kV). Three magnetic lenses image the first crossover, or narrowest cross section of the beam issuing from the filament tip, at a greatly reduced scale on the specimen. A pair of deflection coils actually mounted within the bore of the final lens serves to deflect the electron probe formed by the lenses over the surface of the specimen, the specimen and the cathode-ray displays being scanned in synchronism. The secondary electrons generated at the point of impact of the electron beam on the specimen are accelerated by 12 kV and impinge on a plastic scintillator optically coupled by a plastic light pipe to a multiplier phototube. The output current of this phototube controls the beam current, and hence the brightness in the display. *See* SCINTILLATION COUNTER.

The whole scanning electron microscope can be regarded as a slow-scan television system, with the microscope column constituting a very high-resolution camera tube, the specimen serving as camera tube target. Resolution is determined by the electron probe size at the specimen and magnification by the ratio of the scan amplitude in the display to that at the specimen. Thus the magnification can be varied, for example, in a range of 20–100,000, without any refocusing. The effective probe size is determined not only by the electron lens properties but also by the scattering of the electrons within the specimen. Since the range of electron diffusion decreases as the accelerating voltage of the electrons is reduced, relatively low operating voltages are desirable. The optimal resolution cited for the instrument in Fig. 6 is better than 20 nm, the normal operating resolution about 50 nm. At the same time, the great depth of focus (about 300 times that of a light microscope) endows images with considerable extent in depth with an extraordinarily lifelike, three-dimensional quality. This is lacking in light microscope images of the same magnification, since the image is sharp only at a rather sharply defined planar section through the object surface. *See* CATHODE-RAY TUBE; TELEVISION CAMERA TUBE.

Since the electron probe currents are very small (10^{-9} to 10^{-13}), the recording of noise-free high-resolution images demands relatively long scan periods, ranging from 2 to 1000 s; for visual observation

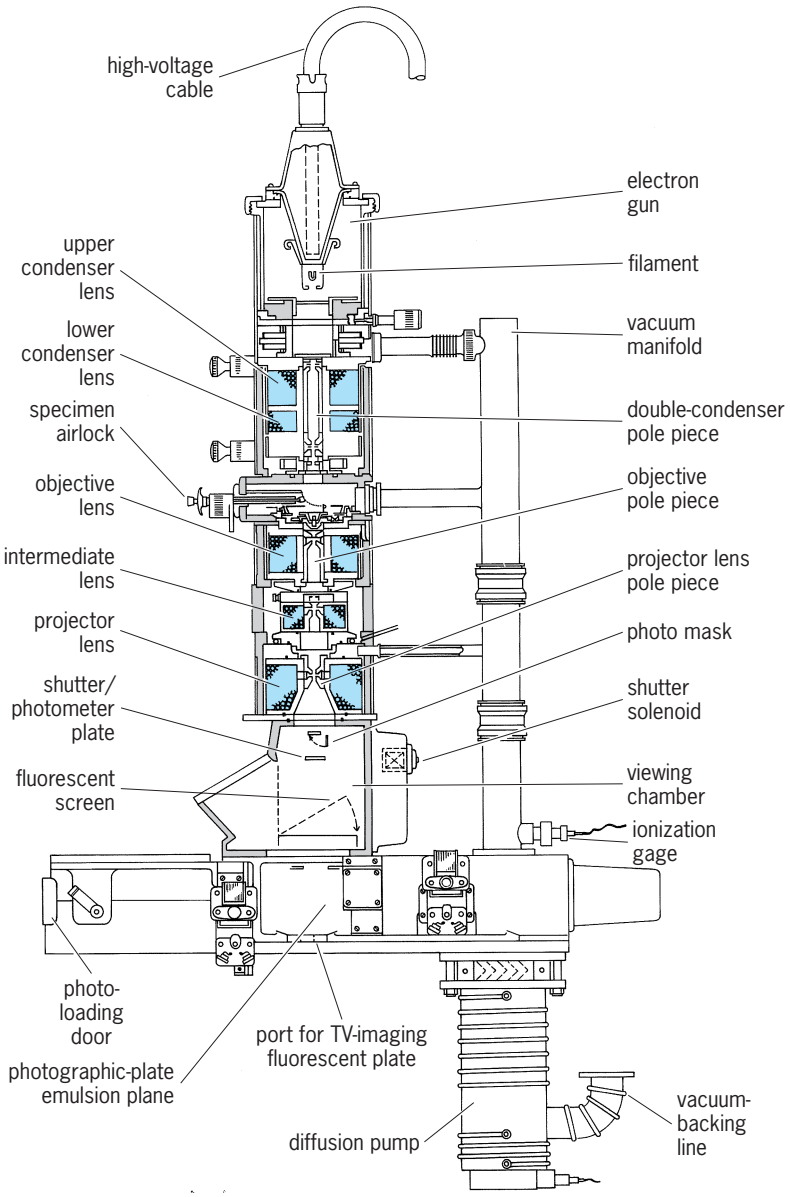


Fig. 3. Section through column of an electron microscope. This instrument has a normal magnification range from 1400 to 200,000. (*Radio Corporation of America*)

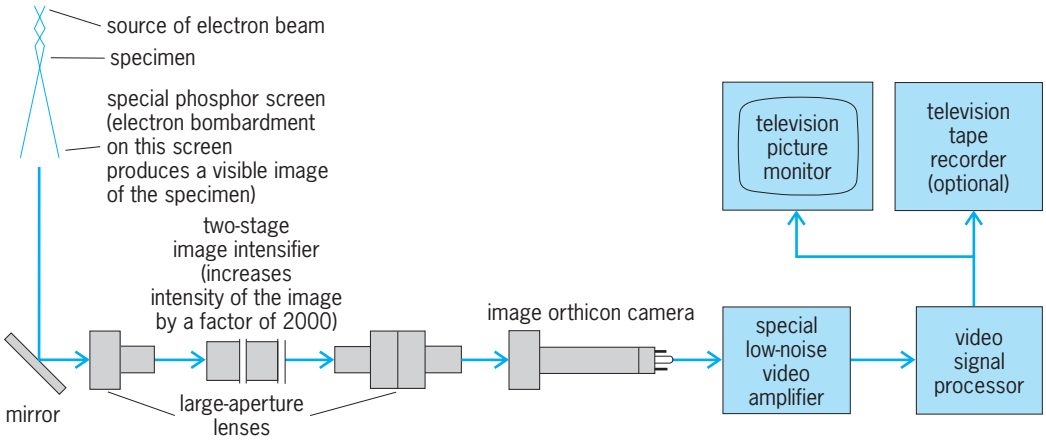


Fig. 4. Scanning electron microscope. (*Cambridge Instrument Co., Ltd.*)

on the long-persistence screen the frame period is reduced to 0.1 to 12 s.

The signal pickup system can be modified in various ways to show different properties of the object. In the examination of integrated circuit elements a comparison of the image obtained in the conductive mode (**Fig. 8**), the signal being derived from an appropriately chosen electrode on the specimen, with that obtained in the normal emissive mode is frequently useful in testing the operation of the system and locating defects. *See* INTEGRATED CIRCUITS; SCANNING ELECTRON MICROSCOPE.

Emission electron microscopes. These may take several forms (**Fig. 9**). In the immersion electron microscope, for which a rudimentary form is shown in Fig. 9*a*, the specimen is a flat conducting surface which may be heated, illuminated, or bombarded by high-velocity electrons or ions so as to emit low-velocity thermionic, photo, or secondary electrons. These are accelerated to a high velocity in an immersion objective or cathode lens and imaged as in a transmission electron microscope.

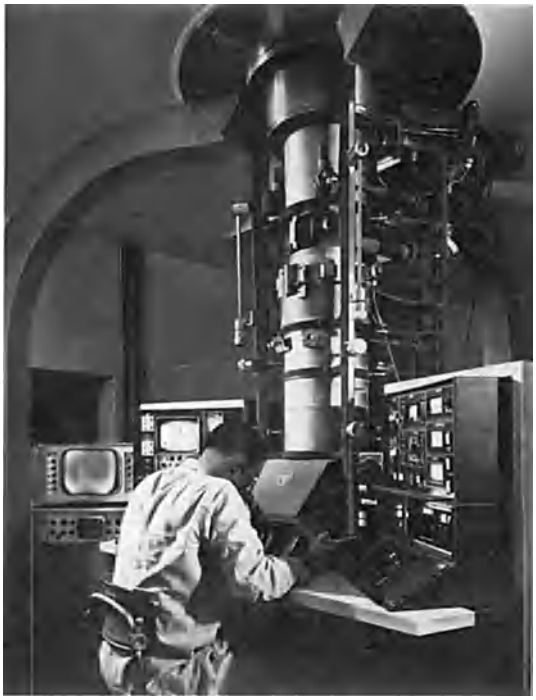


Fig. 5. TV image intensification system for the electron microscope. (*Radio Corporation of America*)

Other emission microscopes employ simple projection without lenses. One example (Fig. 9*b*) is the cylindrical microscope used for examining the thermionic emission from a thin wire stretched along the axis of a cylindrical tube whose walls at high positive potential are coated with a fluorescent substance. Another (Fig. 9*c*) is the field-emission microscope, in which field electron currents are drawn from an extremely fine, rounded point on the end of a wire of a metal such as tungsten. The electrons drawn out of the point follow nearly straight lines toward a positive-potential screen. The display on the screen shows variations in the emission characteristics (as determined by the work function and surface nonuniformities) over the surface of the point. *See* FIELD-EMISSION MICROSCOPY; THERMIONIC EMISSION.

Low-temperature techniques. Apart from freeze-etching, low-temperature techniques find several other applications in electron microscopy.

One of these is employment of a superconducting objective lens, as demonstrated by H. Fernandez-Moran. The lens field is produced by an air-core solenoid of niobium-zirconium alloy, cooled by a liquid-helium bath and short-circuited after excitation so as to remain in the persistent-current mode. While the immersion of the microscope column in liquid helium presents a very serious inconvenience, such superconducting solenoids may generate higher magnetic fields and can thus provide, in principle, lenses of shorter focal length and lower aberration than conventional iron pole-piece objectives. Furthermore, they are characterized by extraordinary stability. *See* SUPERCONDUCTIVITY.

Liquid air–cooling of the specimen chamber also serves to suppress contamination deposited on the specimen through the interaction of the electron beam with traces of organic vapor; without special precautions, contamination buildup can quickly

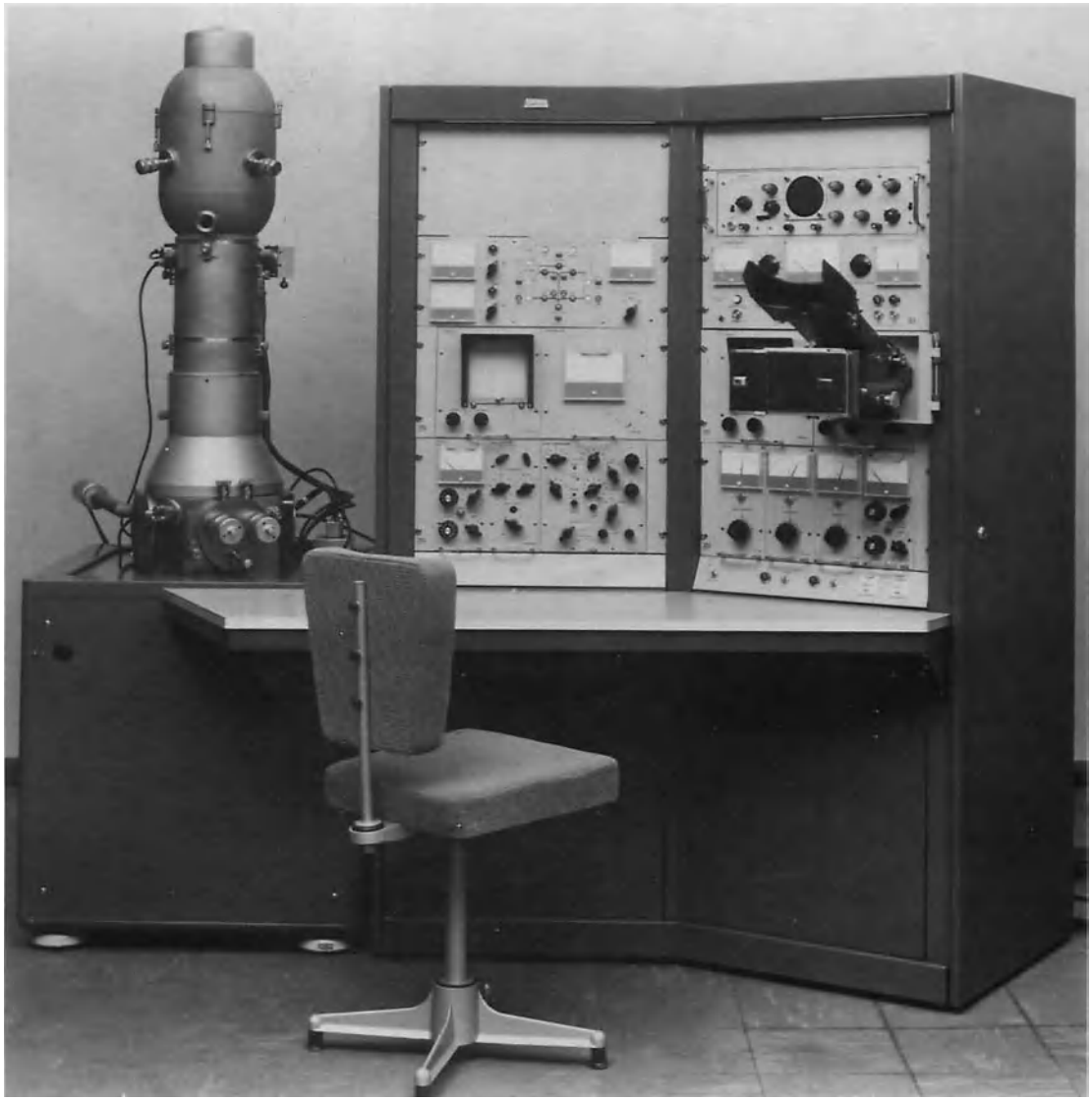


Fig. 6. Megavolt electron microscope at U. S. Steel research laboratory at Monroeville, Pennsylvania. (*Radio Corporation of America*)

obscure fine specimen detail at very high magnifications.

Finally, freeze-drying and freeze substitution provide ways of preparing organic specimens for observation with minimal distortion of their cell structure. In the former instance the specimen may be cooled extremely rapidly by immersion in liquid air–cooled propane and then be warmed slowly in vacuum, so that the ice sublimes rather than melts. The infusion of the embedding material then takes place either in vacuum or a dry-nitrogen atmosphere. In freeze substitution ice is removed by diffusion into a liquid solvent instead of sublimation in vacuum. Edward G. Ramberg

Electron Microscopy

The application of the electron miscroscope to examination and investigation of the ultramicrostructure of materials has become so extensive that there is hardly an area in biological and nonbiological research where electron microscopy does not play a role. In biological and medical research the development of sectioning techniques has extended electron microscopy down to observations at the macromolecular level for delineation of the complex organization of cell components such as membranes, mitochondria, endoplasmic reticulum, and ribosomes. Other techniques have been developed for studies on the structure of virus and even individual proteins and nucleic acids. Nonbiological solid materials have also become objects of extensive and fruitful investigation. Diffraction microscopy, together with the development of adequate techniques of sectioning and thinning crystalline materials such as metals, now makes the observation of defect structure in solids an important aspect of electron microscopy.

The very different character of the interaction of electrons and light with matter produces the greatest practical differences between electron microscopy and optical microscopy. Contrast in the optical microscope image is produced largely by

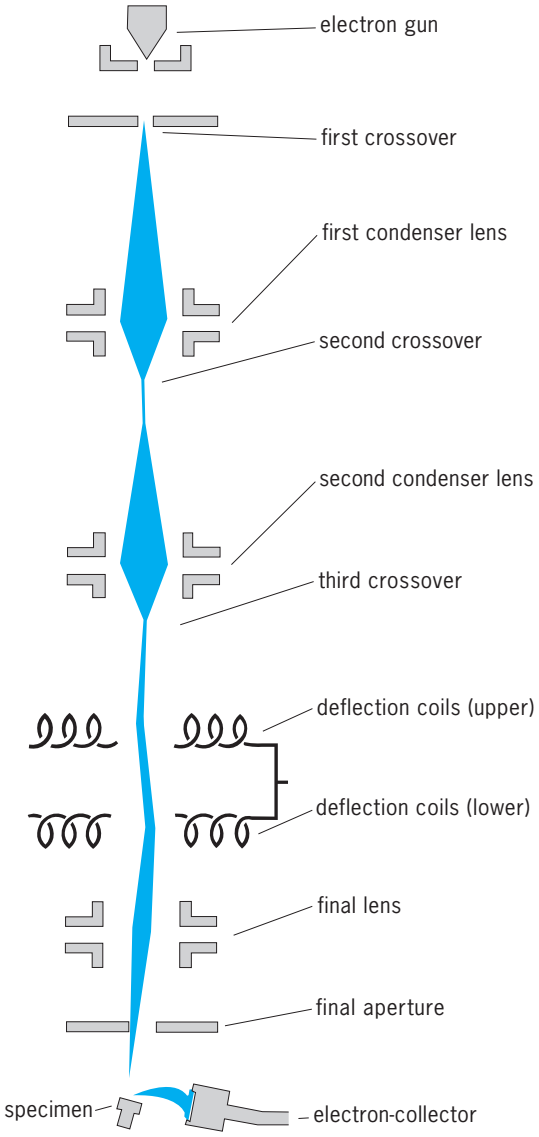


Fig. 7. Scanning electron microscope column. (*Cambridge Instrument Co., Ltd.*)

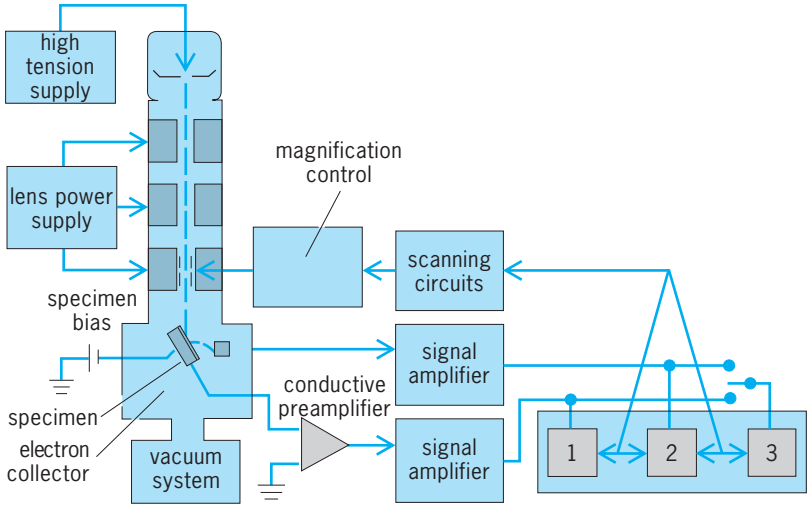


Fig. 8. Alternative operation of scanning electron microscope in conductive and emissive mode. (*Cambridge Instrument Co., Ltd.*)

spectral adsorption of light passing through the specimen. Contrast in the electron microscope image is produced primarily by the scattering of electrons from the specimen. Electrons in the 50–100-kV range, used in conventional commercial electron microscopes, are strongly scattered by matter. It is this strong interaction that determines the requirements for specimen preparation and places serious limitations on the examination of materials. As indicated previously, to avoid scattering of electrons by ambient gas molecules, an electron beam must be confined to a column evacuated to a pressure of the order of 10^{-4} to 10^{-5} mmHg (10^{-7} to 10^{-8} atm or 10^{-2} to 10^{-3} Pa). Thus, specimens must be so prepared that they can be placed in a high vaccum. Dehydration is required, and this requirement immediately places a severe restraint on all biological systems. Moreover, the total amount of material in a specimen must be very small to prevent excessive heating and the specimen must also be extremely thin. Substrates for mounting specimens are typically of the order of 10 nm thick. Sections of solid materials should be no thicker than the order of 0.1 micrometer. Biological tissues are usually sectioned to thicknesses of about 0.05 μm to obtain high-resolution detail. Metal foils are usually thicker, ranging from 0.1 to 0.2 μm to obtain more information on their defect characteristics. As thickness is increased, multiple scattering of electrons by matter destroys the contrast in the image; finally, energy transferred from the beam by inelastic collisions becomes so large that the specimen may be destroyed.

High-voltage electron microscopes have been developed in which the electrons are accelerated to energies of 500 keV to over 1 MeV, giving them greater penetrating power. Some of the severe limitations indicated above are relaxed to a degree in these high-voltage instruments. Thicker sections can be examined; less damage occurs in the specimen, and small specimen wet chambers with hydrated biological materials can be utilized. Development of high-voltage electron microscopy began in the mid-1960s. The problems and complexities of high-voltage electron microscopes limit their application, and most investigations will continue to be done within the limitations set by the 50–100-kV instruments.

Biological applications. Techniques and applications of electron microscopy are so specialized that it is convenient to discuss them in separate categories of biological and nonbiological, even though there are basic techniques that overlap both fields such as substrate preparation and shadow-casting.

Substrates. The material to be examined in the electron microscope must, in general, be mounted on an extremely thin supporting membrane which has a mass density per unit area less than that of the object being observed. In practice a membrane of the order of 10 nm thick is mounted on a fine electrodeposited metal mesh with openings 50–100 μm on a side, and the specimen material is deposited on these mounted substrates. Cellulose nitrate films prepared by casting a film on water from a solution of cellulose nitrate in amyl acetate produces a suitable substrate.

Another substrate material is Formvar, a polyvinyl formal plastic, cast on glass from solutions of the plastic in dichloroethane (ethylene dichloride). More durable substrates can be prepared by evaporating carbon on a glass slide, stripping off the carbon film by floating it onto a water surface, and mounting the film on the standard mesh specimen screens. Carbon-film substrates as thin as 5 nm can be produced and handled in this fashion.

Shadow-casting. Materials of low mass density will not produce appreciably greater scattering than the substrate film upon which they must be mounted. Contrast in electron images from these specimens will therefore be low, and since this low differential of mass density applies especially to organic materials, it presents a large obstacle in imaging very small biological objects. The technique of shadow-casting developed by R. C. Williams and R. W. G. Wyckoff is a very powerful means of enhancing contrast in electron microscopy of small objects of low mass density. It is now possible to resolve globular macromolecular material with diameters as small as 2–3 nm and linear structures such as strands of nucleic acid less than 1 nm thick. Resolution in shadow-cast specimens is set by the granularity of the shadow-casting material.

The technique involves vacuum deposition of an extremely thin metal layer on the specimen at an oblique angle. Any projection above the surface receives a heavier deposit of the metal on the side facing the metal evaporating source and casts a shadow in which no metal is deposited (**Fig. 10**). The metal shadowing layer produces a topographical representation of the surface of the specimen and, as discussed below, is an essential step in all replication of surfaces for study by electron microscopy.

After shadow-casting, the metal layer is the specimen imaged by the electron microscope, for the scattering from the shadow-cast specimen is predominantly that of the deposited metal layer. The areas in the shadow where no metal is deposited will scatter relatively few electrons, and the respective areas of the image will appear dark on the photographic plate that has been exposed to the high electron density incident in these portions of the image. Projecting areas facing the evaporation source will have heavy layers of the metal and will scatter most electrons, thus appearing light on the photographic record of the image. An example of a shadow-cast specimen is shown in **Fig. 11**, in which the image is reproduced in the customary manner with the same relative light and dark as on the original electron micrograph negative. The effect is as though strong illumination were falling obliquely on the specimen.

The metals most generally used are uranium, platinum, or chromium. The procedures developed for producing thin films by metal evaporation are applied. A tungsten filament charged with the metal to be evaporated is heated in a vacuum of about 10^{-5} mmHg (10^{-8} atm or 10^{-3} Pa) to a temperature at which the metal has a vapor pressure of the order 10^{-2} mmHg (10^{-5} atm or 1 Pa), and a layer of mean

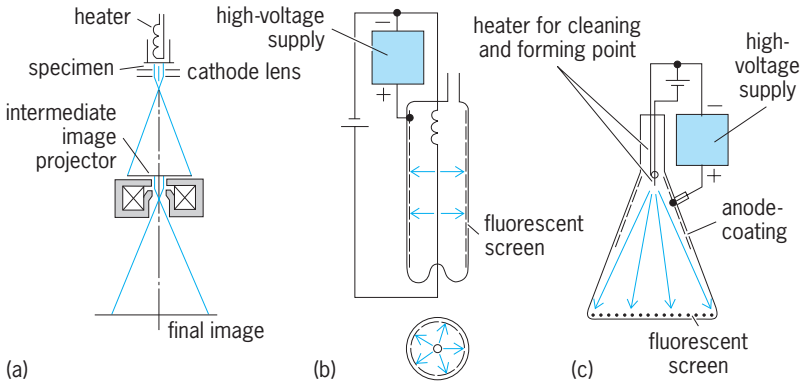


Fig. 9. Emission electron microscopes. (*a*) Immersion electron microscope. (*b*) Cylindrical thermionic-emission microscope. (*c*) Field-emission microscope.

thickness of about 0.5 nm is deposited on the specimen at the desired angle.

Particulate materials. Examination of particulate biological materials such as bacteria, virus, or mitochondria of cells requires proper dispersion of a very small sample of the material on the substrate. Methods must be utilized that prevent aggregation and surface tension effects of drying.

Particles already dispersed and suspended in an aqueous medium can be deposited directly on plastic or carbon substrate films. The effects of surface tension on drying, the aggregation of the particles, and the effects of changing salt concentration and pH on drying must be considered. A method of minimizing these effects is to spray the particle suspension on the substrate in very small droplets of the order of 5–10 μm in diameter. If buffers are needed, volatile ones such as ammonium acetate, ammonium carbonate, or veronal acetate must be used. A further advantage of the spray technique is the rapid drying of the small droplets; this minimizes the effects of changing concentration on the biological materials. By depositing the droplets on substrates cooled to liquid nitrogen temperatures and subliming the ice

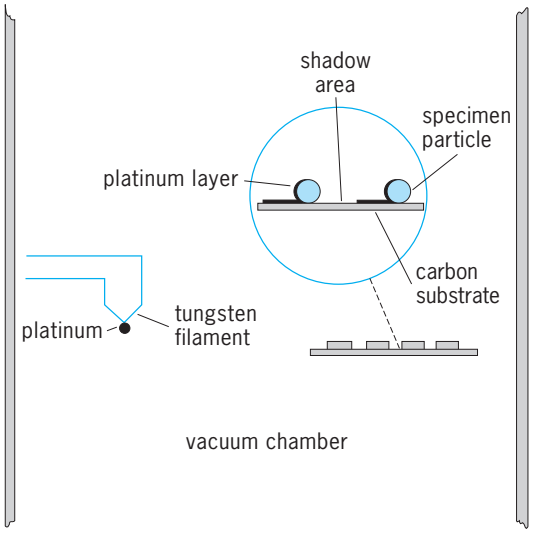


Fig. 10. Schematic of shadow-casting procedure.

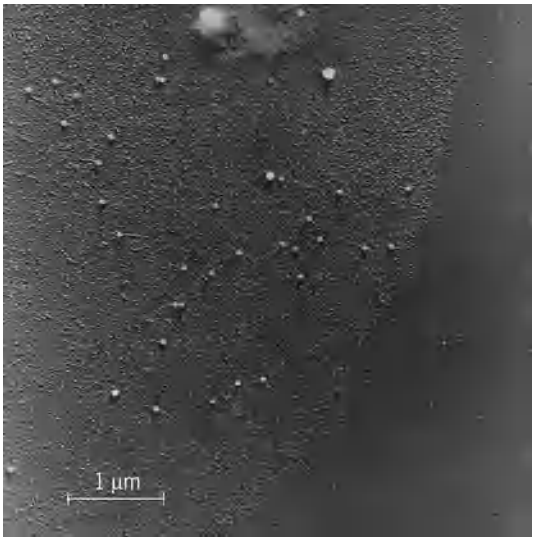


Fig. 11. Edge of spray droplet with bacteriophage.

under vacuum, the surface tension effects of drying can be further reduced.

The spray technique can be used for quantitative work, and assays can be made on suspensions of particles of unknown concentration in which a direct count is obtained on the number of particles per unit volume. Since each droplet is an aliquot sample of the total specimen, when the volume of the droplet is known, a count of the particles in a drop pattern gives the assay. By mixing the suspension to be assayed with a given volume of a known concentration of monodispersed polystyrene latex particles and counting the number of polystyrene in each drop, the volume of the drop is obtained. When a highly purified material is available for assay, it is possible to take an aliquot portion to dryness, to weigh it, and from the electron microscope particle count to calculate the molecular weight of the particle. **Figure 12** shows tobacco mosaic virus (rods) being assayed using the polystyrene spheres.

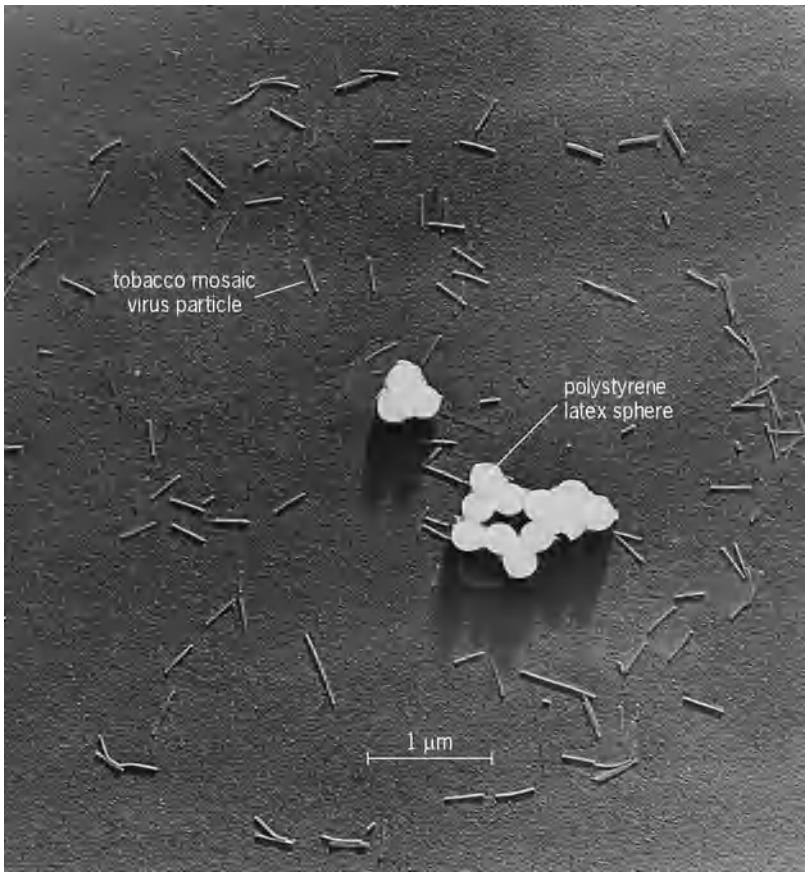


Fig. 12. Spray droplet pattern for assay of tobacco mosaic virus (rods) using polystyrene latex particles (spheres).

The electron microscope is capable of resolving macromolecules of biological interest. Spraying a purified sample of macromolecular material onto a very smooth surface such as freshly cleaved mica produces a suitably dispersed specimen. This specimen is shadowed directly on the mica with platinum, is backed with an evaporated substrate of carbon, and is stripped off the mica by using methods described below.

Negative staining is another important technique used to enhance the image contrast obtained at very high magnifications from biological specimens of a particulate nature such as subcellular components, bacterial surface structure, virus particles, and protein molecules. The particulate material is treated much as described above but with a soluble heavy metal salt added to the aqueous suspension medium. A typical negative staining procedure uses a 2% solution of potassium phosphotungstate buffered and adjusted to a neutral pH of 6.8 to 7.4. A droplet of the suspension either sprayed or deposited on the substrate film dries with an embedding matrix of the heavy metal salt surrounding it. Examination in the electron microscope produces an image of reverse contrast. The surrounding matrix strongly scatters the electron beam, producing a dark ground, while particles such as viruses which should not react with the stain appear light. If the virus or other particle under examination has substructure, the heavy metal matrix can penetrate the interstices and delineate the subunits of the particle (**Fig. 13**).

Tissues and cells. Successful development of techniques for sectioning biological tissues for examination at high resolution provided a means of directly observing the complex ultrastructure of organisms. The understanding obtained of the organization of the cell and the structure of its organelles profoundly changed the fields of histology and cytology and is the most significant contribution of electron microscopy (**Figs. 14** and **15**). *See* CELL (BIOLOGY).

The first successful thin sectioning using a modified conventional microtome was reported in 1948 by D. C. Pease and R. F. Baker. After that techniques developed rapidly and the methods for study of the ultrastructure of cells and tissues became highly developed. Microtomes, using fractured glass or polished diamonds for knives, cut serial sections for reproduction down to thicknesses as low as a few hundred angstroms, or in special cases as thin as 5 nm. Techniques for fixation, dehydration, embedding,

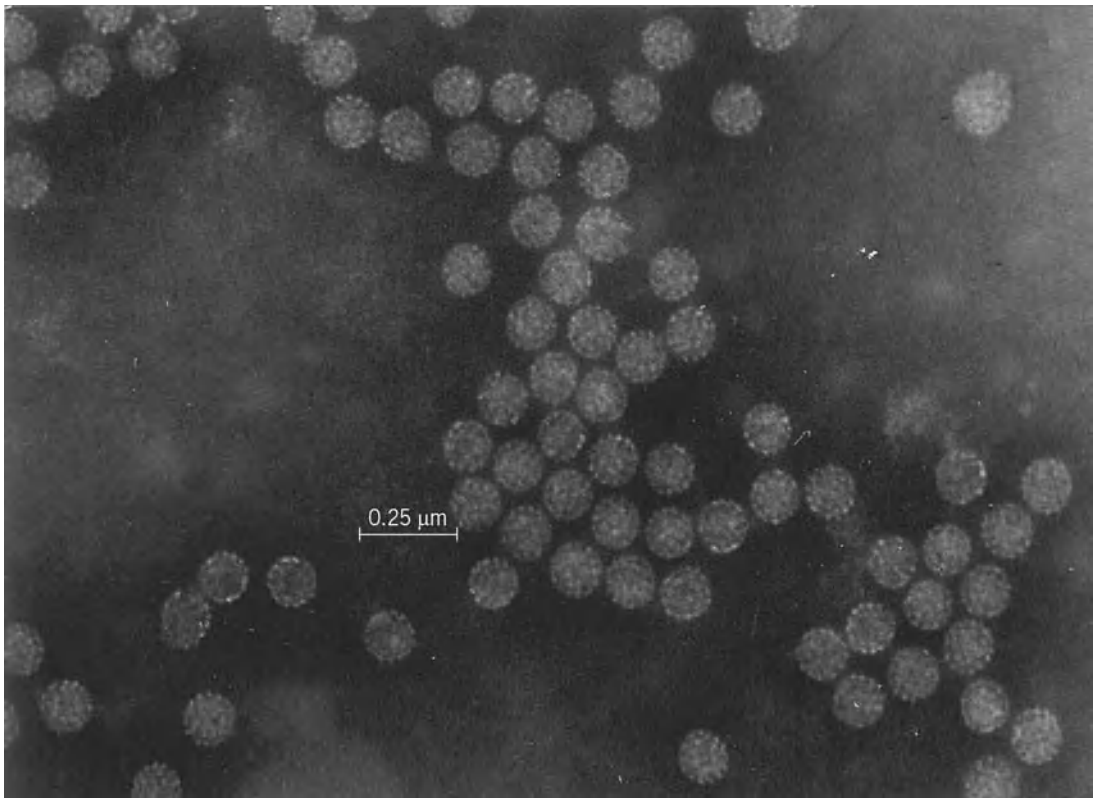


Fig. 13. Purified polyoma virus, negative stain 2% phosphotungstic acid buffered to pH 6. (*Courtesy of S. S. Breese, Jr.*)

and staining of tissues were developed which are adequate for a large range of observations on the complex detail of cell structure. The widely used fixatives are based on osmium tetroxide, potassium permanganate, or the aldehydes properly buffered to control pH. Embedding media were extended to a range of polymeric materials, the most successful being the epoxy resins and polyester resins. Staining of tissues either before or after sectioning can be very helpful to enhance contrast in very thin tissue sections. The general principle of staining for electron microscopy is to introduce heavy atoms such as lead or uranium into the tissue to make it more effective in scattering the electron illumination and thus to increase the visibility of those parts of the tissue to which the heavy atoms are attached. The specificity of stains available is limited, and the electron microscopist does not have available to him the range of specific staining materials with easily identifiable color contrasts that play such important roles in histological work with the light microscope. Among staining materials widely used are such heavy metal compounds as phosphotungstic acid, uranyl acetate, and lead hydroxide as well as the fixative osmium tetroxide.

There are special techniques for localization of specific entities in a cell. Immunochemical staining was developed by adapting the fluorescent antibody technique of light microscopy. Ferritin-labeled antibodies have been prepared that react with the homologous antigen and thus can be located in the cell by the increased contrast associated with this entity in the electron microscope image of the sectioned tissue. *See* FLUORESCENCE MICROSCOPE; IMMUNOCHEMISTRY.

Another technique adapted from light microscopy and applied with increased resolution is autoradiography. A tissue is labeled with a radioactive tracer material by injecting the animal or the nutrient medium before the tissue is taken. Tritium incorporated in thymidine or similar constituents is the most widely used label. The thin section containing the radioactive element is covered with a very thin layer of photographic emulsion and allowed to stand in the dark for weeks or months. Ionizing radiation exposes the photographic emulsion producing a latent image in the silver halide grains of the emulsion at those sites at which radioactive decay has occurred. After development and fixation of the photographic emulsion and staining of the tissue section, the composite section and developed photographic emulsion is examined in the electron microscope. The developed silver grains locate the sites at which the label was attached in the section with resolutions of better than 100 nm (Fig. 15). *See* AUTORADIOGRAPHY.

Great depth of focus, a characteristic of electron microscopes in general and of the scanning microscope in particular, gives the scanning electron microscope a decided advantage over the light microscope for topographical observations on biological structures (**Fig. 16**).

Nonbiological materials. The most severe limitation in observation of nonbiological materials is their bulk nature and the fact that only very small and extremely thin specimens can be examined. Since solid material often loses its bulk properties when reduced to

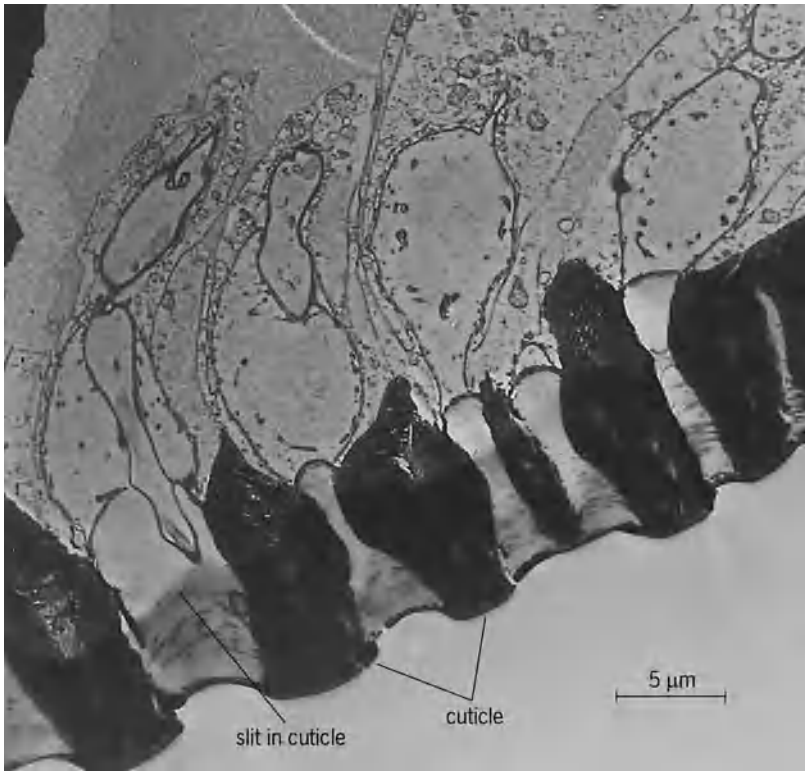


Fig. 14. Electron micrograph of longitudinal section of a spider's leg just above tarsal-metatarsal joint cut through the vibration receptor (ear). (*Courtesy of M. Salpeter*)

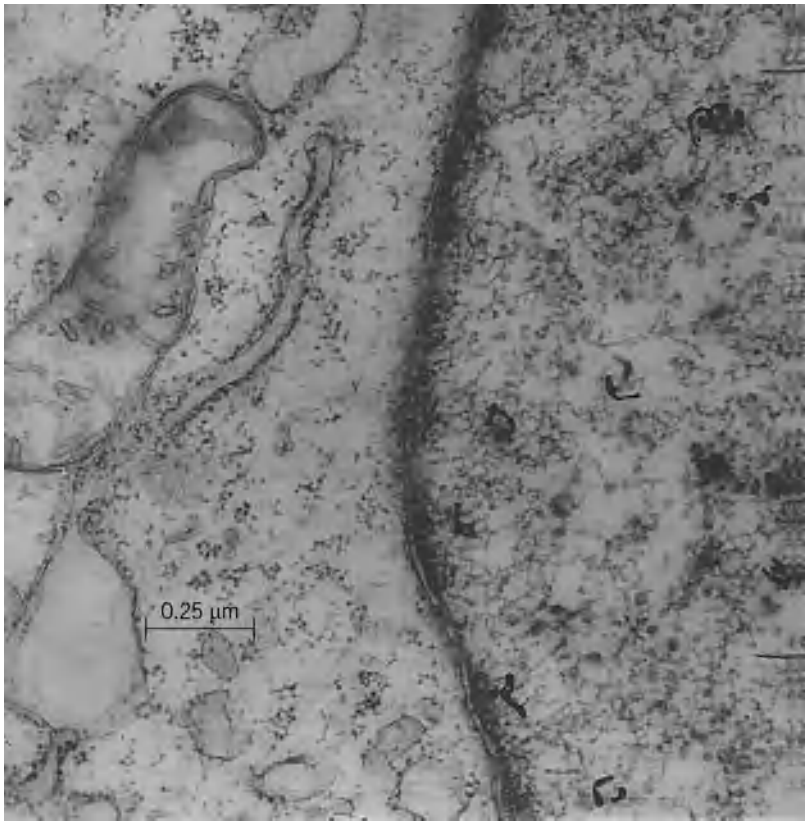


Fig. 15. Electron microscope autoradiogram of thin section of a mesenchymatous cell from the regenerating limb of adult newt (*Triturus*) labeled with tritiated thymidine. The dense black bodies over the nucleus on the right are developed silver grains, indicating radioactive decay in the tissue. (*Courtesy of M. Salpeter*)

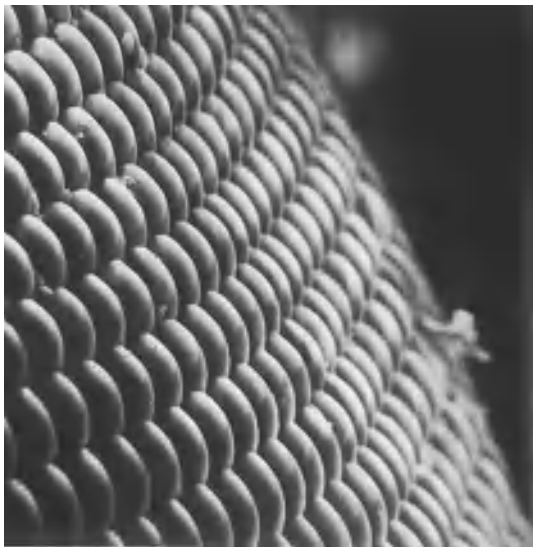


Fig. 16. Scanning electron micrograph of the surface of a fly's eye. (*RCA Research Labs, Princeton, New Jersey*)

very thin films or is too finely dispersed, high-voltage electron microscopes are used most actively for studies on nonbiological materials.

Particulate materials. Proper dispersion of the sample is the main consideration. If the material to be examined is a dry powder, a dispersion technique must be employed. One method used extensively is to mill the powder in a viscose solution of cellulose nitrate in amyl acetate until the aggregates are properly dispersed. The suspension is then diluted with amyl acetate to appropriate concentration for casting a suitably thin film with the particles suspended in it. Shadow-casting is usually employed to enhance the contrast or to give information on particle heights or shapes from the characteristics of their shadows.

Surfaces by replication. Surfaces of materials too thick to be placed directly in the electron microscope can be examined in the transmission electron microscope by preparing a thin film replica of the surface. The shadow-casting technique which produces contrast effects characteristic of the topography of the surface on which it is deposited greatly enhances the results obtainable by replication (**Fig. 17**).

A number of variations in method are employed in producing suitable replicas of surfaces. Each type of surface requires the proper choice from a variety of techniques and materials developed for application to different surfaces. The materials used for thin-film replicas include cellulose nitrate, Formvar, and evaporated carbon films. Carbon replicas are preferred for their greater strength and stability. A relatively smooth surface may be replicated directly by coating a film of Formvar on it by flooding the surface with a 0.5% solution of Formvar in ethylene dichloride. For greater ease of stripping, the replica may be backed by coating a cellulose nitrate film on the Formvar film. The backed replica can be wet-stripped, either by allowing the plastic film to float off on water if it will easily separate from

Fig. 17. Replica of a fractured surface of a magnesia ceramic. (*Courtesy of T. T. Sheng*)

the surface being replicated or the material being replicated can be dissolved away with a suitable solvent or acid to release replicating film. The replica film is mounted on specimen grids with the replication surface up, shadowed, and the backing film, if present, dissolved away in a solvent that does not attack the replica film. In the case of a Formvar replica backed with cellulose nitrate, amyl acetate is satisfactory. Dry stripping is accomplished by sandwiching the specimen grids between the film and Scotch tape and pulling off the film. Sometimes a stripping layer of a hydrophilic material is coated on the surface to be replicated. A detergent material such as Victawet can be evaporated onto the surface to be replicated, the metal shadowing deposited directly over the Victawet coat in the same vacuum chamber and, to complete the replica, a carbon substrate deposited by evaporation. This pre-shadowed replica is perhaps the most satisfactory and is used if the material to be replicated is suitable. On very rough surfaces, such as wood and fractured materials, it may be necessary to form a thick negative replica of a material such as polystyrene which can be molded to the surface by heat and pressure. This can then be replicated by shadowing, evaporating a carbon film on it, and releasing the replica by dissolving the polystyrene away with a suitable solvent. Materials containing water have been successfully replicated by using polyvinyl alcohol as the material for the negative replica. Purified crystalline virus in aqueous suspension and similar materials can be replicated by a rapid freezing technique. A frozen suspension is fractured to expose a fresh surface, placed in a vacuum chamber, and warmed to -110°F (-80°C) until the ice matrix is etched away slightly, leaving the material to be replicated in relief. This is cooled again to liquid nitrogen temperatures, the surface is replicated by shadowing, and a carbon substrate is evaporated on.

Crystalline films. The electron microscope finds wide application in investigations of crystalline materials. Two techniques are used to obtain specimens of bulk materials thin enough for direct transmission microscopy. Bulk samples can be reduced to a foil of 0.004–0.008 in. (0.1–0.2 mm) thick by processes such as jet machining, spark machining, or cutting and grinding. A final thinning is carried out by chemical or electrochemical methods which process the specimen to thicknesses of 0.1–0.2 μm for direct observation in the electron microscope. The other method utilizes various methods of deposition to grow films of suitable thicknesses and characteristics (**Fig. 18**).

Crystalline films obtained by these methods will strongly diffract the electron beam, and image contrast is determined by the nature of these diffraction interactions. When any part of the crystalline specimen is oriented so that a strong Bragg reflection occurs, the diffracted beam is scattered outside of the

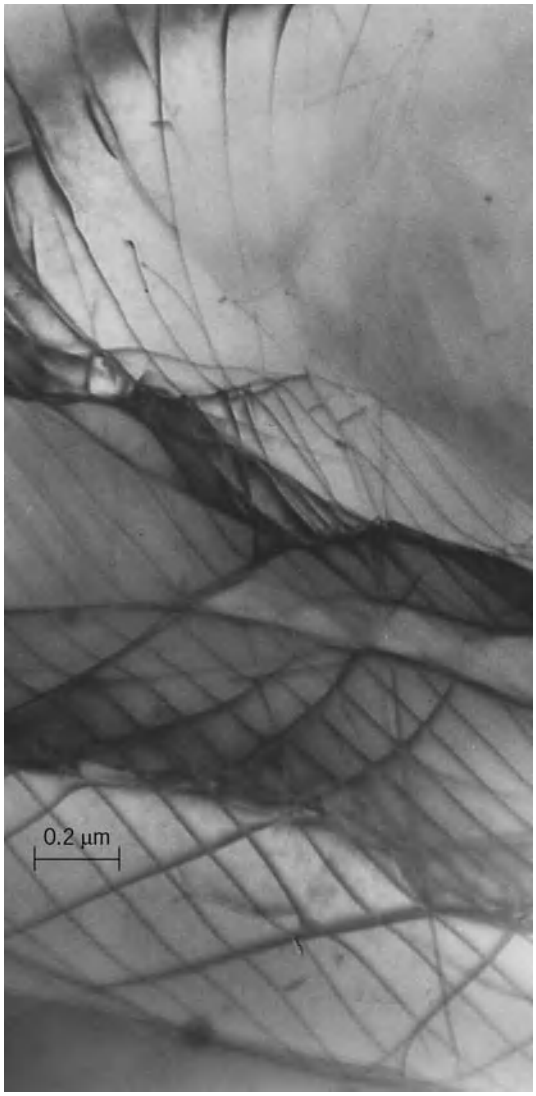


Fig. 18. Thin magnesia foil showing dislocation networks. (*Courtesy of T. T. Sheng*)

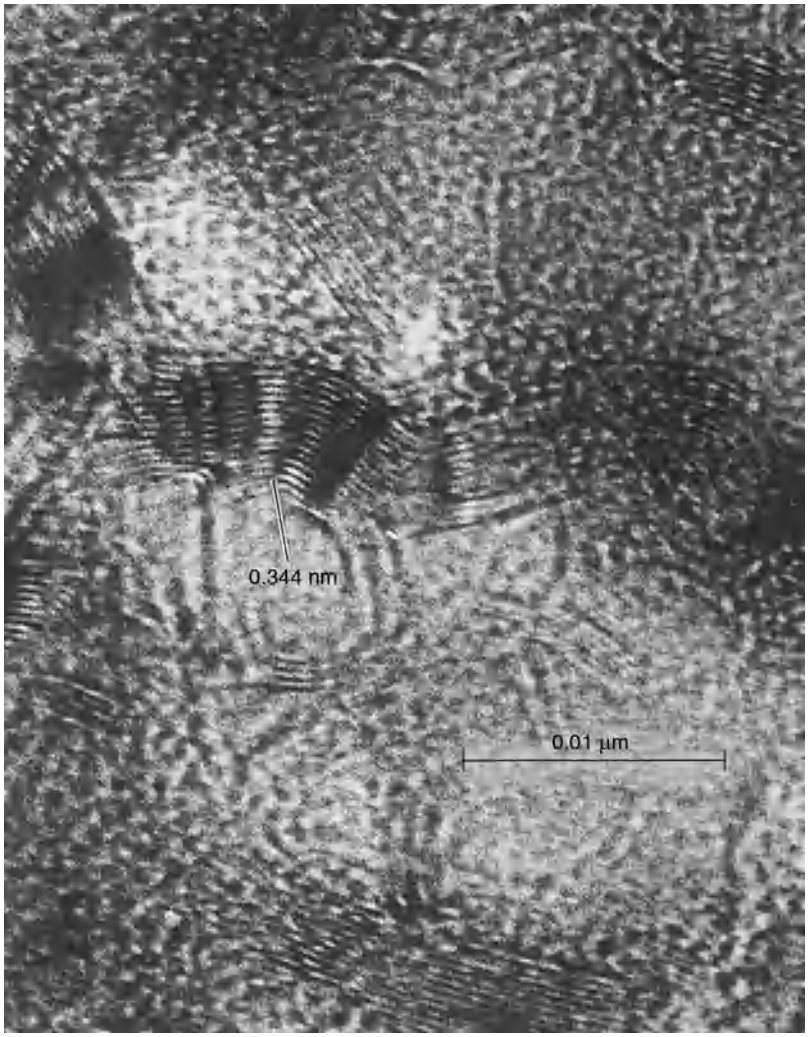


Fig. 19. Graphitized carbon showing lattice resolution of the 0.344-nm lattice spacing.

objective aperture and is no longer available in the imaging process. The image of this specimen then appears dark in areas corresponding to regions where Bragg reflection has occurred, so that high contrast is produced in the electron micrograph. Thus the crystal structure and its imperfections produce contrast effects in the imaging process that make this application of electron microscopy an important tool for study of defect structure in solids and the structure of thin films.

The electron optics of the electron microscope can be adjusted by a simple change of the focal length of one lens to produce the electron diffraction pattern from the same small area of specimen under observation with the normal electron optics for microscopy. *See* ELECTRON DIFFRACTION.

When the aperture stop of the objective lens is adjusted to allow one or two of the diffracted beams to contribute to the image formation in addition to the undeviated electron beam, a lattice resolution image is obtained that resolves lattice planes in crystalline material. Very thin deposited films or crystals which grow as thin platelets are required as specimens for these extremely high-resolution observations. By careful adjustment two or more lattice planes can be very highly resolved to spacings of less than 0.2 nm. If there are imperfections in the lattice, these defects will also be observed under certain orientations of the crystal. Graphitized carbon can also produce lattice resolution images as shown in **Fig. 19**.

Limitations on observations. Specimen artifacts, instrumental defects, and image characteristics are factors in the limitations on observations.

Specimen artifacts. Artifacts introduced in biological specimen preparation are primarily caused by the dehydration process necessary to bring the material under the vacuum of the electron microscope. Where effects such as distortion from surface tension and chemical changes from changing ionic concentration are severe, observations must be supplemented with other methods and careful evaluation made of the extent to which preparation artifacts may have changed a living, dynamic system.

The specimen will also be altered when exposed to the electron beam. Careful procedures will eliminate changes in the micromorphology of the material. Irradiation of a specimen with a low-intensity beam will cause chemical changes in the material, so that, for example, cellulose nitrate film becomes insoluble in all solvents that normally dissolve it. Also, the polymerized butyl methacrylate matrix used to embed tissue sublimes out when thin sections are exposed to low-intensity beams, leaving the cytological material and enhancing the contrast observed in the electron micrograph. On the other hand, exposure to high-intensity electron beams can produce a rise in temperature of hundreds of degrees Celsius in the specimen material. If a thin section with a polymeric embedding medium is subjected directly to a high-intensity beam, the polymer will melt, causing considerable distortion in the section. Thermal effects can cause considerable change in some materials, and care must be exercised in avoiding irradiation with too-high-intensity beams. Examination at high electron optical magnifications ($\times 1{,}000{,}000$) requires very intense beams, and special illuminating systems such as those employing a double condenser are utilized to minimize heating effects.

Contamination of the specimen also occurs under exposure to the electron beam. A layer of low-density, inert, carbonaceous material is laid down, sometimes at rates as high as several angstroms per second. The rate of deposition is a function of beam intensity, specimen temperature, and the amount of hydrocarbons present from pump oils and gasket grease. By surrounding the specimen with a chamber kept at -110°F (-80°C) the deposition can be reduced to a minimum. Special chambers designed to achieve this effect are essential for all high-resolution work where detail and contrast would soon be obliterated by deposition of a contamination layer.

Instrumental defects. High-resolution electron microscopes can be adjusted to attain an instrumental resolving power of better than 0.5 nm. Resolution obtained is usually limited by the resolution inherent in the specimen, especially for biological materials. Achievement of such a high level of instrumental performance requires a high degree of skill in the

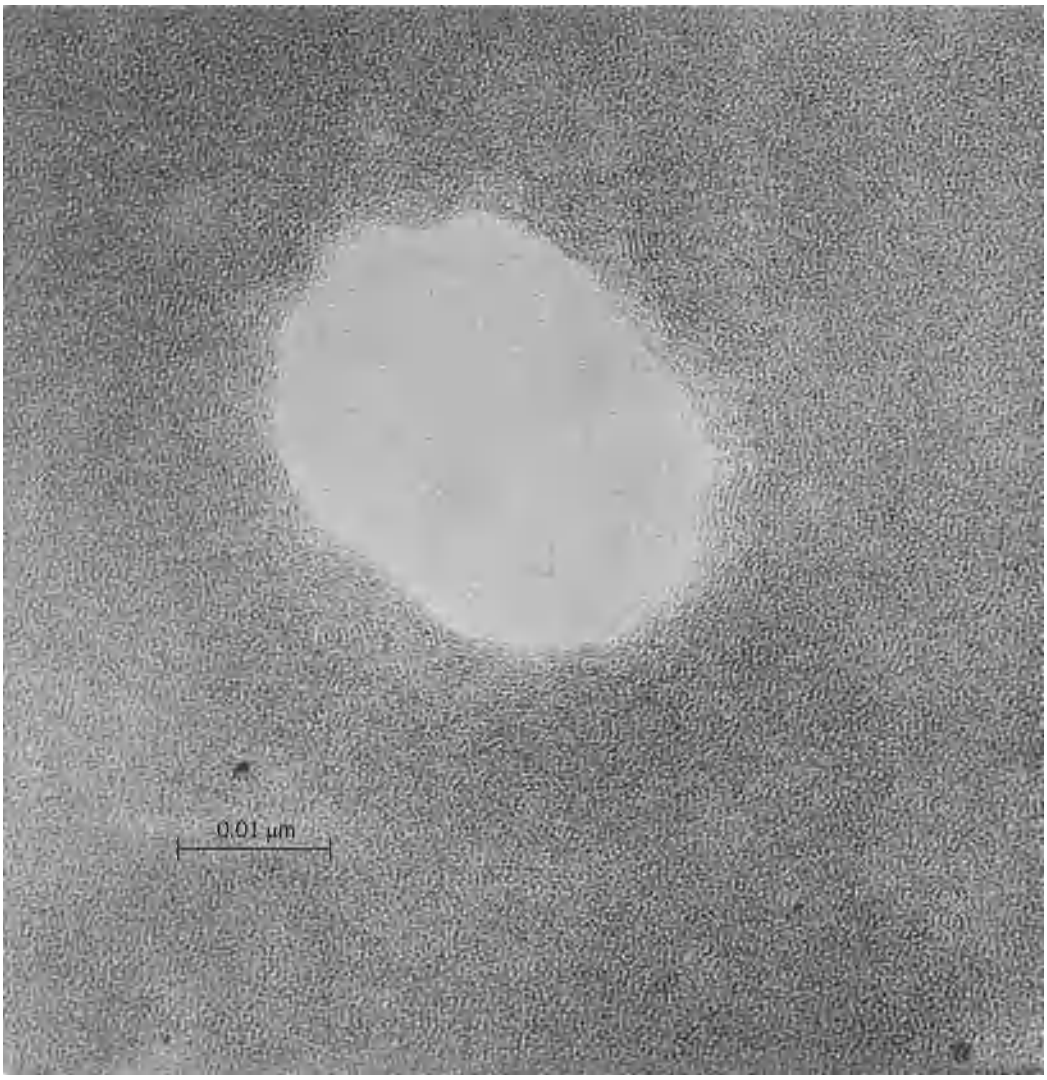


Fig. 20. Evaporated carbon film with small hole. Very slight underfocus shows Fresnel fringe inside hole. Phase contrast produces structural detail in film.

adjustment by the electron microscopist. The effect which first limits the resolving power even in a well-adjusted instrument is a deviation from axial symmetry in the objective lens producing an astigmatic image. Correction is made by introducing a compensating cylindrical field at right angles to the existing one, either by shims of ferromagnetic material oriented at the correct azimuth in the gap of the pole piece or by an electrostatic or magnetic field of correct orientation and magnitude. Externally adjustable stigmators, of both magnetic and electrostatic types, are available which enable the operator to compensate the electron microscope and obtain resolutions of 0.5 nm or better.

Contrast in the electron image can be enhanced by using a stop in the objective lens limiting the angular apertures that are necessary when relatively thick specimens such as sections and replicas are examined. However, an objective aperture can cause deterioration of resolving power, especially if asymmetric contamination forms on the aperture and causes increased astigmatism in the image. Good practice requires careful checking on instrumental performance using test specimens of a noncrystalline material such as carbon film with small holes to check asymmetry or astigmatism and the phase contrast in the out-of-focus image.

Image characteristics. One of the most prominent characteristics of the electron microscope is the contour effect. An instrument with a good, small effective source of illumination produces an image where Fresnel diffraction is clearly observable when the objective lens is out of true focus. In practice only the first-order fringe is strongly in evidence outlining any sharp boundaries in the image; hence the so-called contour effect. When the objective-lens current is too low, so that the focal length of the lens is too long, the image is said to be underfocused. At an edge where there is a sharp change in contrast, the outside of the edge is bordered by a dark contour. Where focal length is too short, the image is said to be overfocused and the contour now appears inside the edge. At true focus the Fresnel fringe vanishes and a soft image is obtained. Most electron

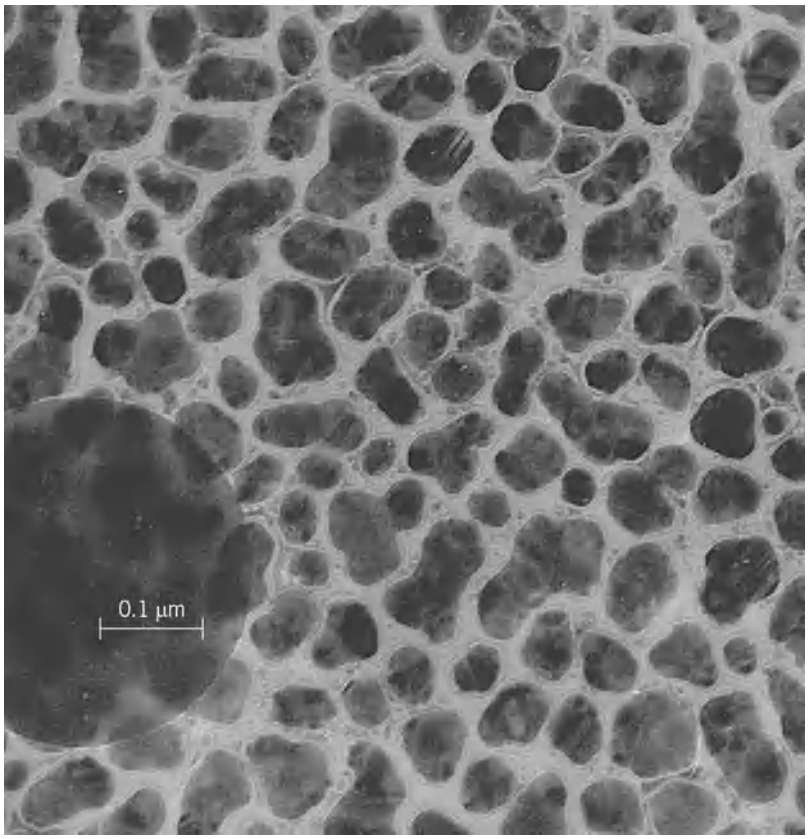


Fig. 21. Transmission electron micrograph of evaporated copper deposited on mica and transferred to a carbon substrate. Contrast effects are caused by different orientations and lattice imperfections.

micrographs are taken at a slightly underfocus setting of the objective lens, but for critical high-resolution work a focal series should be taken. Again, a carbon film with holes is a good test specimen (**Fig. 20**).

Crystalline material can produce anomalous contrast effects. If a crystal is oriented so that the incident electron beam is at an angle satisfying the Bragg condition for diffraction, the crystal can scatter the total incident beam into the Bragg angle while the same crystal when not oriented at the Bragg angle will scatter a fraction of the incident beam, depending on its thickness and its density.

The images obtained under the two different conditions of orientation can be of very different photographic density, and care must be taken in interpreting contrast effects of images of crystalline materials. **Figure 21** shows the strong contrast effects which can be produced by the electron interference interactions when scattered from a crystal lattice. The particles are of approximately uniform thickness, and the observed contrast effects are caused by different orientations and lattice imperfections.

The extension of the working resolving power of the electron microscope shows the significant role that diffraction phenomena play in the imaging characteristics of the instrument. Proper understanding of image contrast can be achieved only with a wave optical treatment of the imaging process. At levels of resolution of less than 0.5 nm, phase contrast plays a most important role. Phase contrast occurs with a defocusing of the image to adjust for the phase shift of the scattered electron wave by the spherical aberration of the objective lens. Phase contrast produces the contrast observed in the image of the carbon film shown in Fig. 19. Crystalline structures such as thin graphite layers show strong phase-contrast effects. The lattice resolution method is essentially a phase-contrast image. Biological macromolecular detail as small as 0.4–0.5 nm can be resolved by phase-contrast imaging when the specimen is suspended on the edge of a fiber support. Benjamin M. Siegel

Bibliography. P. B. Bell, Jr., *Scanning Electron Microscopy of Cells in Culture*, 1984; H. Bethge and J. Heydenreich (eds.), *Electron Microscopy in Solid State Physics*, 1987; R. Dickersin, *Diagnostic Electron Microscopy: A Text-Atlas*, 2d ed., 2000; V. A. Drits, *Electron Diffraction and High-Resolution Electron Microscopy of Mineral Structures*, 1987; H. Fujita (ed.), *History of Electron Microscopy*, 1987; P. Goodhew and F. Humphreys (eds.), *Electron Microscopy and Analysis*, 3d ed., 2000.

Electron motion in vacuum

Motion of electrons in a space freed sufficiently from matter so that collisions with other particles play a negligible role. The motion of electrons in vacuum is controlled by electric and magnetic fields whose force on the electrons is proportional to their magnitude. Electric and magnetic fields may arise from the presence of electrodes, currents, and magnets surrounding the evacuated space in which a particular electron moves, as well as from the presence of other charged particles within this space. This article deals with the nonrelativistic motion of electrons in static electric and magnetic fields, the effect of space charge on the electron paths, and motion in the time-varying fields that are encountered, for example, in the cathode-ray oscilloscope. For a discussion of the motion of electrons and other charged particles in cases where the velocities approach that of light *see* RELATIVISTIC ELECTRODYNAMICS

Static electric fields. An electron moving in a plane of symmetry of an electric field will remain indefinitely in that plane because the electrical forces acting on the electron lie within the plane. A plane of symmetry is here defined as one for which the mirror image of the potential distribution in front of the plane coincides with the potential distribution in back of the plane. Newton's second law of motion for the electron moving in the plane with rectangular coordinates x, y takes the form of Eqs. (1) and (2). Here $-e$ and m are the charge and mass of the

$$m\frac{d^2x}{dt^2} = -eE_x = e\frac{\partial\phi}{\partial x} \tag{1}$$

$$m\frac{d^2y}{dt^2} = -eE_y = e\frac{\partial\phi}{\partial y} \tag{2}$$

electron, t is time, E_x and E_y are the x and y components of the electric field, and ϕ is the electric

potential, normalized as stated by Eq. (3). Newton's

$$\frac{m}{2}\left[\left(\frac{dx}{dt}\right)^2+\left(\frac{dy}{dt}\right)^2\right]=e\phi \tag{3}$$

law as stated in Eqs. (1) and (2) implies that the speed of the electron is small enough so that its mass can be regarded as constant. For an electron having an energy of 10 kilovolts, the increase in mass is about 1%.

Path equation. Elimination of time from Eqs. (1) and (2) leads to the path equation (4). If potential distri-

$$\frac{d^2y}{dx^2}=\frac{1}{2\phi}\left[1+\left(\frac{dy}{dx}\right)^2\right]\left(\frac{\partial\phi}{\partial y}-\frac{dy}{dx}\frac{\partial\phi}{\partial x}\right) \tag{4}$$

bution $\phi(x,y)$ is known, and if position and velocity of the electron at one point within the field are also known, the electron's path can be determined by integrating Eq. (4). For simple electrode structures, the potential distribution can be determined analytically by solving Laplace's equation in the form of Eq. (5). More generally, it can be found by construct-

$$\nabla^2\phi=\frac{\partial^2\phi}{\partial x^2}+\frac{\partial^2\phi}{\partial y^2}+\frac{\partial^2\phi}{\partial z^2}=0 \tag{5}$$

ing a large-scale model of the electrode structure and immersing it in an electrolytic tank so that the surface of the liquid (usually slightly acidified tap water) coincides with the plane of symmetry of interest. With potentials proportional to the actual potentials applied to the model electrodes, an equipotential line on the surface can be found by determining the points at which a probe at the potential in question draws no current.

The path equation (4) leads to Eq. (6) for the ra-

$$R=\frac{2\phi}{\partial\phi/\partial n} \tag{6}$$

dius of curvature R of the paths. Here $-\partial\phi/\partial n$ is the component of the electric field normal to the electron path. If an equipotential plot has been prepared, this relation permits graphical plotting of an electron path (**Fig. 1**). The path is approximated by a series of circular arcs, the radius of curvature between successive equipotential lines being computed from the preceding relation for R.

Paraxial-ray equation. Electrostatic fields having not only a plane of symmetry but an axis of symmetry, which represents the intersection of an infinite family of planes of symmetry, have particular practical importance. Equation (4) still applies here, provided that y is identified with r, the distance from the axis, and x with z, the distance measured along the axis. The Laplace equation in the new coordinates z, r takes the form of Eq. (7). Equation (7) is solved quite

$$\frac{\partial^2\phi}{\partial r^2}+\frac{1}{r}\frac{\partial\phi}{\partial r}+\frac{\partial^2\phi}{\partial z^2}=0 \tag{7}$$

generally by the series shown as Eq. (8), where $\Phi(z)$

$$\phi(z,r)=\Phi(z)-\frac{r^2}{4}\frac{d^2\Phi}{dz^2}+\frac{r^4}{64}\frac{d^4\Phi}{dz^4}\cdots \tag{8}$$

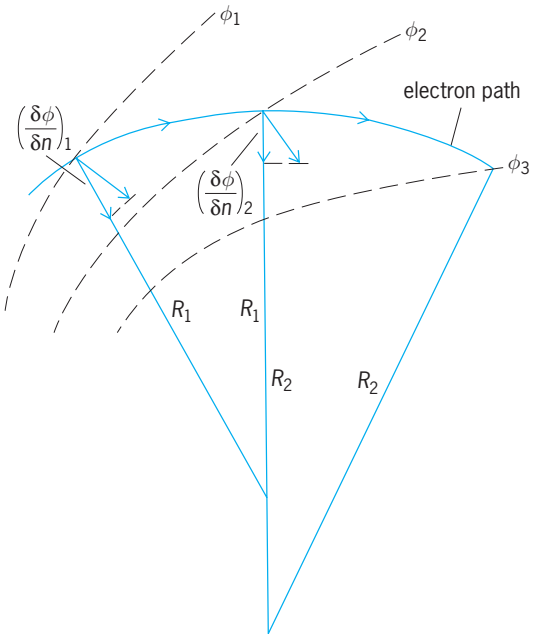


Fig. 1. Path plotting in an electrostatic field. Broken lines ϕ_1, ϕ_2, and ϕ_3 are equipotential lines. R_1 and R_2 are radii or curvature of the electron path.

is the potential on the axis of symmetry. Thus, the potential everywhere within the axially symmetric electrode structure is fully determined by the potential variation along the axis. Substitution of ϕ from Eq. (8) into Eq. (4) with retention of terms of the first order only in r and dr/dt leads to the paraxial ray equation (9). Equation (9) applies to electrons

$$\frac{d^2r}{dz^2}+\frac{1}{2}\frac{d\Phi}{dz}\frac{dr}{dz}+\frac{1}{4}\frac{d^2\Phi}{dz^2}r=0 \tag{9}$$

whose paths depart relatively little, both in slope and in distance, from the axis of the field.

Equation (9) is linear in r. Thus, if one electron path intersects the axis in two points, all electron paths passing through one of the points also pass through the other. In brief, the electric field images the one point into the other. It can be shown further that this imaging property is not limited to the axis but applies to extended areas about the axis, so that the axially symmetric electric field acts on the paths of electrons in the same manner as glass lenses act on light rays. Departures of the exact path from the paraxial equation result in image defects or

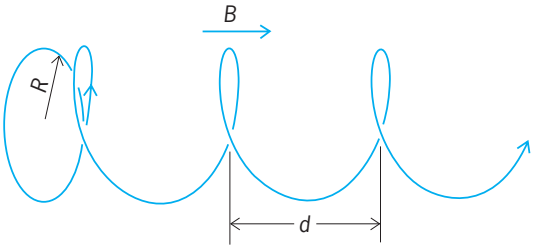


Fig. 2. Motion of an electron in a uniform magnetic field. Its path is in general a helix with pitch d, radius R, and axis parallel to the field.

aberrations similar in character to those observed for glass lenses. *See* ELECTROSTATIC LENS.

Magnetic fields. A magnetic field exerts on an electron of velocity **v** a force **F** perpendicular to both the direction of motion and the direction of the field. In vector notation this Lorentz force is given by Eq. (10). Here, **b** is the magnetic induction.

$$\mathbf{F} = -e(\mathbf{v} \times \mathbf{b}) \tag{10}$$

The components of the Lorentz force are given by Eqs. (11)–(13). Because this force is perpendicular

$$F_x = e\left(b_y \frac{dz}{dt} - b_z \frac{dy}{dt}\right) \tag{11}$$

$$F_y = e\left(b_z \frac{dx}{dt} - b_x \frac{dz}{dt}\right) \tag{12}$$

$$F_z = e\left(b_x \frac{dy}{dt} - b_y \frac{dx}{dt}\right) \tag{13}$$

to the direction of motion, it does no work on the electron, whose velocity consequently remains unchanged in magnitude. A uniform magnetic field parallel to the z axis is described by Eqs. (14). Newton's

$$b_z = B \qquad b_x = b_y = 0 \tag{14}$$

second law leads to a constant z component of the velocity. The magnitude of the velocity component v_{xy} in the xy plane is similarly constant, the square of the total velocity being equal to the sum of the squares of the components, For the motion projected on the xy plane, Newton's second law thus takes the form of Eq. (15). Here R is the radius of curvature of the

$$\frac{mv_{xy}^2}{R} = ev_{xy}B \tag{15}$$

projected path. R is seen to be a constant, so that the projected path is a circle with radius given by Eq. (16). Here α is the angle which the electron path

$$R = \frac{mv_{xy}}{eB} = \frac{\sin\alpha}{B}\left(\frac{2m\phi}{e}\right)^{1/2} = \frac{3.37\phi^{1/2}}{B}\sin\alpha \tag{16}$$

makes with the field direction, and ϕ is the accelerating potential of the electrons. If B is measured in gauss, and ϕ in volts, R is in centimeters. The frequency with which the circle is traversed by the electron is given by Eq. (17). This frequency, in s^{-1},

$$f = \frac{v_{xy}}{2\pi R} = \frac{eB}{2\pi m} = 2.8 \times 10^6 B \tag{17}$$

the cyclotron frequency, thus depends only on the magnetic field strength.

The complete motion of the electron (**Fig. 2**) is thus a helix about a magnetic line of force, with a pitch d in centimeters given by Eq. (18). All electrons passing through a point with equal axial velocity components pass through a series of points separated by d on the same magnetic field line. An initially divergent electron beam is held together by a uniform magnetic field, because an electron path which intersects a particular field line can never depart from it by more than twice the radius R of the helix. Uniform magnetic fields are widely used for keeping electron beams from spreading, for example, in traveling-wave tubes and klystrons. Electron beams will also follow magnetic field lines, if these are gently curved. This property is utilized in the magnetic deflection of beams in certain television camera tubes, such as the image orthicon and the vidicon. In these tubes, a weak transverse

$$d = \frac{v_z}{f} = \frac{2\pi\cos\alpha}{B}\left(\frac{2m\phi}{e}\right)^{1/2} = 21.08\,\frac{\phi^{1/2}}{B}\cos\alpha \tag{18}$$

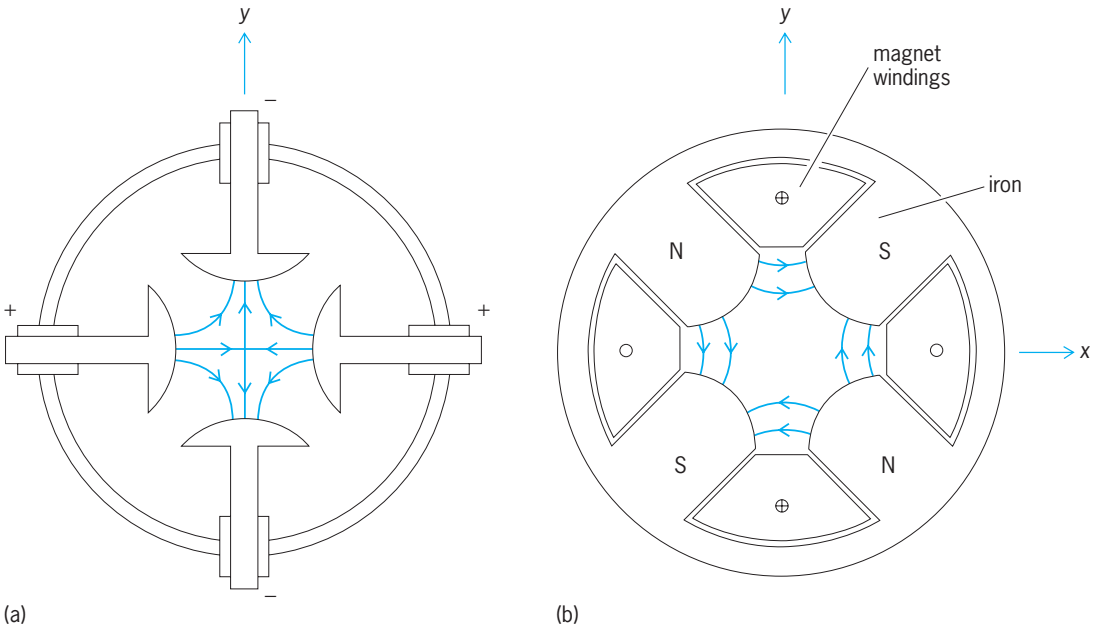


Fig. 3. Two types of quadrupole lens. (*a*) Electrostatic quadrupole. (*b*) Magnetic quadrupole.

magnetic deflection field is superposed on a strong longitudinal magnetic focusing field. *See* TELEVISION CAMERA TUBE; VAN ALLEN RADIATION.

Motion in nonuniform magnetic fields with axial symmetry is conveniently treated as a special case of motion in combined electric and magnetic fields.

Combined fields with axial symmetry. Motion is now expressed by Eqs. (19)–(21). Coordinates z, r, and θ

$$m\frac{d^2z}{dt^2} = e\left(\frac{\partial\phi}{\partial z} + b_r r\frac{d\theta}{dt}\right) \tag{19}$$

$$m\left[\frac{d^2r}{dt^2} - r\left(\frac{d\theta}{dt}\right)^2\right] = e\left(\frac{\partial\phi}{\partial r} - b_z r\frac{d\theta}{dt}\right) \tag{20}$$

$$m\frac{1}{r}\frac{d}{dt}\left(r^2\frac{d\theta}{dt}\right) = e\left(b_z\frac{dr}{dt} - b_r\frac{dz}{dt}\right) \tag{21}$$

represent distance along the axis, perpendicular distance from the axis, and azimuthal angle about the axis. Terms b_z and b_r are the axial and radial components of the magnetic induction. From Eqs. (19)–(21), Eq. (22*a*), a path equation expressing the variation of the radial distance r with the axial distance z, is derived, with ϕ^* defined by Eq. (22*b*). Here ϕD^2 is a shorthand symbol for the last term in Eq. (22*b*) and C is, except for a universal multiplying constant, the angular momentum of the electron about the axis at a point where the magnetic field vanishes, given by Eq. (22*c*). Here A is the magnetic vector potential, which is numerically equal to the magnetic flux through a circle about the axis through the reference point divided by the circumference of that circle.

$$\frac{d^2r}{dz^2} = \frac{1}{2\phi^*}\left[1 + \left(\frac{dr}{dz}\right)^2\right]\left(\frac{\partial\phi^*}{\partial r} - \frac{dr}{dz}\frac{\partial\phi^*}{\partial z}\right) \tag{22a}$$

$$\phi^* = \phi\,(1 - D^2) = \phi - \left[\frac{C}{r} + \left(\frac{e}{2m}\right)^{1/2} A\right]^2 \tag{22b}$$

$$C = r^2\frac{d\theta}{dz}\phi^{1/2}\left[\left(\frac{dr}{dz}\right)^2 + r^2\left(\frac{d\theta}{dz}\right)^2 + 1\right]^{-1/2} - \left(\frac{e}{2M}\right)^{1/2} rA \tag{22c}$$

At the same time, the axiumth θ of the electron changes according to Eq. (23). Equations (22) can be solved by graphical and numerical methods useful for determining electron paths in electrostatic fields. The general paraxial equation is obtained by substituting the expansion of Eq. (24) into Eq. (22*b*).

$$\theta = \phi_0 + \int_{z0}^{z} \frac{D}{r(1 - D^2)^{1/2}} \times \left[1 + \left(\frac{dr}{dz}\right)^2\right]^{1/2} dz \tag{23}$$

$$A = \frac{r}{2}B(z) - \frac{r^3}{16}\frac{d^2B(z)}{dz^2}\cdots \tag{24}$$

Here $B(z)$ is the magnetic induction along the axis. Substitution of Eq. (24) and that for the electrostatic potential and retention of terms of the first order in r and $drPdz$ only lead to Eqs. (25). With B in gauss, ϕ in volts, and z in centimeters, $e/8m$ equals 0.022 volt/(gauss-cm)^2.

$$\frac{d^2r}{dz^2} + \frac{1}{2\Phi}\frac{d\Phi}{dz}\frac{dr}{dz} + \left(\frac{1}{4\Phi}\frac{d^2\Phi}{dz^2} + \frac{eB^2}{8m\Phi} - \frac{C^2}{\Phi r^4}\right)r = 0 \tag{25a}$$

$$\theta = \theta_0 + \int_{z0}^{z}\left[\frac{C}{r^2\Phi^{1/2}} + \left(\frac{e}{8m\Phi}\right)^{1/2} B\right]dz \tag{25b}$$

$$\frac{C}{r_0{}^2\Phi_0{}^{1/2}} = \left(\frac{d\theta}{dz}\right)_0 - \left(\frac{e}{8m\Phi_0}\right)^{1/2} B(z_0) \tag{25c}$$

Quadrupole fields. The principal use of axially symmetric electrostatic and magnetic fields is to converge electron pencils; thus, in the cathode-ray tube, electrons diverging from a point in front of the emitter, called the crossover, are converged to a small spot on the viewing screen. Similarly, in an electron-imaging device such as the electron microscope, electrons diverging from a point on an object are brought to focus at a corresponding point of the

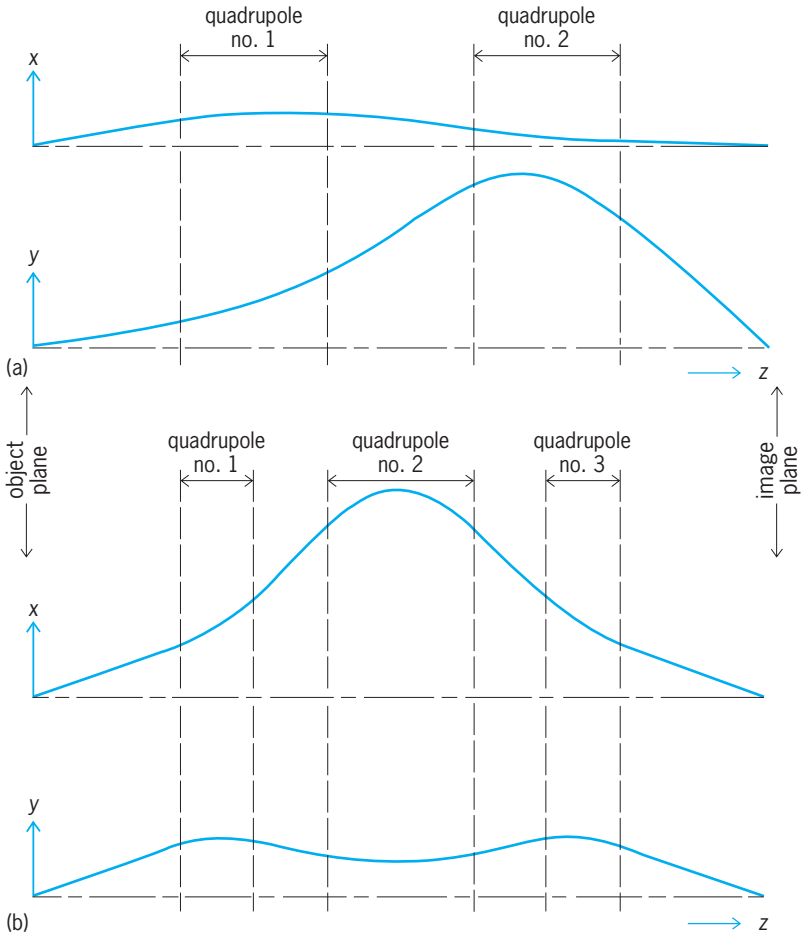


Fig. 4. Diagrammatic representation of electron paths in principal planes of image-forming symmetrical (*a*) doublet and (*b*) triplet.

image. However, this converging effect is only secondary; the primary effect of an axially symmetric electrostatic field is to accelerate or decelerate the electrons parallel to the axis, and the primary effect of an axially symmetric magnetic field is to give electrons diverging from a point on the axis a rotation about the axis. Correspondingly, the converging action or refractive power of a conventional electrostatic lens is proportional not to the first power but to the square of the ratio of the electrostatic field to the accelerating voltage, whereas the converging action of a conventional magnetic lens is proportional to the ratio of the square of the magnetic field to the accelerating potential at low electron energies (<0.5 MeV) and to the square of the ratio of the field to the accelerating potential at high electron energies (>0.5 MeV). Because breakdown phenomena place practical limits on the magnitude of electrostatic fields, and because pole-piece saturation limits (static) magnetic fields, conventional electron lenses are relatively ineffective in converging high-energy particles; they can be properly characterized as weak-focusing lenses.

Quadrupole fields, on the other hand, constitute strong-focusing lenses, insofar as their converging action is directly proportional to the electrostatic or magnetic field respectively. Consequently they (particularly magnetic quadrupole fields) have assumed great importance in the development of high-energy particle accelerators. In this application they permit the narrow confinement of the particle beam and consequently greatly reduce the cost of construction. Apart from the concentration and focusing of high-energy particle beams, strong-focusing lenses have found application as projector lenses (but not as objectives) of electron microscopes. Weak quadrupole fields (stigmators) also are used to compensate residual asymmetries in conventional electron lenses. *See* ELECTRON MICROSCOPE; PARTICLE ACCELERATOR.

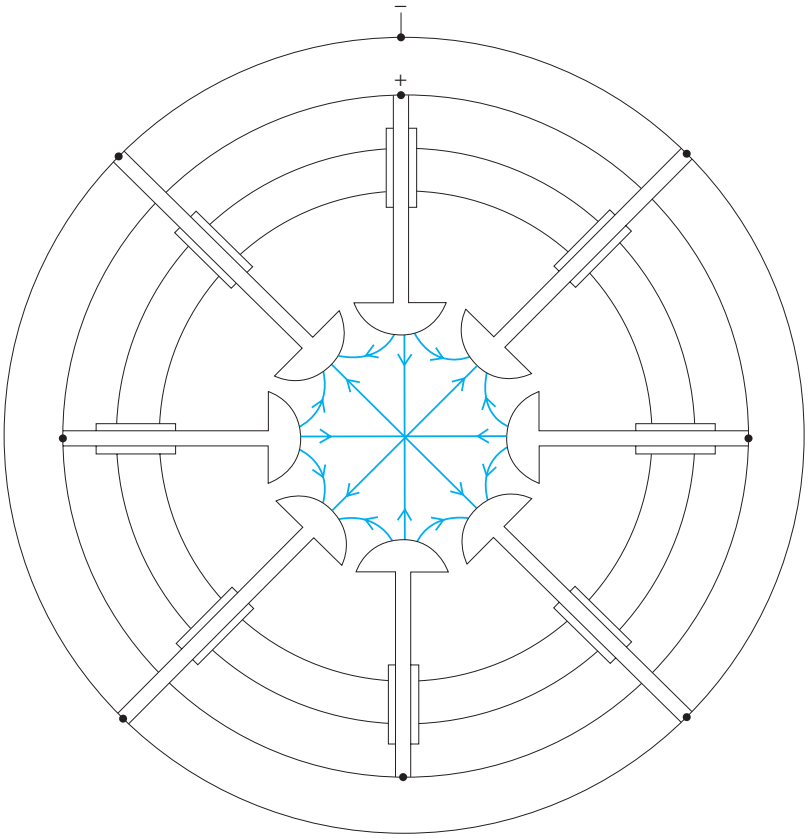


Fig. 5. Electrostatic octupole.

Figure 3 shows sections normal to the lens axis of an electrostatic and a magnetic quadrupole lens. In the first instance the electrostatic potential is symmetric with respect to the two principal (xz and yz) planes and antisymmetric with respect to the planes through the axis forming an angle of 45° with respect to the principal planes. In the second instance the magnetostatic potential is antisymmetric with respect to the principal planes and symmetric with respect to the two 45° planes. Fulfillment of Laplace's equation (5) now demands the form of Eq. (26) for

$$\phi(x, y, z) = \frac{a}{2}(x^2 - y^2) - \frac{d^2a}{dz^2}2(x^4 - y^4)\cdots \quad (26)$$

the electrostatic potential. Magnetic field B, on the other hand, may be written as Eqs. (27)–(29). In Eqs. (26)–(29), a and A are functions of z. Neglecting the

$$B_x = Ay - \frac{1}{6}\frac{d^2A}{dz^2}y(3x^2 + y^2)\cdots \quad (27)$$

$$B_y = Ax - \frac{1}{6}\frac{d^2A}{dz^2}x(x^2 + 3y^2)\cdots \quad (28)$$

$$B_z = 0 - \frac{1}{6}\frac{d^3A}{dz^3}xy(x^2 + y^2)\cdots \quad (29)$$

higher-order terms, which become negligible for a long lens, one finds that the force exerted by the field on the electron is given by Eqs. (30)–(32).

Electrostatic lens		*Magnetic lens*	
$F_x = -eax$	(30a)	$F_x = -eAv_2x$	(30b)
$F_y = eay$	(31a)	$F_y = eAv_zy$	(31b)
$F_z = 0$	(32a)	$F_z = 0$	(32b)

The force is thus, to a first approximation, entirely in a plane normal to the axis. In one principal plane it is such as to produce convergence (direction of force opposite to that of the displacement from the axis), and in the other, such as to produce divergence (direction of force equal to that of the displacement). Two successive quadrupole fields of opposite polarity (a doublet) are required to form a sharp, real image of an object on the axis. This image will have different magnifications in the two principal directions. If the image is to be undistorted, a triplet of three successive quadrupole fields is required. These electron paths are illustrated in **Fig. 4**.

The shaping of the poles influences only terms of the fifth and higher orders in the x and y coordinates in the field expressions. For long lenses all terms except the linear terms vanish if the poles are rectangular hyperbolic cylinders. In practice, circular-cylinder pole caps are more easily realized and give practically the same results. Like axially symmetric lenses, quadrupole lens systems exhibit aberrations, although these are more complex in character.

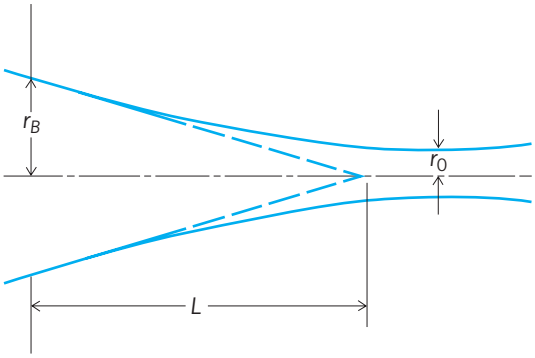


Fig. 6. Widening of electron beam by space-charge repulsion as it traverses from left to right.

Octupole fields (**Fig. 5**), which produce transverse forces proportional to the third power of the displacement from the axis, opposite in polarity in the principal planes and in the 45° planes, are commonly employed to correct aberrations of quadrupole lens systems.

Effect of space charge. Space charge of either positive or negative sign can influence the paths of electrons. Space charge of positive sign is formed by electron beams passing through an imperfectly evacuated space. The beam electrons collide with gas atoms and ionize them. The heavy ions remain in the path of the electron beams for some time and prevent it from spreading. The luminous nodular or thread beams which are produced in this way are favorite objects for demonstration. *See* SPACE CHARGE.

Electron beams in high vacuum, on the other hand, are subject only to the mutually repulsive forces between the electrons themselves. The repulsion is reduced, but never canceled, by the action of the magnetic fields that surround charges in motion; for two electrons moving with the same velocity v parallel to each other, the ratio of the magnetic attractive force to the electrostatic repulsive force is v^2/c^2, where c is the velocity of light. Hence, the magnetic force is significant only for electrons of very high energy.

The action of the remainder of the electrons in the beam upon any one electron can be approximated adequately by that of a continuous charge distribution equal to the average space-charge distribution. The behavior of the edge ray of a uniform circular beam of current I aimed at a point of convergence a distance L from the initial cross section of radius r_B may serve as an example (**Fig. 6**). If the variation of the potential along the axis of the beam is neglected (that is, if ϕ is assumed to be constant), and the charge density ρ is regarded as uniform within any beam cross section, ρ is given by Eq. (33) and the

$$\rho = \frac{I}{\pi r^2}\left(\frac{m}{2e\phi}\right)^{1/2} \tag{33}$$

path equation becomes Eq. (34). Here ϵ is the dielectric constant of vacuum. As the result of the repulsive force of space charge, the ray under consideration does not cross the axis, but reaches a minimum separation r_o from the axis and diverges from this point on. Integration of the differential equation, Eq. (34), gives the radius as Eq. (35). For example, if $r_B =$

$$\frac{d^2r}{dz^2} = \frac{\pi\rho}{\epsilon\phi}\,r = \left(\frac{m}{2e}\right)^{1/2}\frac{I}{\phi^{3/2}}\frac{1}{\epsilon r} \tag{34}$$

$$r_0 = r_B\,\exp\left[-\epsilon\left(\frac{e}{2m}\right)^{1/2}\frac{r_B{}^2\phi^{3/2}}{L^2I}\right]$$

$$= r_B\,\exp\left(-3.3\times10^{-5}\frac{r_B{}^2\phi^{3/2}}{L^2I}\right) \tag{35}$$

1 mm, $\phi = 10{,}000$ volts, $L = 10$ cm, and $I = 0.001$ ampere, then $r_o = 0.037$ mm.

Time-varying fields. In the preceding discussion, it was assumed that the electric and magnetic fields traversed by the electrons were constant in time. The total energy of the electron, or the sum of the kinetic energy and the potential energy, is then a constant. Because the potential energy is a function of position only, so is the kinetic energy. This is no longer true if the fields change appreciably in a period corresponding to the transit time of the electrons. *See* KLYSTRON; MAGNETRON; PARTICLE ACCELERATOR; TRAVELING-WAVE TUBE.

Beam deflection in the cathode-ray oscilloscope ceases to be proportional to the potential difference V applied to the deflection plates if V changes appreciably during the passage of the electron beam through the deflection field. If $V = V_0\cos 2\pi ft$, an integration of the transverse impulse impressed on the electron passing between two parallel plates of length l and separation d expresses the deflection angle α in the form of Eq. (36). Here ϕ is

$$\tan\alpha = \frac{V_0 l}{2\phi d}\frac{\sin u}{u}\cos(2\pi ft) \tag{36}$$

the accelerating potential of the beam, and t is the time the electron passes through the center of the

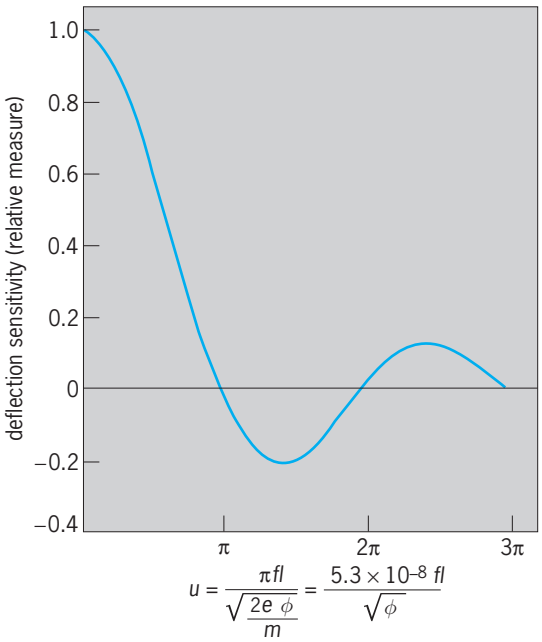


Fig. 7. Deflection sensitivity of cathode-ray oscilloscope as a function of frequency *f*.

deflection field. If f (the frequency of the applied voltage) is measured in s^{-1}, l in centimeters, and ϕ in volts, then Eq. (37) is applicable. The quantity

$$u = \pi f l/(2e\phi/m)^{1/2} = 5.3 \times 10^{-8} f l/\phi^{1/2} \quad (37)$$

$(\sin u)/u$ represents the ratio of the deflection sensitivity at frequency f to that at low frequencies ($f \rightarrow 0$) (**Fig. 7**). Thus, for a 10-kV beam and deflection plates 1 cm in length, the response is 95.4% at 1000 MHz (10^9 s^{-1}), 82.3% at 2000 MHz, 40.2% at 4000 MHz, and it drops to zero at about 5930 MHz. In this discussion, the deflection field is assumed to be a sharply cutoff uniform field, with the effects of fringe fields neglected. Edward G. Ramberg

Bibliography. D. A. DeWolf, *Basics of Electron Optics*, 1990; P. W. Hawkes and E. Kaspar (eds.), *Principles of Electron Optics*, 1996; M. Szilagy, *Electron and Ion Optics*, 1988.

Electron paramagnetic resonance (EPR) spectroscopy

The study of the resonant response to microwave- or radio-frequency radiation of paramagnetic materials placed in a magnetic field. It is sometimes referred to as electron spin resonance (ESR). Paramagnetic substances normally have an odd number of electrons or unpaired electrons, but sometimes electron paramagnetic resonance (EPR) is observed for ions or biradicals with an even number of electrons. EPR spectra are normally presented as plots of the first derivative of the energy absorbed from an oscillating magnetic field at a fixed microwave frequency versus the magnetic field strength. The dispersion may also be detected.

To overcome the intrinsic low sensitivity of the magnetic dipole transitions responsible for EPR, samples are placed in resonant cavities. Routine experiments are carried out in the steady state at a fixed microwave frequency of approximately 9 gigahertz by slowly sweeping the magnetic field through resonance. Free electrons resonate in a magnetic field of 3250 gauss (325 millitesla) at the microwave frequency of 9.1081 GHz, whereas organic free radicals resonate at slightly different magnetic fields characteristic of each particular molecule. *See* ELECTRON SPIN; MAGNETIC RESONANCE; PARAMAGNETISM.

Applications. EPR spectroscopy is used to determine the electronic structure of free radicals as well as transition-metal and rare-earth ions in a variety of substances, to study interactions between molecules, and to measure nuclear spins and magnetic moments. It is applied in the fields of physics, chemistry, biology, archeology, geology, and mineralogy. It is also used in the investigation of radiation-damaged materials and in radiation dosimetry.

The basic physics of transition-metal ions and rare-earth ions present in low concentrations in diamagnetic host crystals has provided a theoretical basis for how electronic structure is modified by the surrounding atoms. Particular applications include probing phase transitions in solids and studies of pairs and triads of magnetically interacting ions.

Applications of EPR in chemistry include characterization of free radicals, studies of organic reactions, and investigations of the electronic properties of paramagnetic inorganic molecules. Information obtained is used in the investigation of molecular structure. EPR is used widely in biology in the study of metal proteins, for nitroxide spin labeling, and in the investigation of radicals produced during reaction processes in proteins and other biomacromolecules. EPR has proved to be an important technique for interdisciplinary investigations of photosynthetic systems. By means of EPR, more than 20 proteins that function in the mitochondrial respiratory chains of mammals have been identified, and details regarding their electron transfer processes have been elucidated.

Sensitivity and resolution. EPR spectra due to pure transition-metal compounds or crystals show broad lines due to interactions between electron magnetic moments. In contrast, in nuclear magnetic resonance (NMR) the interactions between the much smaller nuclear magnetic moments are about 1 million times less intense. Small linewidths and usable resolution require use of dilute solutions in diamagnetic solvents or low concentrations of paramagnetic ions in isomorphous diamagnetic crystal lattices. Concentrations of typically less than 10^{16} molecules per cubic centimeter usually suffice for organic free radicals, but concentrations 10,000 times greater may be needed for inorganic molecules and for transition-metal and rare-earth ions in solids because of the greater linewidths, even for dilute systems. The number of spins required depends on the detectable limit of commercial spectrometers of 10 billion spins with a linewidth of 1 gauss (0.1 tesla). The larger the linewidths, the larger the number of spins required.

The observation of EPR spectra depends on spin-lattice relaxation, which is the exchange of magnetic energy with the thermal motion of the crystal or molecule. For transition-metal ions and rare-earth ions, experiments often require operation at or near liquid helium temperature (4 K; −269°C; −452°F). Organic free radicals can usually be studied successfully at room temperature. *See* MAGNETIC RELAXATION.

Solids. EPR spectra from single crystals clearly provide the greatest amount of information. These include crystals containing small concentrations of paramagnetic ions substituting for the regular ions in the crystal or, for organic molecules, small fractions of free radicals produced by ionizing radiation. Spectra which may contain up to several hundred lines are often highly anisotropic; that is, they change with the orientation of the magnetic field direction in the crystal. Transition-metal-ion and rare-earth-ion EPR spectra in crystals are generally much more anisotropic than free radicals due to the intrinsic anisotropy of the electron magnetic moments, and of other effects that are important when there is more than one unpaired electron.

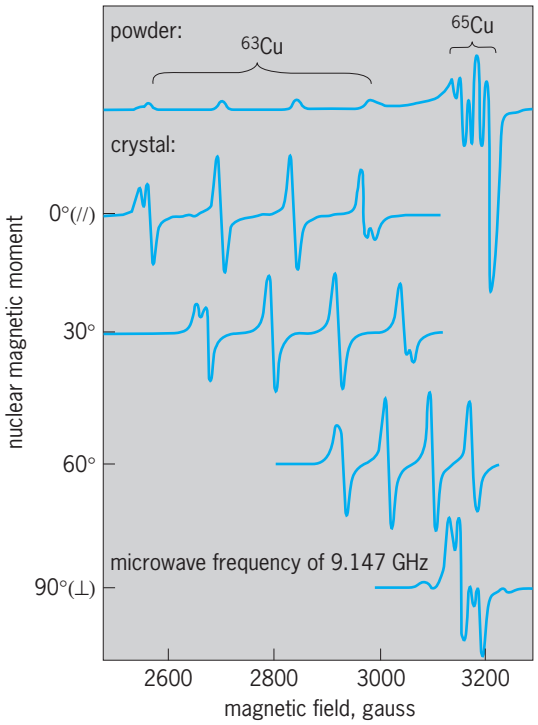


Fig. 1. Powder and single-crystal EPR spectra due to Cu^{2+} ions in calcium cadmium acetate hexahydrate at room temperature. The single-crystal spectra, with the magnetic field making angles of 0°, 30°, 60°, and 90° with the symmetry axis (//), illustrate anisotropy of the electron magnetic moment and the four-line copper hyperfine structure. The presence of the two copper isotopes, ^{63}Cu and ^{65}Cu, is revealed by the partly resolved features. The outer lines belong to ^{65}Cu, which has the larger nuclear magnetic moment. The lack of a characteristic four-line pattern at 90° is well understood. The powder spectra show absorptionlike features associated with 0° crystal spectra and evidence for the second copper isotope (^{65}Cu). (*After J. R. Pilbrow, Transition Ion Paramagnetic Resonance, Oxford University Press, 1990, © J. R. Pilbrow*)

The occurrence of many lines is due to interactions of the orbital motions of electrons with the electric potential of the local surrounding atoms, and to hyperfine interactions between the paramagnetic electrons and nuclear magnetic moments of the paramagnetic ion and surrounding atoms. In the case of free radicals, symmetric or nearly symmetric characteristic hyperfine patterns are observed. From knowledge of hyperfine interactions with nuclei whose spins and magnetic moments are known, the electron distribution throughout a molecule may be determined. Since hyperfine interactions vary as the reciprocal of the cube of the distance between the center of the free radical and the nucleus, structural information may be obtained in addition to electron densities.

The relative values of nuclear moments of different isotopes of a given element often can be determined from the ratio of the hyperfine splittings produced in the same chemical environment (**Fig. 1**).

Liquids and motional averaging. Spectra in solution due to free radicals are often quite simple as a result of motional averaging, and this clearly gives less information than would be obtained from a single-crystal investigation. Linewidths are very narrow (approximately 0.1 gauss or smaller). By varying the temperature above and below room temperature, EPR spectra range from the frozen solution at low temperatures, with a powderlike spectrum, to rapid motional averaging at room temperature where anisotropies are averaged out.

The intermediate region can provide information about slow molecular motions, which is especially important for nitroxide spin labels selectively attached to different parts of macromolecules such as the components of natural and synthetic phospholipid membranes, liquid crystals, and proteins. Such measurements have revealed important structural and functional information. Since the nitrogen hyperfine interaction in nitroxide free radicals has characteristic anisotropy, molecular reorientation times in the range 10^{-9} to 10^{-5} second can be determined. A typical free-radical spectrum in solution shows hyperfine splittings that split into two lines (**Fig. 2**). These splittings are due to two inequivalent nitrogen-14 (^{14}N) nuclei, a single phosphorus-31 (^{31}P) nucleus, and a weak coupling to a single hydrogen (^{1}H) nucleus (or proton). Other commonly occurring nuclear hyperfine splittings in free-radical EPR spectra may be due to deuterium (^{2}H), boron-11 (^{11}B), carbon-13 (^{13}C), or nitrogen-15 (^{15}N). When there are many equivalent protons, the splitting pattern can be calculated from binomial coefficients.

Electron transfer between molecules. Studies of chemical reaction processes involving electron transfer between molecules take advantage of stable free radicals with resolved hyperfine structure, where each line is associated with a particular precession frequency of the unpaired electron spin coupled to a particular configuration of nuclear spins. If an electron jumps to a different molecule nearby with a different arrangement of nuclear spins, its electron magnetic moment will then precess at a slightly different frequency. If the jump rate is small compared with the difference in precession frequencies, the observed linewidths will increase by the jump rate (in hertz). When the jump rates are fast compared with the difference between the precession frequencies, spectra characteristic of each molecule are observed. This latter situation also applies to inorganic transition-metal ion dimer complexes in crystals and frozen solutions where averaging is not observed and the anisotropic spectra provide useful structural clues. *See* ELECTRON-TRANSFER REACTION.

Electron nuclear double resonance (ENDOR). This multiple-resonance method involves simultaneous irradiation at both microwave frequencies and radio frequencies, the latter usually being swept over a range covering the nuclear frequencies. It can be understood as the detection of nuclear magnetic resonance (NMR) through a change in the EPR signal as the nuclear frequencies are swept through one or more resonances. ENDOR patterns for equivalent nuclei are much simpler than in a normal EPR spectrum. For example, for ^{1}H, the ENDOR pattern consists of two lines for any number of equivalent nuclei, whereas the normal EPR spectrum would

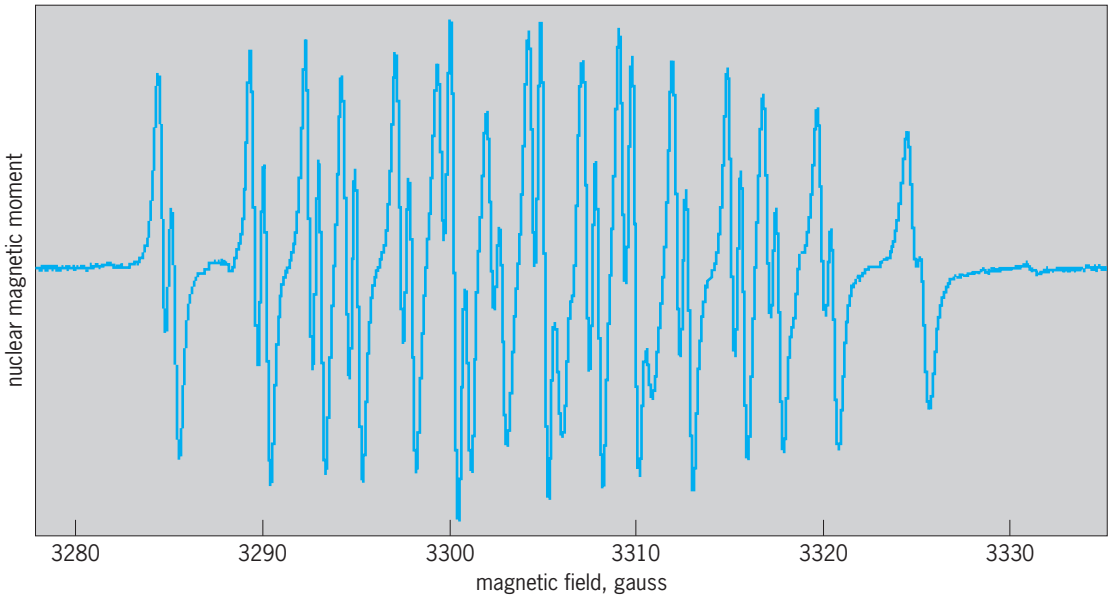


Fig. 2. Room-temperature solution spectrum of the betaine free-radical cation in toluene at a microwave frequency of 9.28877 GHz. The spectrum would be more symmetric at higher microwave frequencies. (*Courtesy of Dr. G. R. Hanson*)

yield multiline spectra with intenstites determined by the binomial coefficients. Linewidths are typically as low as 10 kHz compared with typical values in EPR of approximately 1 MHz for free radicals and 10 MHz for transition-metal ions. ENDOR has been successfully applied to determine properties of hydrogen atoms trapped in alkali halide and calcium fluoride crystals; to locate paramagnetic defects in diamonds, semiconductors, ferroelectrics; and to study organic free radicals.

Electron-electron double resonance (ELDOR). This method involves irradiation at two different microwave frequencies. The changes in the spectrum resulting from either sweeping the second frequency or the magnetic field are monitored through detection at the first frequency. This has been important for studies of slow motions using nitroxide spin labels.

Optical detection of magnetic resonance (ODMR). This method takes advantage of the much higher sensitivity of electric dipole transitions. In ODMR, paramagnetic states are optically excited, and the EPR signal is detected through changes in the optical absorption as the magnetic field is swept through one or more resonances. It has been particularly important for studying spin-triplet states that have nonmagnetic ground states and defects in solids, particularly in semiconductors. *See* NUCLEAR MAGNETIC RESONANCE (NMR).

Pulsed methods. It is often of interest to study the EPR spectra of many chemical or biological systems as a function of time after the microwave power is suddenly switched off. In practice, this is usually accomplished with controlled short high-power pulses of microwave radiation. Pulses as short as 2 billionths of a second have been used. The free induction decay (FID) contains all of the spectral information, and this is mathematically transformed into the normal spectrum using the Fourier transform. The microwave circuit in EPR spectrometers places a limit on measurements, typically until at least 70 billionths of a second following the microwave pulse. While the free induction decay can be observed for many free radicals, it is normally difficult to observe for transition-metal and rare-earth ions. Use of the spin echo is resorted to in these cases.

Electron spin echoes occur whenever two or more short, intense pulses of microwave radiation are applied to a sample with inhomogeneously broadened lines due to many overlapping narrow lines resulting from resonances corresponding to different arrangements of nuclear spin splittings. In the simplest (two-pulse) echo, during the interval τ between the two pulses, the magnetic moments become dephased and lose their coherence. The second pulse is arranged so as to refocus the magnetic moments, thus producing an echo at a later time τ. This means that, provided relaxation times permit, the experimenter has considerable control as to when echoes will occur. If τ is increased in small steps in a two-pulse echo, or suitable time intervals are incremented for multiple-pulse echoes, the echo intensity as a function of time often shows one or more periodic modulations known as electron spin echo envelope modulation (ESEEM). These modulated echo patterns may be mathematically transformed to yield the spectrum of nuclear frequencies. Two-dimensional (2D) forms of spectroscopy are also possible by increasing two different pulse intervals. The resulting 2D plots enable precise correlation of spectra to structural and electronic properties of the active paramagnetic species.

ESEEM and ENDOR provide complementary information. In samples where the unpaired electron in a molecule, or from an ion in a crystal, has high

symmetry, ENDOR spectra are strongest near or along principal, or symmetry, directions; whereas ESEEM has its greatest intensities away from principal directions.

ESEEM has been used effectively to probe weak hyperfine interactions in solids, in assisting with the determination of molecular structure, and in illuminating the local environment of paramagnetic ions at active sites in proteins. J. R. Pilbrow

Bibliography. L. Kevan and M. K. Bowman (eds.), *Pulsed and Continuous Wave ESR*, 1990; J. R. Pilbrow, *Transition Ion Electron Paramagnetic Resonance*, 1990; C. P. Poole and H. A. Farach, *Handbook of Electron Spin Resonance*, 1994; J. A. Weil, J. R. Bolton, and J. E. Wertz, *Electron Paramagnetic Resonance: Elementary Theory and Practical Applications*, 1993.

Electron-positron pair production

A process in which an electron and a positron are simultaneously created in the vicinity of a nucleus or subatomic particle. Electron-positron pair production is an example of the materialization of energy predicted by special relativity and is accurately described by quantum electrodynamics. Pair production usually refers to external pair production, in which the positron (positively charged antielectron) and electron are created from a high-energy gamma ray as it passes through matter. Electron-positron pairs are also produced from internal pair conversions in nuclei, decays of unstable subatomic particles, and collisions between charged particles. *See* QUANTUM ELECTRODYNAMICS.

External pair production. In external conversion, the energy of an incoming gamma ray (a high-energy electromagnetic photon) is directly converted into the mass of the electron-positron pair (**Fig. 1**). The photon energy $h\nu$ (where h is Planck's constant and ν is the photon frequency) must therefore exceed twice the rest mass of the electron $2m_0c^2$, equal to 1.022 MeV (m_0 is the electron mass, c the velocity of light). In order to conserve both energy and momentum in this process, the pair must be created near a nucleus, which recoils to balance the momentum of the incoming photon with the momenta of the created electron and positron. Because the nucleus is so much heavier than the electron, it carries away almost no energy from the pair, and the energy of the photon in excess of $2m_0c^2$ is shared unequally as kinetic energy by the positron and electron. Individually, the electron and positron each exhibit a distribution of kinetic energies ranging from zero to the maximum available energy, $E_{max} = h\nu - 2m_0c^2$, correlated with one another so that their sum is equal to E_{max}. Similarly, the positron and electron are emitted over a broad range of angles, although they exhibit a tendency to move in the same direction, which reflects the momentum of the incoming photon. For incident photon energies above 5 MeV, external pair production is the dominant mechanism by which gamma rays are absorbed in matter. *See* GAMMA-RAY DETECTORS; GAMMA RAYS; PHOTON.

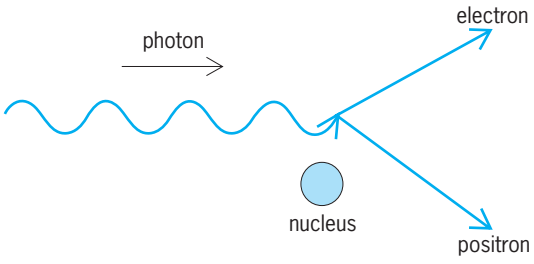


Fig. 1. External electron-positron pair production.

When a highly energetic charged particle collides with matter or with another particle, positron-electron pairs may be copiously produced by the external conversion of bremsstrahlung photons emitted as the incident particle is decelerated. This process is repeated as the created particles in turn produce pairs, until the average available energy is below the pair production threshold, giving rise to an electromagnetic cascade or shower. Showers are commonly observed in cosmic-ray interactions and are often used to detect high-energy electrons and photons produced by particle accelerators. *See* COSMIC RAYS.

Internal pair creation. Internal pair creation differs from external conversion in that the positron and electron are created directly from energy liberated by the deexcitation of an excited nucleus (produced, for example, in radioactive decay or nuclear collisions) to a state of lower energy, if the transition energy exceeds the pair mass threshold of $2m_0c^2$. Internal pair creation usually occurs only 10^{-3} times as often as deexcitation by gamma-ray emission, although the exact pair creation probability, as well as the angular correlation between the emitted positron and electron, depends on the nuclear charge and upon the energy and multipolarity of the nuclear transition. *See* RADIOACTIVITY.

Particle decay. Many unstable subatomic particles, such as the neutral Z boson and J/psi meson, decay into a positron-electron pair alone or with other particles. Since the decaying parent particle is massive, momentum is conserved without the presence of an additional nucleus as is required in external conversion. Decay into a single pair alone creates a positron and electron with equal and opposite momenta. They share the available kinetic energy equally, that is, the equation below is satisfied, where

$$E_{e^+} = E_{e^-} = \frac{Mc^2 - 2m_0c^2}{2}$$

M is the mass of the parent, and they are emitted back to back in the rest frame of the parent particle. *See* ELEMENTARY PARTICLE; MESON.

Superheavy nuclear collisions. In the 1980s, a new mode of positron-electron pair production was discovered in collisions of very heavy nuclei at high energies (**Fig. 2**). Narrow positron and electron peaks

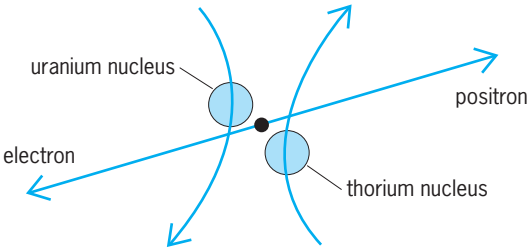


Fig. 2. Anomalous positron-electron pair creation in superheavy quasiatomic collisions.

were detected in coincidence from 1.4-GeV uranium-on-thorium quasiatomic collisions; three sets of correlated structures were found. In each case, the corresponding peak in the distribution of the sum of the positron and electron energies is narrower than the individual Doppler-broadened lines. This suggests the correlated cancellation of the positron and electron laboratory Doppler shifts, which is characteristic of an 180° back-to-back emission-angle correlation in the rest frame of their source. Explanations for these structures involving instrumental backgrounds, uncorrelated emission of separate positron and electron lines, external or internal pair production, and multibody final states are unable to describe the observed narrow width of the sum-energy peak. The data appear to reflect the production and decay of at least three previously undetected neutral objects with masses of 1.63, 1.77, and 1.84 MeV/c^2. Attempts to explain these data in terms of one or several new elementary particles have been unsuccessful so far. *See* POSITRON; QUASIATOM.

Thomas E. Cowan

Bibliography. M. L. Burns and A. K. Harding (eds.), *Positron-Electron Pairs in Astrophysics*, 1983; R. D. Evans, *The Atomic Nucleus*, 1955, reprint 1982; W. Greiner (ed.), *Physics of Strong Fields*, 1987; K. Siegbahn (ed.), *α-, β-, and γ-Ray Spectroscopy*, 1968; J. Thompson, *Soft Lepton Pair and Photon Production*, 1991.

Electron-probe microanalysis

A method used for determining the elemental composition of materials, based on the x-rays emitted by different elements when bombarded with high-energy electrons. It is a micro method that can detect x-ray photons emitted by the atoms within a small volume excited by an electron beam focused to 10 nanometers diameter or less. In biology electron-probe microanalysis can be used to determine the composition of cell organelles without isolating them and therefore altering the distribution of diffusible elements Studies with electron-probe analysis have thrown light on the major problems of excitation-contraction coupling in muscle, visual transduction in retinal rods, and the functions of the endoplasmic reticulum and mitochondria in the regulation of cytoplasmic calcium.

High-energy electrons can ionize atoms, ejecting an inner-shell electron. To fill the resultant vacancy, an outer-shell electron falls into the ionized shell (**Fig. 1**); the atom remains in the higher-energy excited state. The emission of a characteristic x-ray by the ionized atom is one of the mechanisms for releasing its excess energy. Another mechanism is the emission of Auger electrons; the probability of ejecting an x-ray instead of an Auger electron is the fluorescence yield. In the case of electron-probe microanalysis the source of the exciting electrons is the electron gun, which, in modern electron microscopes equipped with field-emission guns, can produce a focused beam narrower than 1 nm. The x-rays emitted as the result of atomic ionization are called characteristic x-rays, because their energy is characteristic of the core shell of the ionized element and of the shell from which the electron relaxed into the vacancy. X-rays excited by ionization of the innermost (K) shells are called K x-rays, those due to L-shell ionization, L x-rays. Typical energies of characteristic K x-rays of the biologically important elements are given in the **table**. *See* AUGER EFFECT; SURFACE PHYSICS; X-RAY FLUORESCENCE ANALYSIS.

X-ray detection and data processing. At the great sensitivity required for biological applications at high spatial resolution, x-rays are detected with energy-dispersive detectors. They are solid-state devices in which incident x-rays create electron-hole pairs. In those most commonly used, lithium-drifted silicon (Si-Li) detectors, each 3.8 eV of energy discharged by an x-ray photon creates one electron-hole pair. Hence, following the arrival of a high-energy x-ray, the charge deposited and then swept off the detector is a measure of the energy of the incident x-ray. The charge after pulse and digital processing, can be stored in one of the channels of a multichannel analyzer. This procedure builds up a spectrum of the number of emitted x-rays as a function of their

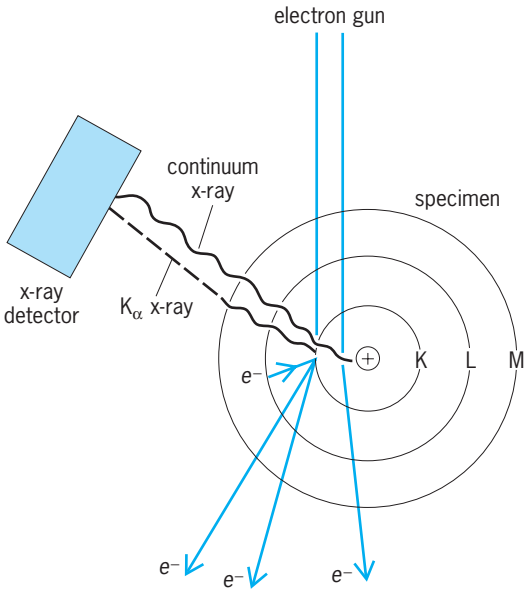


Fig. 1. Ionization of an atom in a specimen and production and detection of the characteristic and continuum x-rays generated by the incident electrons. Unscattered and elastically or inelastically scattered electrons are shown leaving the specimen plane.

Energies of K x-rays

Element	Energy, keV
Sodium	1.041
Magnesium	1.253
Phosphorus	2.013
Sulfur	2.306
Potassium	3.310
Calcium	3.691

energies (**Fig. 2**). The energy resolution of modern x-ray detectors, usually specified at the Kα manganese line, is approximately 150 eV. This represents significant broadening of the intrinsic (4 eV or less) width of the x-ray lines by electronic noise and other factors, and can cause overlap of adjacent x-ray peaks. The analysis of overlapping peaks, required for accurate determinations, is performed by computer.

Quantitative analysis. The concentrations of elements are determined from the relationship, given in the equation below, between the number of atoms

$$I_x = \omega_x \cdot q_x \cdot N_x \cdot T_x$$

(N_x) in an irradiated space and the number of characteristic x-rays (I_x) emitted. Here ω_x is the flouresscence yield, q_x the ionization cross section (the probability of an atom being ionized), N_x the number of atoms of element X in the space irradiated, I the probe current, and T_x the detector transmission efficiency. Thus, the number of x-rays generated in a given space is directly proportional to the number of atoms and the number of exciting electrons. The two major influences on the performance of the detector are the geometric efficiency with which the emitted x-rays reach it and the extent of x-ray absorption by its window. The geometry of electron microscope lenses is such that less than 1% of emitted x-rays reaches an energy-dispersive detector. Low-energy x-rays are absorbed by the detector window: elements below sodium in atomic number are not detected, and the detection efficiency for sodium and magnesium is less than that for, say, potassium. Elements of low atomic number can be analyzed with windowless x-ray detectors or electron energy-loss spectroscopy. *See* ELECTRON MICROSCOPE; ELECTRON SPECTROSCOPY.

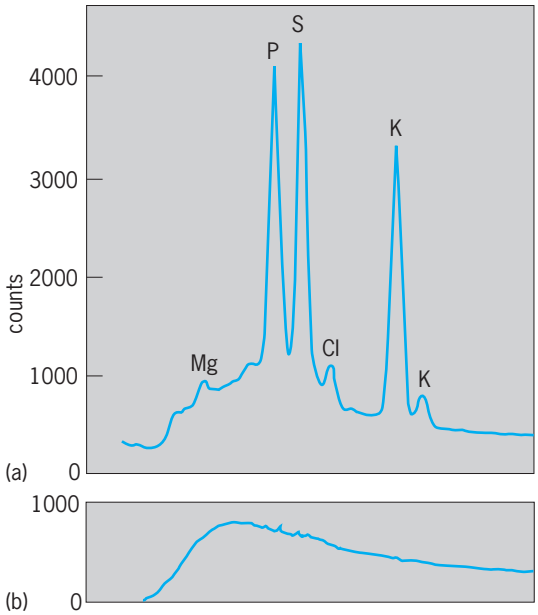


Fig. 2. Typical cellular x-ray spectrum, indicating high concentration of potassium: (*a*) a series of characteristic elemental peaks superimposed on (*b*) a background of x-ray continuum, shown with the characteristic peaks removed by computer processing.

X-ray photons are also absorbed by the material being analyzed, complicating the analysis of thick and bulky specimens. Spatial resolution is also reduced in thick specimens through broadening of the probe by electron scattering. Biological studies at the high spatial resolution required for analyzing cell organelles are performed on sections less than 200 nm thick.

The concentration of elements is also best determined in ultrathin specimens, because excitation of an atom with electrons also creates continuum x-rays, which result from inelastic scattering by atomic nuclei. Continuum x-rays are distributed over a broad energy band (Fig. 2*a*); their number is proportional to the total mass within the irradiated space. The same energy-dispersive x-ray detectors simultaneously detect characteristic and continuum x-rays, and the ratio of the two is a measure of elemental concentration, usually expressed as mmol-kg, dry weight. With modern instrumentation, concentrations of calcium as low as 0.3 mmol/kg can be detected within a 100-nm^3 volume or less.

X-ray mapping. X-ray mapping, a second mode of electron-probe microanalysis application, yields a two-dimensional map of elemental distribution. X-ray maps are obtained by scanning an area of a specimen with the electron beam and collecting the x-ray spectrum at each picture point (pixel) as the beam is rastered. In this manner, the concentration of different elements at each picture point can be determined, and the resultant values displayed as relative intensities on x-ray maps of, for example, phosphorus or calcium. Powerful electron sources are required with the small probes that give high spatial resolution, to obtain low-noise x-ray maps of biological specimens. Field-emission guns are the strongest available and provide the highest current density in a focused beam. The distribution of several elements in specific granules of the atrium of a rat heart, mapped with a field-emission gun, is shown in **Fig. 3**. *See* X-RAY MICROSCOPE.

Specimen preparation by rapid freezing. Extreme care must be taken to prevent the loss or translocation of diffusible elements during specimen preparation. Freezing at rates of almost 10,000°C/s (18,000°F/s) is used almost universally, followed by vacuum dehydration. Valuable information, albeit at lower resolution, has been obtained by analysis of frozen hydrated (ice) specimens of secretory epithelia.

Biological applications. Important biological structures have been analyzed with electron-probe

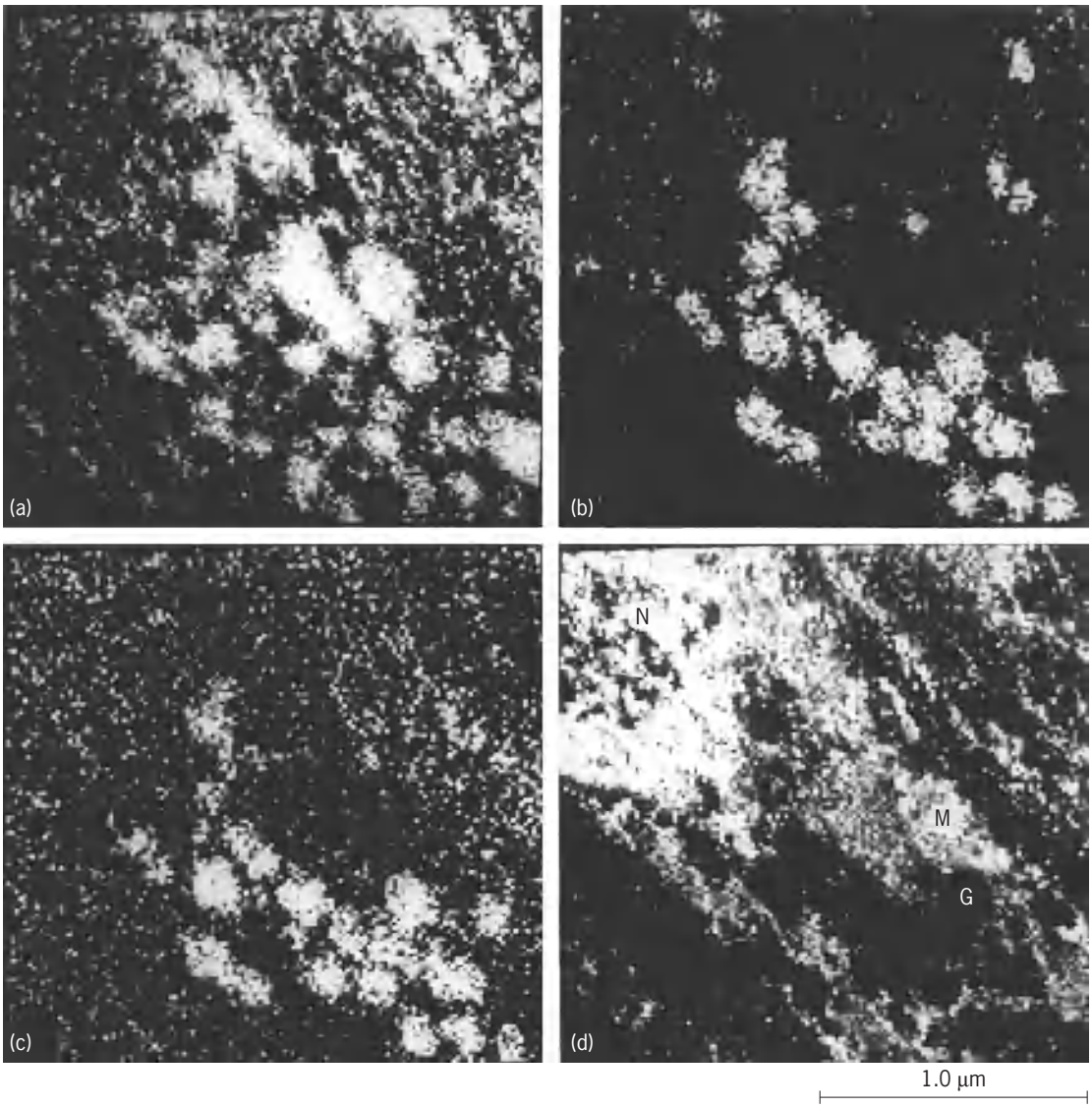


Fig. 3. X-ray maps of a rat heart cryosection, showing the distribution of (*a*) sulfur, (*b*) calcium, (*c*) chlorine, and (*d*) phosphorus in a 2-μm^2 region of the nuclear pole, containing mitochondria (M), atrial granules (G), and condensed chromatin inside the nuclear (N) envelope. The intensity of gray indicates distribution; black represents the fewest and white the most numerous atoms of each element.

microanalysis, including sarcoplasmic and endoplasmic reticulum, mitochondria, and nuclei. Largely as a result, mitochondria are now known to accumulate significant amounts of calcium only under pathological conditions when cytoplasmic calcium rises to abnormally high levels; in normal cells the endoplasmic (in muscle the sarcoplasmic) reticulum, in conjunction with plasma membrane calcium pumps, is the principal regulator of cytoplasmic calcium. Electron-probe microanalysis has revealed subcellular movements of magnesium in and out of mitochondria and sarcoplasmic reticulum that were not anticipated from studies of isolated organelles. Electron-probe microanalysis of bacteria and bacterial spores has revealed the distribution of magnesium, calcium, and other elements, as well as an increase in cell calcium during bacterial division. The application of this technique to abnormal red blood cells in sickle cell anemia has shown that their high calcium content is concentrated in small vacuoles. *See* MASS SPECTROMETRY; SECONDARY ION MASS SPECTROMETRY (SIMS); SPECTROSCOPY.

Andrew Paul Somlyo

Bibliography. A. Le Furgey, M. Bond, and P. Ingram, Frontiers in electron probe microanalysis: Application to cell physiology, *Ultramicroscopy*, 24:185–219, 1988; A. P. Somlyo, Cell calcium measurement with electron probe and electron energy loss analysis, *Cell Calcium*, 6:197–212, 1985; A. P. Somlyo, Compositional mapping in biology: X-rays and electrons, *J. Ultrastruc. Res.*, 83:135–142, 1985.

Electron spectroscopy

A form of spectroscopy which deals with the emission and recording of the electrons which constitute matter—solids, liquids, or gases. The usual form of spectroscopy concerns the emission or absorption

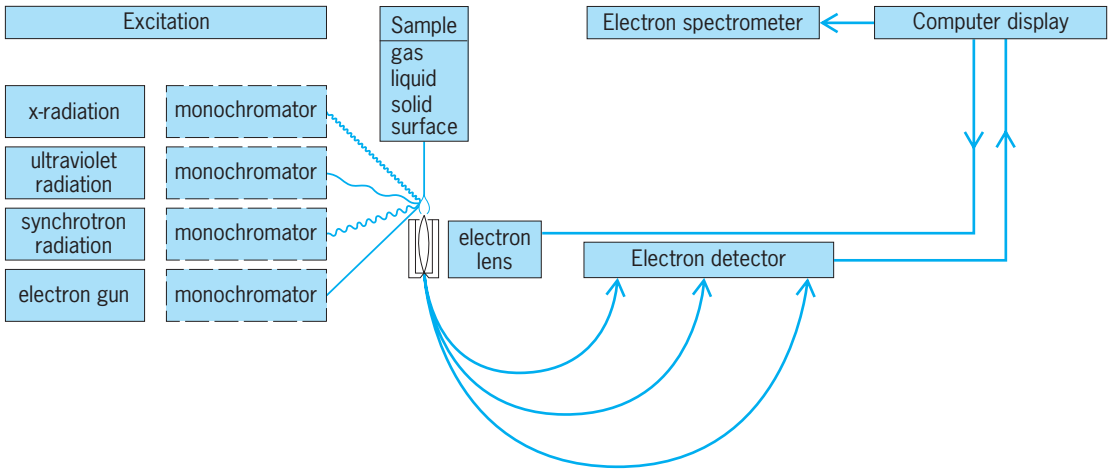


Fig. 1. Excitation of electron spectra recorded with high-resolution instruments. (*After H. Siegbahn and L. Karlsson, Photoelectron spectroscopy, Handbuch der Physik, vol. 31, 1982*)

of photons [x-rays, ultraviolet (uv) rays, visible or microwave wavelengths, and so on]. Electron spectra can be excited by x-rays, which is the basis for electron spectroscopy for chemical analysis (ESCA), or by uv photons, or by ions (electrons; **Fig. 1**). By means of x-ray or uv photons with energy E_{bv}, photoelectron spectra (PES) are produced when electrons with binding energies E_b are emitted with energy E_{kinetic} from bound molecular states, according to the equation below. E_r is a usually negligible small recoil

$$E_{\text{kinetic}} = E_{bv} - E_b - E_r - \varphi$$

energy, and φ a small work function correction which can be attributed to contact potentials. For insulating solid materials, precautions are taken for

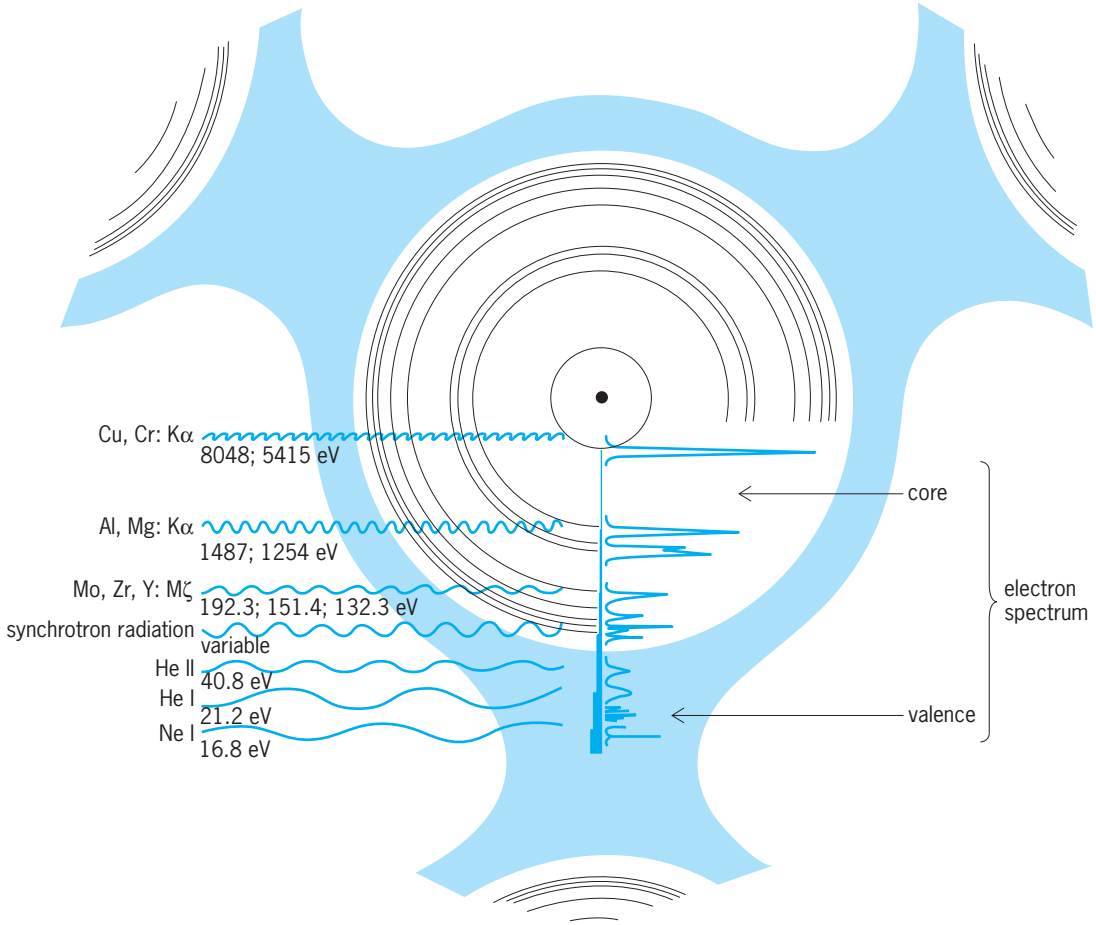


Fig. 2. The system of levels of an atom in a molecule can be divided into a valence electron region and an atomic-core region. Excitation of electron lines from the various regions can be made at different photon energies.

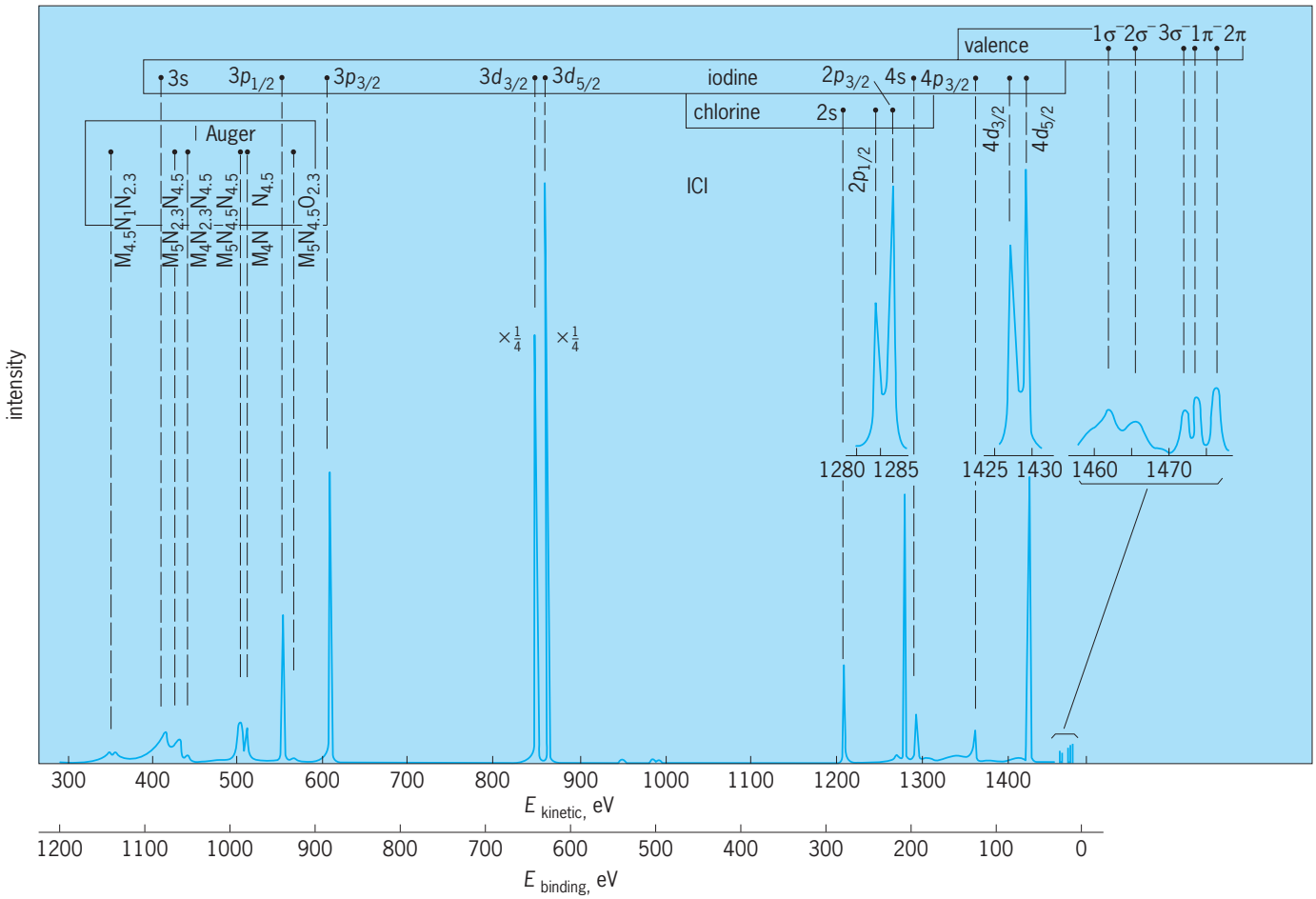


Fig. 3. Electron spectrum of the vapor of iodine chloride at a pressure of 0.04 torr (5 Pa), excited monochromatic Al Kα radiation with the ESCA instrument (scan spectrum). The M- and N-shell levels are excited in iodine, and the L levels in chlorine. At low binding energies (high kinetic energies), the valence spectrum of the molecule is recorded. The MNN Auger electron lines of iodine can also be seen.

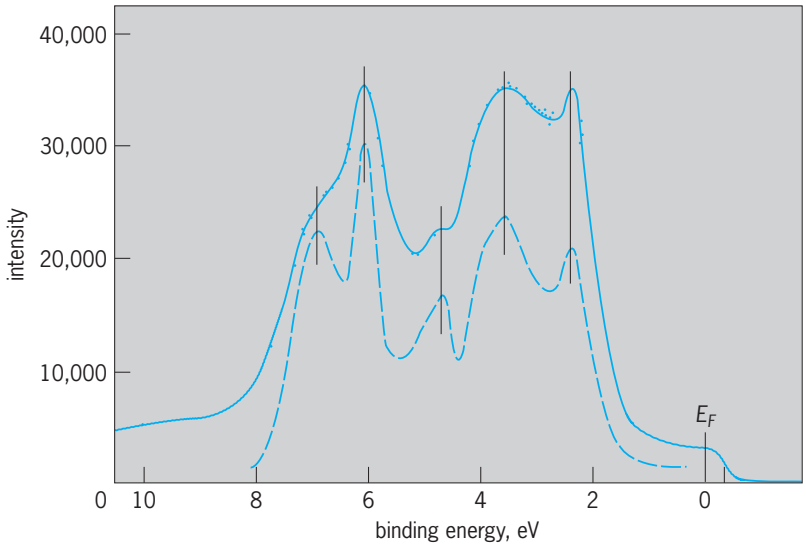


Fig. 4. Valence-band electron spectrum of gold excited by monochromatized Al Kα radiation (solid curve). Comparison is made with a theoretically calculated density of states (broken curve) which has been folded with a 0.25-eV gaussian corresponding to the experimental resolution. The energy scale of the theoretical density-of-states curve has also been expanded by 7.5%, which leads to a near-perfect correspondence between theoretical and experimental band profiles. E_F is the Fermi edge. $h\nu = 1486.7$ eV. Foil, $\theta = 45^\circ$. (*After U. Gelius et al., Experimental spectrum, Nucl. Instr. Meth., B1:85 1984; and N. E. Christensen and B. O. Seraphin, Theoretical density of states, Phys. Rev., B4:3321, 1971*)

stabilizing the surface potential. For solids, the binding energies are preferably referred to the Fermi level, and for gases, to the vacuum level.

Modes of excitation. By means of ESCA, complete sets of photoelectron lines can be excited from the internal (core) levels as well as from the external (valence) region (**Fig. 2**). Also, complete sequences of the Auger electron lines are automatically obtained in this mode. A convenient source of excitation is the x-radiation from Al at 1487 eV. With the use of spherically bent quartz crystals this radiation can be further monochromatized to an inherent width of 0.2 eV. The commonly used light source for uv excitation of photoelectron spectra within the valence electron region is a helium lamp that produces highly monochromatic radiation at 21.2 eV (internal width less than 10 mev) and (with less intensity) at 40.8 eV. Intermediate in energy for excitation are ultrasoft x-rays (Y Mζ at 132 eV). Another source of excitation in the intermediate region is provided by synchrotron radiation, which can be continually varied by means of a suitable monochromater. Excitation by means of an electron beam is an alternative mode to obtain Auger electron lines that has the advantage of ease of production. Ordinary photoelectron

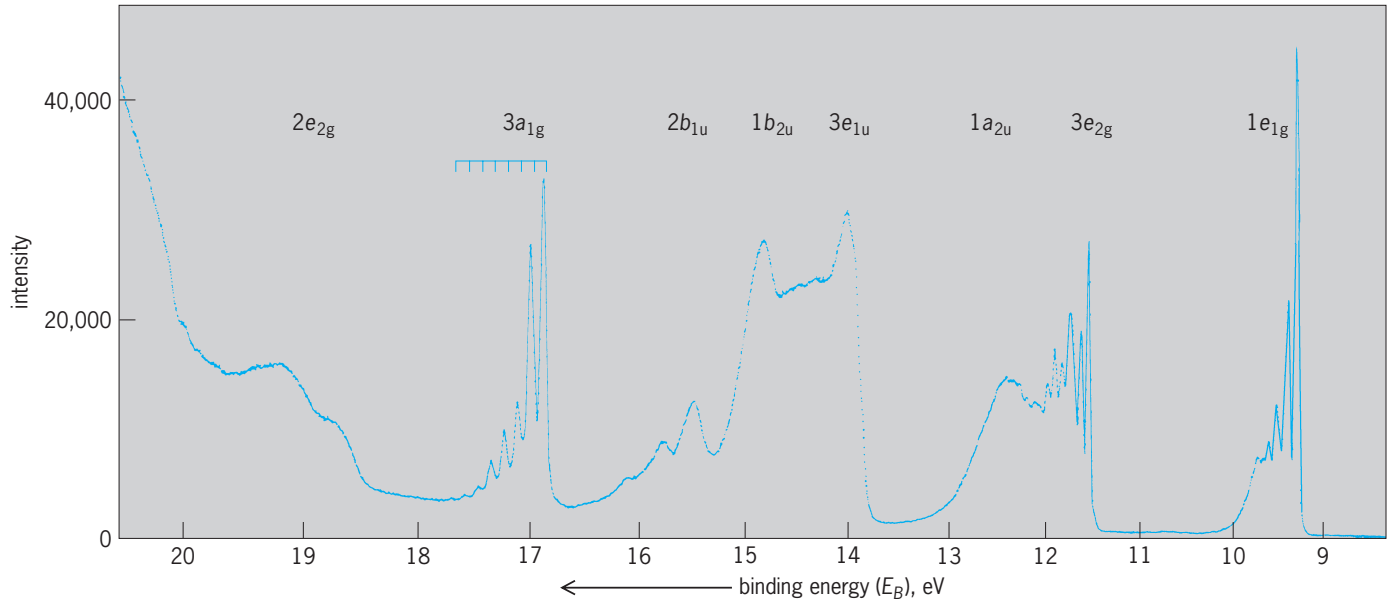


Fig. 5. Benzene valence electron spectrum excited by HeI radiation at 21.22 eV, showing vibrational structures.

lines are not produced by that radiation. In order to compensate for the low-signal-to-background ratio when Auger electron spectra are excited by an electron beam impinging on a solid surface, the spectral distribution is frequently differentiated. *See* AUGER EFFECT.

Applications. In ESCA, only electrons which are expelled from a surface layer of less than 50 nanometers of a solid material contribute to the electron line with the kinetic energy given above. Electrons from the interior of the material are scattered out from the line, and form a low background which does not interfere with the line character of the ESCA spectrum. The electron lines are extremely sharp and well suited for precision measurements. With a high- resolving ESCA spectrometer which has a magnetic or electrostatic focusing dispersive system, the electron lines have widths which are set by the limit caused by the uncertainty principle (the "inherent" widths of atomic levels). With a suitable choice of radiation, electron spectroscopy reproduces directly the electronic level structure from the innermost shells (core electrons) to the atomic surface (valence or conduction band; **Fig. 3**). Furthermore, all elements from hydrogen to the heaviest ones can be studied even if the element occurs together with several other

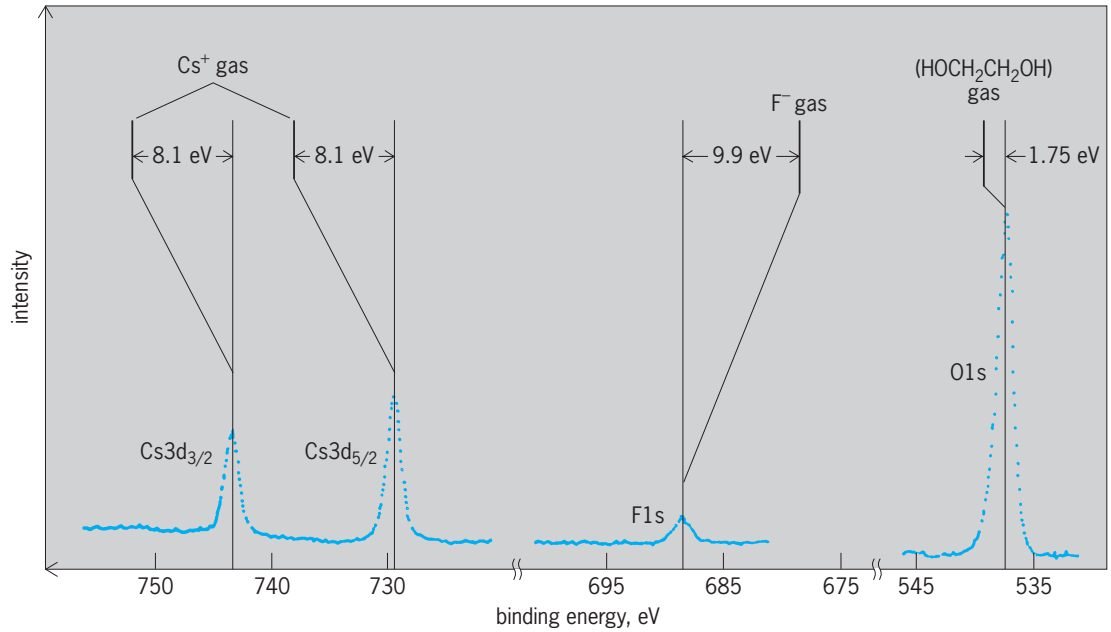


Fig. 6. ESCA spectrum of cesium fluoride (CsF) in glycol ($HOCH_2CH_2OH$) solution. Comparison is made with the corresponding electron-binding energies in the gas phase. (*After H. Siegbahn et al., Phys. Scripta, 27:431, 1983*)

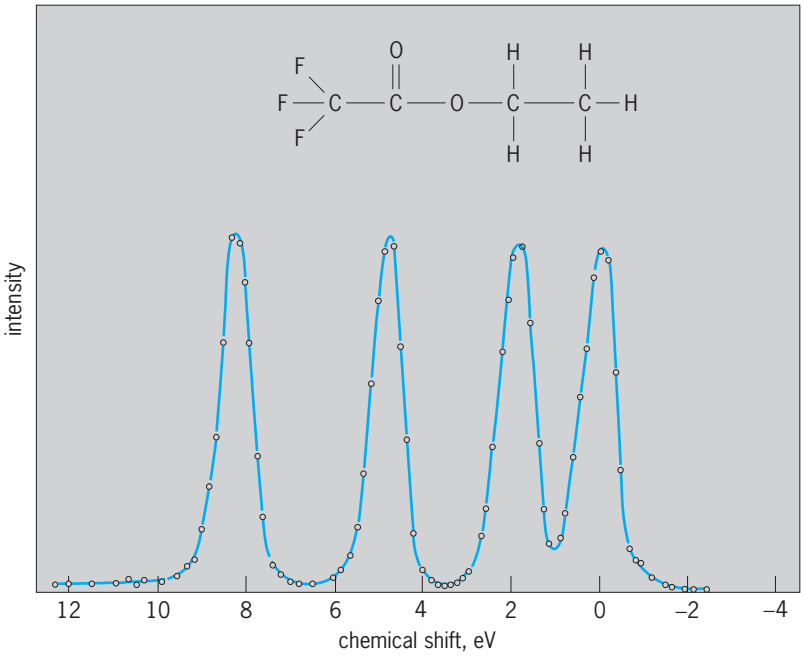


Fig. 7. Electron spectrum from the 1*s* level of carbon in ethyl trifluoroacetate; binding energy = 291.2 eV.

elements and even if the element represents only a small part of the chemical compound. *See* LINE SPECTRUM; UNCERTAINTY PRINCIPLE.

When applied to solid materials, ESCA is a typical surface spectroscopy with applications to problems such as chemical surface reactions, for example, corrosion or heterogeneous catalysis. In such cases, ultrahigh vacuum conditions are usually required, with a vacuum less than 10^{-9} torr (10^{-7} pascal). ESCA also reproduces bulk matter properties such as valence electron band structures (**Fig. 4**). With uv excitation, the resolution is sufficient to resolve vibrational structures in valence electron spectra for gases (**Fig. 5**). In fact, electron spectroscopy can supply a detailed knowledge of the valence orbital structure for all molecules which can be brought into gaseous form with pressures of 10^{-5} torr (10^{-3} Pa) or more. A convenient pressure region is 10^{-2} torr (10 Pa). Liquids and solutions of various compositions can also be studied by ESCA techniques (**Fig. 6**).

ESCA chemical shifts. When atoms are brought close together to form a molecule, the electronic orbitals of each atom are perturbed. Inner orbitals, that is, those with higher binding energies, may still be regarded as atomic and belonging to specified atoms within the molecule, whereas the outer orbitals combine to form the valence-level system of the molecule. These orbitals play a more or less active part in the chemical properties. The chemical bonds affect the charge distribution so that the original atoms can be regarded as charged to various degrees; a neutral molecule has a net charge of zero. The individual atoms in the molecule can be regarded as spheres with different charges set up by the transfer of certain small charges from one atomic sphere to the neighboring atoms taking part in the chemical bond. Inside each charged sphere the atomic potential is constant, in accordance with classical electrostatic theory. The result of this atomic potential is to shift the whole system of inner levels in an atom by a small amount, the same amount for each level. Levels belonging to different atoms in the molecule are shifted differently, however, and if the ESCA chemical shifts for individual atoms in the molecule are measured, a mapping can be made of the distribution of charge or potential in the molecule. This is a reflection of the chemical bondings between the atoms, which in turn can be described by the orbitals in the valence-level system.

A unique feature of ESCA is that, if the exact position of the electron lines characteristic of the various elements in the molecule is measured, the area of inspection can be moved from one atomic species to another in the molecular structure. **Figure 7** shows the electron spectrum from the 1*s* level of the carbon in ethyl trifluoroacetate. All four carbon atoms in this molecule are distinguished in the spectrum. The lines appear in the same order from left to right, as do the corresponding carbon atoms in the structure that is shown in the figure. If the structure of the molecule is known, the charge distribution can be estimated in a simple way by using, for example, the electronegativity concept and assuming certain resonance structures. More sophisticated quantum-chemical treatments can also be applied. Conversely, if, by means of ESCA, the approximate charge distribution is known, conclusions about the structure of the molecule can be drawn.

Experimental evidence obtained so far for various elements in a large number of molecules indicates strong correlations between chemical shifts and calculated atomic charges. A typical correlation curve for the 2*p* level in sulfur obtained from more than 100 compounds containing this element is shown in **Fig. 8**. Similar curves are obtained for other elements such as carbon, nitrogen, oxygen, and phosphorus. Chemical shifts are also observed in the electron lines

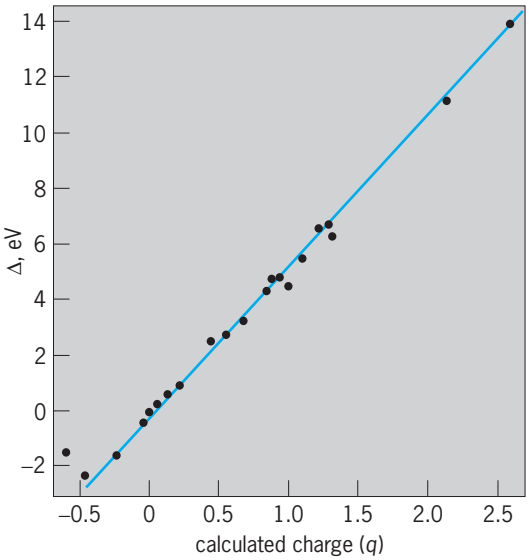


Fig. 8. Binding-energy shifts for the sulfur 2*p* electrons versus the calculated charge. The points indicate averages from more than 100 compounds.

due to the Auger effect. Second-order chemical shifts from groups situated farther away in the molecule (inductive effects) are also observed.

The theoretical importance of the chemical shift effect lies in the study of molecular electronic structure. In other contexts, its usefulness lies precisely in its ability to specify chemical composition, in particular, at surfaces. For example, a metal and its oxide give two distinctly different lines because of the chemical shift effect. It is often easy to follow the rate of oxidation at a surface, or the adsorption of a gas layer onto it. The chemical composition of a surface or the changes due to various chemical or physical treatments are amenable to detailed examination. *See* ATOMIC STRUCTURE AND SPECTRA; CHEMICAL BONDING; ELECTRON CONFIGURATION; ELECTRONEGATIVITY; MOLECULAR ORBITAL THEORY; SPECTROSCOPY.

Kai Siegbahn; Hans Siegbahn

Bibliography. T. L. Barr, *Electron Spectroscopy–Chemical Analysis (ESCA)*, 1994; C. R. Brundle and A. D. Baker (eds.), *Electron Spectroscopy, Theory, Techniques and Applications,* vols. 1–5, 1977–1984; M. Cardona and L. Ley (eds.), *Photoemission in Solids*, vols. 1 and 2, 1978, 1979; K. Siegbahn et al., *ESCA Applied to Free Molecules*, 1969; K. Siegbahn et al., *ESCA: Atomic, Molecular and Solid State Structure Studies by Means of Electron Spectroscopy*, 1967; G. C. Smith, *Surface Analysis by Electron Spectroscopy*, 1994; E. L. Wolf, *Principles of Electron Tunneling Spectroscopy*, 1985, reprint 1989.

Electron spin

That property of an electron which gives rise to its angular momentum about an axis within the electron. Spin is one of the permanent and basic properties of the electron. Both the spin and the associated magnetic dipole moment of the electron were postulated by G. E. Uhlenbeck and S. Goudsmit in 1925 as necessary to allow the interpretation of many observed effects, among them the so-called anomalous Zeeman effect, the existence of doublets (pairs of closely spaced lines) in the spectra of the alkali atoms, and certain features of x-ray spectra. *See* SPIN (QUANTUM MECHANICS).

All theory that concerns itself with electronic, nuclear, atomic, and molecular phenomena includes the electron spin in its formulation to obtain a theoretical structure consistent with experimental observation. The electron thus possesses the intrinsic property of spin angular momentum (rotational motion about an axis), in addition to the intrinsic properties of charge and mass. *See* ELECTRON.

The spin quantum number is s, which is always $^1/_2$. This means that the component of spin angular momentum along a preferred direction, such as the direction of a magnetic field, is $\pm^1/_2\hbar$, where $\hbar = h/2\pi$ and h is Planck's constant. The spin angular momentum of the electron is not to be confused with the orbital angular momentum of the electron associated with its motion about the nucleus. In the latter case the maximum component of angular momentum along a preferred direction is $l\hbar$, where l is the angular momentum quantum number and may be any positive integer or zero. The total orbital angular momentum is equal to $\sqrt{l(l+1)}\hbar$. In this discussion the terms angular momentum or magnetic dipole moment describe the maximum component of these quantities along a field direction. *See* ANGULAR MOMENTUM; QUANTUM NUMBERS.

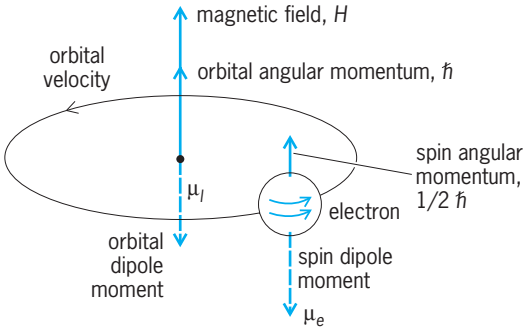


Fig. 1. Diagram of an electron in circular motion in a magnetic field, showing spin and orbital angular momenta and resulting magnetic moments.

Electron magnetic moment. The electron has a magnetic dipole moment by virtue of its spin. The approximate value of the dipole moment is the Bohr magneton μ_0 which is equal, in SI units, to $eh/4\pi m = 9.27 \times 10^{-24}$ joule/tesla, where e is the electron charge measured in coulombs, m is the mass of the electron, and c is the velocity of light. The orbital motion of the electron also gives rise to a magnetic dipole moment μ_l that is equal to μ_0 when $l = 1$. In **Fig. 1** is shown the simple case of an electron in circular motion in a magnetic field. The electron is shown with a spin angular momentum parallel to the orbital angular momentum. (Physically it could equally well be in the opposite direction.) Since the electron has a negative charge, the directions of the orbital dipole moment and spin dipole moment (both are vectors) are opposite to that of the orbital angular momentum. For a positron (a positively charged particle having the same mass and magnitude of charge as the negatively charged electron), the magnetic moments are positive, that is, in the same direction as the angular momentum. *See* MAGNETON; POSITRON.

The orbital magnetic moment of an electron can readily be deduced with the use of the classical statements of electromagnetic theory in quantum-mechanical theory; the simple classical analog of a current flowing in a loop of wire describes the magnetic effects of an electron moving in an orbit. The spin of an electron and the magnetic properties associated with it are, however, not possible to understand from a classical point of view. The classical radius of the electron is $e^2/(8\pi\epsilon_0 mc^2) = 1.141 \times 10^{-13}$ cm, where ϵ_0 is the permittivity of free space; a reasonable distribution of mass and electric charge within this radius which would lead to a magnetic moment μ_0 leads to the calculation of a peripheral velocity of the electron far greater than the velocity

of light. This is, of course, wholly precluded by the special theory of relativity. No theory of the structure of the electron has been formulated which makes the spin of the electron amenable to simple pictorial understanding. Nevertheless the interpretation of the spectra of atoms and molecules, the magnetic properties of materials, and other phenomena on an atomic scale unambiguously require that the electron have the property of spin. To the extent that physical theory assumes these properties and does not concern itself with questions about the structure of the electron, it is quite adequate to deal with a large range of physical phenomena in a highly quantitative way. *See* RELATIVITY.

In the Landé g factor, g is defined as the negative ratio of the magnetic moment, in units of μ_0, to the angular momentum, in units of $\hbar$. For the orbital motion of an electron, $g_l = 1$. For the spin of the electron the appropriate g value is $g_s \simeq 2$; that is, unit spin angular momentum produces twice the magnetic moment that unit orbital angular momentum produces.

The following discussion is limited to atoms which have a single electron outside of closed electron shells. Both the orbital and spin angular momenta of the electrons within closed shells add up in such a way that their net angular momentum is zero. The single electron outside closed shells may have $l = 0, 1, 2, \ldots$. By the usual rules developed from quantum mechanics the total angular momentum quantum number of the electron, which is called j, is $l \pm s$ (Fig. 1) except when $l = 0$, in which case the total angular momentum is s. For instance, when $l = 1$, $j = {}^1/_2$ or ${}^3/_2$. These relations may be represented by vector diagrams, as shown in **Fig. 2**. Since the revolution of the electron about the nucleus produces a magnetic field at the electron, and since the electron has a magnetic dipole moment, the energy of the atom is different when the vector s is parallel to $l (j = l + {}^1/_2)$ and when s is antiparallel (180°) to $l (j = l - {}^1/_2)$. Thus the spin of the electron causes a doubling of the energy levels in all single-electron atoms except when $l = 0$, in which case level is single. In the case of sodium, the familiar yellow lines (the D lines) comprise a closely spaced doublet that arises from a transition from a state of $l = 1$ to one of $l = 0$. The doubling of the lines is a direct consequence of the existence of the electron spin.

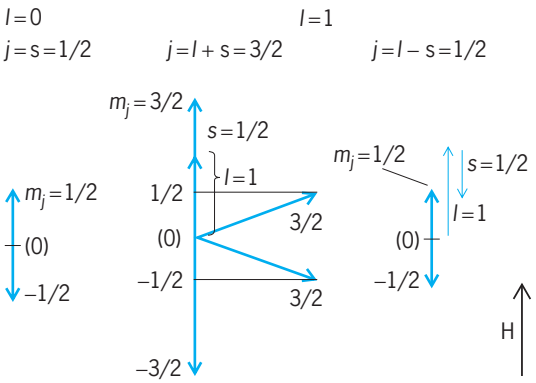


Fig. 2. Diagram illustrating addition of *l* and *s* vectors into a resultant *j* vector. The allowed values of m_j are also shown. In no case does the (0) give an allowed value. *H* = magnetic field.

Energy-level splitting. The total electronic magnetic moment of an atom depends on the state of coupling between the orbital and spin angular momenta of the electron (Fig. 2). In the single-electron case, an atom in the state for which $l = 0$ has only the magnetic moment associated with the spin, and this moment may be oriented in either of two directions with respect to an externally applied magnetic field. However, when the atom is in a state for which $l = 1$, j can be ${}^1/_2$ or ${}^3/_2$ and the Landé g factors for these two states are ${}^2/_3$ and ${}^4/_3$, respectively. That is, the magnetic moment per unit angular momentum is equal neither to that which characterizes the orbital motion nor to that which characterizes the spin motion. In a magnetic field, a single energy level characterized by l and j is split into several energy levels, each described by the component of j, m_j, along a magnetic field, where $m_j = j, j - 1, \ldots, -(j - 1), -j$. The energy of the level is the zero field energy plus the term $m_j g_j \mu_0 H$. In a transition between two energy levels that gives rise to a spectral line, the energy of the emitted or absorbed photon is the zero field energy difference between the two levels plus the difference between two magnetic energy terms as given previously. The resultant splitting of the line, which is single at zero magnetic field, into a line of two or more components is called the Zeeman effect. The Zeeman effect has been called anomalous when it is not explicable purely in terms of the orbital motion of the electron. The introduction of the electron spin allows the interpretation of all observed Zeeman effects. *See* ZEEMAN EFFECT.

Atomic beam measurements. Prior to 1940, spectroscopic measurements had been made only on optical spectra, and Zeeman effects were small effects superimposed on lines of considerable natural width. Exact measurements of the splitting of lines in a magnetic field, therefore, could not be made on such lines, and all data on the Zeeman effect were consistent with the statement that $g_s = 2g_l$. With the development of spectroscopy by the atomic beam method, a new order of precision in the measurement of the frequencies of spectral lines became possible. By using the atomic-beam techniques, it became possible to measure g_s/g_l directly, with the result $g_s/g_l = 2$ (1.001168 ± 0.000005). The magnetic moment of the electron therefore is not μ_0 but $1.001168\mu_0$, or equivalently the g factor of the electron departs from 2 by the so-called g factor anomaly defined as $a = (g_2 - 2)/2$ so that $\mu = (1 + a)_0$. Thus the first molecular beam work gave $a = 0.001168$. *See* MOLECULAR BEAMS.

Calculation of g-factor anomaly. It is not possible to give a qualitative description of the effects which give rise to the g-factor anomaly of the electron. The detailed theoretical calculation of the quantity is in the domain of quantum electrodynamics, and involves the interaction of the zero-point oscillation of the electromagnetic field with the electron.

Comparison of theoretical determination of a with its experimental measurement constitutes the most accurate and direct existing test of the theory of quantum electrodynamics.

Theoretical work on the g-factor anomaly, based on the principles of quantum electrodynamics, began simultaneously with its experimental discovery. The initial prediction of the value of the anomaly was made by J. Schwinger in 1948, who showed that the anomaly could be written as $a = 0.5(\alpha/\pi)$, where α, the fine-structure constant, is given by $\alpha = e^2/\hbar c \simeq 1/137$. Shortly thereafter, it was shown that a could be expressed as a power series in α (more customarily, α/π), that is, one can write $a = A(\alpha/\pi) + B(\alpha/\pi)^2 + C(\alpha/\pi)^3 + D(\alpha/\pi)^4 + \cdots$. Calculation of B, C, and D from quantum electrodynamics has proven to be very difficult, with errors occurring at various stages of the work. However, it is now generally agreed that $A = 0.500$, $B = -0.328478965\ldots$, $C = 1.181241456\ldots$, and $D = -1.7283 \pm 0.0035$. In order to find a theoretical value for a, an accurate value of α is required. There are several values of α, obtained from different sources, which may be used to determine a. The most accurate value of α is based upon results obtained by a measurement of the recoil velocity of rubidium atoms in an optical lattice and is given by $\alpha^{-1}{}_{\mathrm{Rb}} = 137.03599878 \pm 0.00000091$. The result of substituting this value of α into the expression for a gives the most recent theoretical value of a as a(theoretical) $= 0.0011596521887 \pm 0.0000000000077$. The uncertainty quoted for this number is due primarily to error in α. Finally, the weak and strong interactions have been calculated to add about 0.0000000000017 to a(theoretical). *See* ELECTRICAL UNITS AND STANDARDS; FUNDAMENTAL CONSTANTS; HALL EFFECT; QUANTUM ELECTRODYNAMICS; STRONG NUCLEAR INTERACTIONS; WEAK NUCLEAR INTERACTIONS.

Measurement of g-factor anomaly. The initial molecular-beam method for measuring a was based on measurement of g_s/g_l, which yields $1 + a$ rather than a itself. Such experiments have been superseded by two types of experiments which have the major advantages of measuring a directly, and of measuring a for electrons trapped in electric and magnetic fields but not bound to atoms. The second feature means that various corrections due to the presence of an atom are no longer necessary.

The new experiments may be readily understood, qualitatively, by noting that if an electron (spin angular momentum $\hbar/2$) moves at a low velocity $V(V/c \ll 1)$ perpendicular to a magnetic field B, the particle will rotate in the field at a frequency (the cyclotron frequency) given by $f_c = (1/2\pi)\ (eB/m)$, while its magnetic moment (and spin) will precess about the field at a frequency (the spin precession frequency) given by $f_s = (1 + a)f_c$. The difference between these frequencies, often called the $g - 2$ or difference frequency, is given by $f_D = f_s - f_c = af_c$. Thus f_D represents the rate at which the spin of the particle rotates relative to the particle's velocity.

One technique for measuring a consists of trapping very low-energy electrons in a magnetic field and measuring f_D and f_c simultaneously, with a being given by the relation $a = f_D/f_c$.

A slightly different version of the same technique consists of trapping electrons whose velocity is about half the speed of light (electrons accelerated to several hundred thousand volts) in a magnetic field and determining the same ratio. Relativistic effects manifest themselves in both the expressions for f_s and f_c, but cancel exactly when the ratio f_D/f_c is formed. The high-energy technique constitutes an excellent check of the separate measurements, as well as one of the most precise tests extant of the special theory of relativity, which predicts that the frequency ratio should be independent of velocity. *See* RELATIVISTIC ELECTRODYNAMICS.

The most accurate experimental result has been obtained by using the low-energy technique. This result is a(experimental) $= 0.00115965218085 \pm 0.00000000000076$, and it is in agreement, within the limits of error, with a(theoretical). The comparison of a(experimental) with a(theoretical) presented above constitutes the most precise confrontation of any experiment with a theoretical prediction in the history of science, and it provides the value of the fine-structure constant, $\alpha^{-1} = 137.035999710 \pm 0.000000096$, with a relative precision of 0.7 part per billion (10^9), which is an order of magnitude more precise than any other measurement available at present. *See* ATOMIC STRUCTURE AND SPECTRA; GYROMAGNETIC RATIO; QUANTUM MECHANICS.

Arthur Rich; Toichiro Kinoshita

Bibliography. R. Eisberg and R. Resnick, *Quantum Physics of Atoms, Molecules, Solids, Nuclei, and Particles*, 2d ed., 1985; G. Gabrielse et al., New determination of the fine structure constant from the electron g value and QED, *Phys. Rev. Lett.*, 97:030802, 2006; P. J. Mohr and B. N. Taylor, CODATA recommended values of the fundamental physical constants: 2002, *Rev. Mod. Phys.*, 77:1–107, 2005; B. Odom et al., New measurement of the electron magnetic moment using a one-electron quantum cyclotron, *Phys. Rev. Lett.*, 97:030801, 2006; H. C. Ohanian, *Modern Physics*, 2d ed., 1995; J. W. Rohlf, *Modern Physics from α to Z^0*, 1994.

Electron tomography

A powerful technique for determining three-dimensional (3D) structure from a series of two-dimensional (2D) images recorded in a transmission electron microscope. Transmission electron microscopy uses high-energy electrons to illuminate specimens at resolutions that in favorable cases surpass 1.0 nanometer (1/25,000,000 of an inch) and can resolve atoms. Resolution is not as great with electron tomography (4–8 nm) but is still much higher than can be achieved by light microscopy or medical imaging methods. However, the weak penetrating power of electrons requires operation under high vacuum and generally limits the technique to specimens less than about 1 micrometer in diameter. The most common application is

in biological research to study the 3D architecture of subcellular components such as mitochondria, nerve synapses, the Golgi apparatus, muscle cells during contraction, the arrangement of molecular components in bacterial cells, chromosome fibers, the kinetochore, and other structures. Such studies provide important data for understanding how cellular components carry out their vital functions. Electron tomography is also being used to study the structure of certain inorganic crystals used in the semiconductor industry.

Principles of operation. Transmission electron microscopy provides a translucent interior view of the specimen, analogous to what is seen in an x-ray image, rather than the surface view that we are accustomed to seeing in everyday life. Although this enables the microscopist to see inside the specimen, it has the disadvantage that the image formed is a 2D projection of the 3D object with features from different depths superimposed (**Fig. 1***a*). The solution is to reconstruct a 3D image of the original specimen from a series of 2D projections (Fig. 1*b*). This approach is also employed in medical imaging methods such as computerized axial tomography (CAT scanning) and magnetic resonance imaging (MRI).

The series of 2D images required for tomographic reconstructions is obtained by tilting the specimen in the electron beam. This is in contrast to medical imaging in which the patient is kept stationary while the beam source and detector are rotated to provide the necessary views. An important limitation of electron tomography is the inability to collect tilt data over the full 180° range required for complete angular coverage. (A 360° range is not necessary because, in projection imaging, views from the opposite side of the specimen provide the same information.) The difficulty in electron tomography is that typical electron microscope specimens are very thin (50–500 nm) and mounted on a support film that is more than 1000 times wider than the specimen. As a result, images can be recorded only over a limited tilt range of ±60 or 70° (a total range of 120–140°) before the specimen becomes too thick to view. This gives rise to a range of missing angular coverage that produces a well-characterized distortion that must be taken into account when interpreting electron tomographic reconstructions.

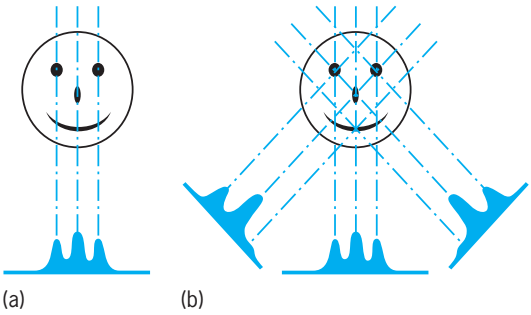


Fig. 1. Projection imaging and tomographic reconstruction. (*a*) Formation of one-dimensional (1D) projection of a two-dimensional (2D) smiley face. To form a true projection, the amount of penetrating radiation that strikes a unit length of the recording medium must be a measure of the amount of mass traversed by the beam in its path through the specimen. This is illustrated by the image density graph superimposed upon the 1D projection, and the broken lines that represent the radiation path through areas of maximum density in the specimen (that is, through the eyes or nose plus the mouth of the face). In transmission electron microscopy, true projections are formed as long as the specimen is not thicker than 0.2–0.5 micrometer. (*b*) Tomographic reconstruction of the 2D smiley face. The first step in tomography is to systematically record a series of 1D images from different viewing directions. The density in each 1D image is then projected back into a 2D area by simply distributing the density evenly along the original projection direction. When these "back-projection rays" from the different 1D images are summed, they intersect and reinforce one another only at the points where there was significant mass in the original object. In this way the 2D image is reconstructed from the 1D images. The same strategy is used to reconstruct a 3D image from 2D projection images. (*Adapted with permission from B. F. McEwen and M. Marko, Three-dimensional transmission electron microscopy and its application to mitosis research, Meth. Cell Biol., 61:82–111, 1999*)

Electron tomographic data sets are generally collected with 1 or 2° tilt angle intervals, because the resolution obtained in the resulting reconstruction is dependent upon how finely the angular range is sampled. Thus, a typical tilt series contains 60–140 images. Recording so many images was time-consuming and tedious with older microscopes, but recent improvements to mechanical design, along with incorporation of direct digital recording of the images and computer control of the microscope, have enabled newer microscopes to automatically record a tilt series of 140 images in less than 30 min. These technical developments in electron microscopy, coupled with the rapidly increasing computational power and the disk storage capacity of modern computers, are transforming electron tomography into a high-throughput tool that can be used ever more routinely in research and diagnostic settings.

Volume analysis and sample application. Perhaps the most surprising aspect of electron tomography is that it generally takes longer to analyze an electron tomographic reconstruction than it does to collect the data and compute the reconstruction. The difficulty is that the investigator is faced with displaying a three-dimensional tomographic reconstruction on a two-dimensional computer screen. However, in contrast to the original imaging situation, one now has a 3D image in the computer that can be taken apart and examined from any direction. Usually the quickest way to locate components of the structure is to view the reconstruction volume as a sequential stack of thin slices. However, understanding the 3D relationships of component parts and communicating the finding to a larger audience generally require constructing a 3D representation. For complex structures the most common approach is to segment select components away from the rest of the tomographic reconstruction by manual tracing and assignment of a color code.

Figure 2 illustrates the application of electron tomography to the study of the mammalian kinetochore. The kinetochore is a specialized structure that forms at a specific location on chromosomes during cell division. It functions to attach chromosomes to a bipolar spindle, generate chromosome motion, and exert control over the events of cell

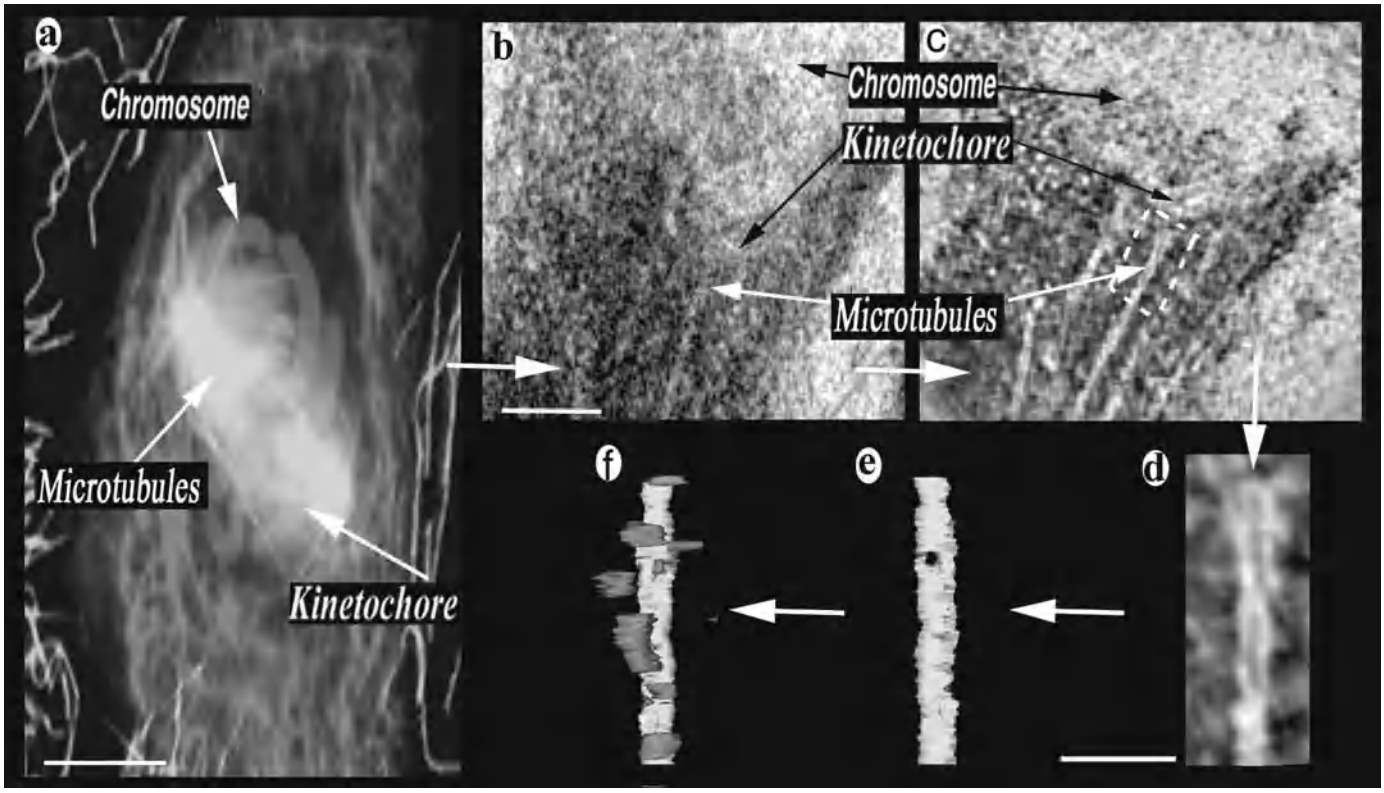


Fig. 2. Using electron tomography to analyze cell division. (*a*) An immunofluorescent light microscopy image of a mammalian cell during cell division. (*b*) Electron microscopy image of a kinetochore from the same type of cell. The kinetochore is the region of the chromosome that attaches to the mitotic spindle. (*c*) Single 2D slice from the 3D electron tomographic reconstruction of the kinetochore in *b*. Note the sharper microtubules in the tomographic reconstruction. (*d*) Single microtubule extracted from the tomographic reconstruction. (*e*) 3D model created by tracing and color-coding (light gray) the microtubule in *d*. (*f*) Same model as in *e* showing kinetochore components (dark gray) that are attached to the microtubule.

division. These functions are essential for proper cell division and the viability of the organism. Figure 2*a* is a gray-scale version of an immunofluorescence light microscopy image of a dividing cell, with arrows indicating a chromosome, location of a kinetochore, and microtubules. Microtubules are the chief structural components of the mitotic spindle. In the actual immunofluorescent image, microtubules are stained green, chromosomes blue, and keratin fibers, which form a cage around the spindle, red. One of the crucial yet poorly understood aspects of chromosome alignment is the interaction between the kinetochore and spindle microtubules. It is clear from Fig. 2*a* that the resolution of light microscopy is inadequate to study this interaction. Although single microtubules are visible in an electron microscopy projection image (Fig. 2*b*), details of the microtubule ends become visible only in single slices from the tomographic reconstruction (Fig. 2*c, d*). Manual tracings from trimmed slices such as in Fig. 2*d* were stacked into a 3D image, which are shown as a surface view in Fig. 2*e*. Kinetochore connections to the microtubules were also traced, and are shown in dark gray along with the microtubule in Fig. 2*f*. These and similar 3D models are being used to test current hypotheses on how the kinetochore interacts with microtubule ends.

Future directions and conclusions. Recent technical advances and the inherent versatility of electron tomography have accelerated its use and expanded the range of its applications. In particular, electron tomography is now being used for high-quality frozen-hydrated preparations of biological specimens. These preparations are made by rapidly freezing the specimen to prevent the formation of normal crystalline ice, which severely damages cellular structure. Such preparations are a snapshot of the specimen in its native hydrated environment and, thereby, superior to conventional preparations made by solvent extraction, heavy-metal staining, and plastic embedding. Previously the sensitivity of frozen-hydrated specimens to damage from electron irradiation had precluded such applications.

Electron tomography has been established as an important method for determining the structure of subcellular components and organelles, and the number and range of applications are likely to continue growing rapidly in the near future. *See* COMPUTERIZED TOMOGRAPHY; ELECTRON MICROSCOPE; FLUORESCENCE MICROSCOPE; MEDICAL IMAGING; OPTICAL MICROSCOPE.

Bruce F. McEwen

Bibliography. W. Baumeister, R. Grimm, and J. Walz, Electron tomography of molecules and cells, *Trends Cell Biol.*, 9:81–85, 1999; B. F. McEwen and M. Marko, The emergence of electron tomography as an important tool for investigating cellular ultrastructure, *J. Histochem. Cytochem.*, 49:553–563, 2001; B. F. McEwen and M. Marko, Three-dimensional transmission electron microscopy and its application to mitosis research, *Meth. Cell Biol.*, 61:82–111, 1999.

Electron-transfer reaction

A reaction in which one electron is transferred from one molecule or ion to another molecule or ion. Electron-transfer reactions are ubiquitous in nature. Some are deceptively simple [for example, reaction (1), where the asterisk is used to identify a specific isotope]; others look very complicated (for example, the long-range electron transfers found in biology). The widespread occurrence of electron-transfer reactions has stimulated much theoretical and experimental work. *See* BIOLOGICAL OXIDATION; CHEMIOSMOSIS; PHOTOSYNTHESIS.

$$^{*}Fe(H_2O)_6^{3+} + Fe(H_2O)_6^{2+} \longrightarrow {}^{*}Fe(H_2O)_6^{2+} + Fe(H_2O)_6^{3+} \quad (1)$$

The development of electron-transfer theory, as well as the measurement of the rates of oxidation-reduction (redox) reactions, began in the late 1940s. A great deal of experimental work has been carried out by inorganic chemists, and during the 1960s and 1970s the reactivity patterns of a large number of complexes were uncovered. Using this body of knowledge as a reference point, bioinorganic chemists studying the rates and mechanisms of biological electron-transfer reactions are seeking parallels with the redox behavior of less complicated metal complexes. *See* BIOINORGANIC CHEMISTRY; COORDINATION COMPLEXES.

Metalloproteins are more than just metal ions in disguise. For example, the redox behavior as well as virtually every property of a protein (excluding its amino acid sequence) are dependent upon the solution pH; redox proteins are very large polyelectrolytes whose prosthetic groups are typically found buried in the protein interior. Thus pH-dependent behavior in a metalloprotein may be much more complex than in a simple metal complex. Another important distinction between the redox reactions of proteins and the electron transfers of small metal complexes is the magnitude of the electron donor-to-acceptor distance. The relevant distance for small molecules is generally taken to be van der Waals contact, which is roughly 0.3 nanometer. In contrast, several experiments have shown that electrons can jump at significant rates across distances of 1 nm or more in protein interiors. *See* INTERMOLECULAR FORCES; PH; PROTEIN.

Experimental investigation of the factors that control the rates of biological redox reactions has not come as far as the study of the electron transfers of metal complexes, because of the large number of variables that must be dealt with, and problems associated with protein purification and stability. Many experimental approaches have been pursued, including the covalent attachment of redox reagents to the surfaces of metalloproteins and the use of mutant proteins (obtained by gene cloning). *See* BIOELECTRONICS.

Self-exchange and cross reactions. The simplest reactions in solution chemistry are electron self-exchange reaction (2), in which the reactants and

$$^{*}A_{ox} + A_{red} \rightarrow {}^{*}A_{red} + A_{ox} \quad (2)$$

products are the same (the asterisk is used to identify a specific isotope). The only way to determine chemically that a reaction has taken place is to introduce an isotopic label. There is no change in the free energy ($\Delta G^{\circ} = 0$) for this type of reaction.

Much more common are cross reaction (3), where

$$A_{ox} + B_{red} \rightarrow A_{red} + B_{ox} \quad (3)$$

A_{ox} is the oxidized reactant, B_{red} is the reduced reactant, A_{red} is the reduced product, and B_{ox} is the oxidized product. For these reactions, $\Delta G^{\circ} \neq 0$. The experimental determination of cross-reaction rates is generally more straightforward than the determination of self-exchange rates. Either the reactants are simply mixed together or a thermodynamically unstable system is generated rapidly (via pulse radiolysis, flash photolysis, or temperature-jump relaxation) to initiate the redox reaction. In almost all cases, electronic absorption spectroscopy has been used to monitor the progress of protein cross reactions. *See* SPECTROSCOPY.

Inner-sphere and outer-sphere mechanisms. Both types of electron-transfer reactions (self-exchange and cross reactions) can be classified broadly as inner sphere or outer sphere. In an inner-sphere reaction, a ligand is shared between the oxidant and reductant in the transition state. An outer-sphere reaction, on the other hand, is one in which the inner coordination shells of both the oxidant and reductant remain intact in the transition state. There is no bond breaking or bond making, and no shared ligands between redox centers. Long-range electron transfers in biology are all of the outer-sphere type.

Electron-transfer theory. The simplest electron transfer occurs in an outer-sphere reaction. The changes in oxidation states of the donor and acceptor centers result in a change in their equilibrium nuclear configurations. This process involves geometrical changes, the magnitudes of which vary from system to system. In addition, changes in the interactions of the donor and acceptor with the surrounding solvent molecules will occur. The Franck-Condon principle governs the coupling of the electron transfer to these changes in nuclear geometry: during an electronic transition, the electronic motion is so rapid that the nuclei (including metal ligands and solvent molecules) do not have time to move. Hence, electron transfer occurs at a fixed nuclear configuration. In a self-exchange reaction, the energies of the donor and acceptor orbitals (hence, the bond lengths and bond angles of the donor and acceptor) must be the same before efficient electron transfer can take place. *See* FRANCK-CONDON PRINCIPLE.

The incorporation of the Franck-Condon restriction leads to the partitioning of an electron-transfer reaction into reactant (precursor complex) and product (successor complex) configurations.

Steps (4)–(6) go from reactants to products: K is the

$$A_{ox} + B_{red} \rightleftharpoons [A_{ox},\ B_{red}] \quad (4)$$

$$[A_{ox}, B_{red}] \xrightarrow{k_{et}} [A_{red}, B_{ox}] \quad (5)$$

$$[A_{red}, B_{ox}] \xrightarrow{\text{fast}} A_{red} + B_{ox} \quad (6)$$

equilibrium constant for the formation of the precursor complex [:A_{ox}, B_{red}], and k_{et} is the forward electron-transfer rate to produce the successor complex [A_{red}, B_{ox}].

Some degree of electronic interaction, or coupling between donor and acceptor, is required if the redox system is to pass from the precursor to the successor complex. (This passage involves an intermediate species known as a transition state or activated complex.) The magnitude of the electronic coupling determines the behavior of the reactants once the transition state is reached. Two cases can be distinguished. If the electronic coupling is very small, as in so-called nonadiabatic reactions, there is a very low probability that the reactants will undergo electron transfer, leading to very little product formation. If the electronic interaction is sufficiently large, as is the case for adiabatic reactions, the probability of electron transfer occurring when the reactants reach the transition state is unity. The degree of adiabaticity of a reaction is characterized by a transmission coefficient κ, whose value ranges from 0 to 1. For systems in which the electronic coupling is sufficiently large, κ equals 1. This situation occurs when the reacting centers are close together, the donor and acceptor orbitals overlap, and no substantial changes in geometry are involved. The transmission coefficient is much less than 1 for electron transfer over long donor-to-acceptor distances, as commonly found in protein electron-transfer reactions and in some electron transfers in large organic molecules.

Long-range electron transfer. The rate of long-range electron transfer between an electron donor (B_{red} and an electron acceptor (A_{ox}) depends on both the electronic coupling between A_{ox} and B_{red} (which is a function of the intersite A_{ox}//B_{red} distance d, the nature of the intervening medium, and the relative A_{ox}//B_{red} orientation, where // represents the protein medium that separates the donor and the acceptor) and an activation energy term. A standard theoretical rate equation (7) expresses k_{et} in terms of these

$$k_{et} = \nu\ [\exp(-\beta d)]\{\exp[-(\lambda + \Delta G^{\circ})^2/4\lambda RT]\} \quad (7)$$

factors; here ν is a frequency factor, β is a medium- and orientation-dependent quantity, d is the intersite λ is the reorganization energy, ΔG° is the reaction free energy of the electron-transfer process, R is the universal gas constant, and T is the absolute temperature of the system. Experiments in several laboratories have been designed to estimate the values of λ and β in modified metalloproteins, rigid organic molecules, and protein-protein complexes.

Modified metalloproteins. Long-range intramolecular electron transfer is being studied in synthetically modified metalloproteins. Ruthenium (Ru) ammines have been chosen as modification agents primarily because they can be attached specifically to surface histidines, and the resulting complexes are kinetically stable in solution in the relevant oxidation states. X-ray crystallographic structure determinations have been completed for the native proteins used in these studies, and so the distance and intervening medium between the donor and acceptor sites in the modified protein are accurately known. [X-ray crystallographic studies of a modified protein show that only minimal structural perturbations result from covalent attachment of ruthenium to a surface histidine (His).] The driving force ΔG° for the electron transfer between donor and acceptor of a modified protein can be determined accurately through standard electrochemical experiments. Thus electron transfer can be studied in a system where two of the important variables in Eq. (7), d and δG°, are reliably known. *See* X-RAY CRYSTALLOGRAPHY.

Sperm whale myoglobin (Mb) has four surface histidines that can be modified by using ruthenium ammine complexes. The pentaammine-ruthenium(III) [a_5Ru] complex has been covalently attached to these histidines. After separation and purification of the reaction products, four singly labeled myoglobin species are obtained, with the closest edge-to-edge distances from the heme ranging from 1.27 nm (His-48) to 2.2 nm (His-12). Electron transfer from the His-48 ruthenium complex to the iron (Fe)–containing heme has been measured by selectively reducing the Ru center on the exterior of the protein using flash photolysis techniques. This is followed by electron transfer from Ru^{2+} to the heme Fe^{3+} in the interior of the protein. The forward electron-transfer rate k_{et} (Ru^{2+} to Fe^{3+}) in a_5Ru-(histidine-48)-modified myoglobin [a_5Ru(48)MbFe] was found to be 0.02 s^{-1}; the reverse rate k_r (Fe^{2+} to Ru^{3+}) was uniquely determined to be 0.04 s^{-1} by employing a method developed to generate the Fe^{2+}//Ru^{3+} mixed-valence species. These kinetic results clearly demonstrate that long-range electron transfer is reversible in this system.

Cytochrome c (from horse heart) also has been modified with the a_5Ru reagent. In contrast to the myoglobin system, cytochrome c has only one surface histidine (His-33) that readily binds ruthenium. The edge-to-edge distance from this histidine to the heme is 1.16 nm. By using flash photolysis, the Ru^{2+}-to-Fe^{3+} electron-transfer rate was found to be approximately 30 s^{-1}. Pulse radiolysis has also been employed on the same system, with a measured electron-transfer rate of 50 s^{-1}. *See* CYTOCHROME.

Driving-force dependences have been addressed in two ways with modified proteins. First, the ammine ligands on the ruthenium label have been varied to produce complexes with different reduction potentials. Second, the iron from the heme has been removed and replaced with zinc, magnesium, and palladium. After photoexcitation, the metalloporphyrin functions as an electron donor, leading to electron transfer to the surface-bonded ruthenium.

With these methods, a wide range of driving forces for intramolecular electron transfer in cytochrome *c* and myoglobin has been investigated. At the highest driving forces in cytochrome *c*, rates of roughly $3 \times 10^6\ s^{-1}$ were found. Values of the reorganization energy λ between 1.3 and 2 eV have been obtained from these experiments.

The distance dependence of k_{et} has been investigated in the a_5RuMbZn system. Values of β in the 8–10-nm^{-1} range have been extracted from these studies, thereby indicating that a protein is a good medium for electronic coupling of a donor and acceptor.

Rigid organic molecules. Long-range electron-transfer reactions have been investigated in molecules containing organic donors and acceptors separated by rigid hydrocarbon units. With hydrocarbon spacers of 1–1.2 nm, electron-transfer rates as high as $10^{10}\ s^{-1}$ have been observed. The rates in the organic molecules are 10,000 times faster than protein electron transfers at comparable donor-to-acceptor distances. In one study involving hydrocarbon spacers of lengths 0.7–1.4 nm, β was found to be 8.5 nm^{-1}.

Long-range electron-transfer reactions in rigid organic molecules have been studied only in organic solvents; electron-transfer theory suggests that the reorganization energy λ should be smaller in nonaqueous media than reorganization energies for comparable long-range electron transfers in aqueous solutions. Results obtained are consistent with theory; a reorganization energy of roughly 1 eV was found in a thorough investigation of driving-force effects on the rate of long-range electron transfer in a donor (hydrocarbon spacer)–acceptor system.

Protein-protein complexes. In biologically relevant partners, both redox centers are buried in a protein matrix; the entire electron transfer occurs through a protein medium and includes a protein–protein interface. The reorganization energy λ has been measured by determining both the temperature dependence and the free-energy (ΔG°) dependence of electron-transfer rates in these systems. These protein–protein complexes also are well suited for the study of the effects of the medium on biological electron-transfer rates. Two approaches have been used. One is to compare electron-transfer rates between homologous (same species) and heterologous (different species) protein redox pairs. The other is to use the technique of site-directed mutagenesis to change amino acids that lie between the two redox centers in the protein–protein complex.

The most commonly used experimental system involves metal substitution of zinc for iron in one of the porphyrin redox centers in the protein–protein complex. This system allows facile initiation of electron transfer through photoexcitation of the zinc porphyrin (ZnP). The excited zinc porphyrin (ZnP^*) may then decay to the ground state or transfer an electron: the electron-transfer reaction is $Fe^{3+}//ZnP^* \rightarrow Fe^{2+}//ZnP^+$. The Fe^{2+} in $Fe^{2+}//ZnP^+$ will transfer an electron back to ZnP^+, thereby regenerating the starting material. The forward rate ($ZnP^* \rightarrow Fe^{3+}$) is k_{et}, and the back rate ($Fe^{2+} \rightarrow ZnP^+$) is k_b. *See* PORPHYRIN.

Hybrid hemoglobins. Hybrid hemoglobin is the best characterized of the protein–protein complexes. In this system, the donor-acceptor separation distance and intervening medium are known from x-ray structure determinations. Both α(Zn)β(Fe) and α(Fe)β(Zn) hybrids have been studied. Electron transfer ($ZnP^* \rightarrow Fe^{3+}$) between the α_1 and β_2 subunits of hemoglobin has been observed: the metal-metal distance is 2.5 nm (edge-to-edge porphyrin distance is 2.0 nm). The ΔG° for the electron transfer is approximately 0.8 V, and at room temperature k_{et} is approximately $100\ s^{-1}$. The temperature dependence of k_{et} has been analyzed by both quantum-mechanical and semiclassical theories; values of λ in the 2–3-eV range have been obtained. *See* HEMOGLOBIN.

Effects of the medium. The effects of the medium on long-range electron transfer have been investigated in the cytochrome *c* peroxidase (CCP)//cytochrome *c* (cyt *c*) complex. By using cyt *c* Fe^{3+}//CCPZn, ΔG° for back electron transfer is approximately 0.85 V. When yeast cytochrome *c* peroxidase was used in this reaction, k_{et} for the complex with yeast cytochrome *c* was found to be 10-fold faster than that for complexes with tuna cytochrome *c* or horse cytochrome *c*. The rate k_b was found to be 1000-fold faster for yeast than for tuna cytochrome *c*. These data indicate that specific medium and interface contact effects may enhance the reaction in the homologous yeast pair. The fact that the homologous and heterologous pairs show such radically different rate enhancements for k_{et} and k_b suggests a conformational change after the forward electron transfer. *See* CONFORMATIONAL ANALYSIS.

By using site-directed mutagenesis, phenylalanine-87, which is located above the heme pocket, has been replaced by tyrosine and serine in yeast cytochrome *c*. Electron-transfer studies with CCPZn have demonstrated that k_{et} is not affected by these mutations. The back electron transfer, on the other hand, is unaffected for the tyrosine mutant but is decreased 10,000-fold for the serine mutant. These data suggest that aromatic amino acids act as mediators for electron transfer. Although reorganization energy effects caused by the mutation to serine cannot be ruled out, these experiments demonstrate that changes in the intervening protein medium can produce dramatic electron-transfer effects.

Driving-force dependence of ket. The ΔG° dependence of the electron-transfer rate has also been studied in cyt *c*//CCP by using a series of substituted horse cytochromes *c*. These experiments yielded an estimate of 1.5–2.0 eV for the reorganization energy λ. The λ values for this protein–protein complex and for hybrid hemoglobin are in reasonable agreement with those found for a_5Ru-modified proteins. The origin of the relatively large reorganization energies for protein electron transfers in aqueous solution is being actively investigated in several laboratories. *See* CHEMICAL THERMODYNAMICS; OXIDATION-REDUCTION.

Bruce E. Bowler; Walther R. Ellis, Jr.; Harry B. Gray;Thomas J. Meade

Bibliography. J. R. Bolton, N. Mataga, and G. M. McClendon (eds.), *Electron Transfer in Inorganic, Organic, and Biological Systems*, 1991; G. L. Closs and J. R. Miller, Intramolecular long-distance electron transfer in organic molecules, *Science*, 240:440–447, 1988; J. Mattay (ed.), *Electron Transfer I*, 1994; S. L. Mayo et al., Long-range electron transfer in heme proteins, *Science*, 233:948–952, 1986; R. A. Scott, A. G. Mauk, and H. B. Gray, Experimental approaches to studying biological electron transfer, *J. Chem. Educ.*, 62:932–938, 1985.

Electron tube

A device in which electrons can travel through a sealed chamber containing at least two electrodes and gas at a very low pressure. The gas pressure usually ranges from about 10^{-6} to 10^{-9} atm (10^{-1} to 10^{-4} pascal). At the low extreme of this pressure range, electron tubes are sometimes referred to as vacuum tubes, and at the high extreme as gas tubes. *See* GAS TUBE; VACUUM TUBE.

Electron emission. At least one of the electrodes must emit electrons, and at least one must collect electrons. The emitting electrode, the cathode, may emit electrons through one or more of four mechanisms: thermionic or primary emission, secondary emission, photoelectric emission, or field emission. Electrons must acquire more energy than they have in the conduction band of a metal in order to escape from the surface of a metal. They acquire this energy, respectively, in the four mechanisms listed above, from heat, electron or ion impact, a photon impact, or an external electric field. Photoelectric emission, first observed by H. Hertz (1887), is used in light-sensing devices, often in combination with secondary electron multiplication to amplify the current. Secondary emission, sometimes in combination with thermionic emission, plays an important role in magnetrons invented by A. W. Hull (1921) and in crossed-field amplifiers. Field emission is used in some experimental amplifiers, flat-panel display devices, and x-ray tubes, but by far the most common type of emitting electrode used in electron tubes, including cathode-ray tubes, is the thermionic cathode. *See* CATHODE-RAY TUBE; FIELD EMISSION; FLAT-PANEL DISPLAY DEVICE; MAGNETRON; PHOTOEMISSION; PHOTOMULTIPLIER; PHOTOTUBE; SECONDARY EMISSION; X-RAY TUBE.

Thermionic emission was first observed by T. A. Edison (1883) as "negative electricity" emitted from an incandescent filament (the Edison effect). J. J. Thompson (1897) identified the particle emitted from the heated filament as one of very small mass (an electron). O. W. Richardson (1902), M. von Laue (1918), and S. Dushman (1923) derived equations that describe the phenomenon. Pure metals such as tungsten and tantalum can reach temperatures sufficient to produce large currents (about 1 A/cm^2) through thermionic emission. Monolayers of elements such as thorium can reduce the filament temperature required for thermionic emission to about 1650°C (3000°F) by reducing the electron energy required of an electron to escape the surface (the work function). Tungsten or nickel coated with mixtures of barium, calcium, and strontium oxides discovered by A. Wehnelt (1903) or porous tungsten matrices impregnated with barium-calcium aluminate (the Philips cathode) can emit large currents at even lower temperatures (800–1150°C or 1500–2100°F). *See* THERMIONIC EMISSION.

Diodes. A diode is a two-electrode tube, with a cathode and a collecting electrode. Edison must have built a diode in order to discover thermionic emission. A. Fleming (1904) developed the first thermionic diode using an oxide cathode. Because the collecting electrode is usually operated at a positive potential with respect to the cathode in order to collect much of the available electron current from the cathode, it is called an anode. Even so, because of the thermal energy of thermionic electrons, the anode can collect some electrons when it has a slightly negative potential. In a diode, with a good vacuum, the electric field of the electrons in the space between the cathode and the anode just cancels the electric field produced by the anode at the cathode surface. Thus the number of electrons in this space must increase in proportion to anode voltage while their kinetic energy also increases in proportion to the voltage. The electron velocity is proportional to the square root of the kinetic energy, and because current is proportional to density times velocity, the current is proportional to the 3/2 power of the voltage. This is referred to as the space-charge-limited current. The proportionality constant is an invariant of the diode design and is known as the perveance. In gas diodes, ion neutralization of the electron space charge permits the maximum cathode current (determined by the supply voltage and the load resistance) to flow even at very low anode voltages. *See* CHILD-LANGMUIR LAW.

Diodes were used in great numbers as rectifiers until about 1950 or 1960. By that time, most had been replaced by silicon rectifiers, often used in series, to replace even very high voltage vacuum rectifiers. I. Langmuir (1923) and K. Blodgett (1924) calculated the space-charge-limited current and the potentials between electrodes in diodes with circular cylindrical and spherical electrodes. This theory remains important because it is used as a starting point in the design of highly convergent electron guns for very high power microwave tubes. *See* DIODE; SEMICONDUCTOR RECTIFIER.

Triodes. L. DeForest (1906) added a third electrode to a diode in order to control the current flow from cathode to anode. This third electrode, the grid, took the form of a fairly open array or mesh made of wires with a diameter small compared to their spacing. In this geometry, much of the electric field from the anode terminates on the grid, and the field from the grid that terminates on the cathode exerts a primary influence on the space-charge current that flows to or through the grid. When the grid is at a negative potential with respect to the cathode, current flows due to the anode field that leaks through the grid, but

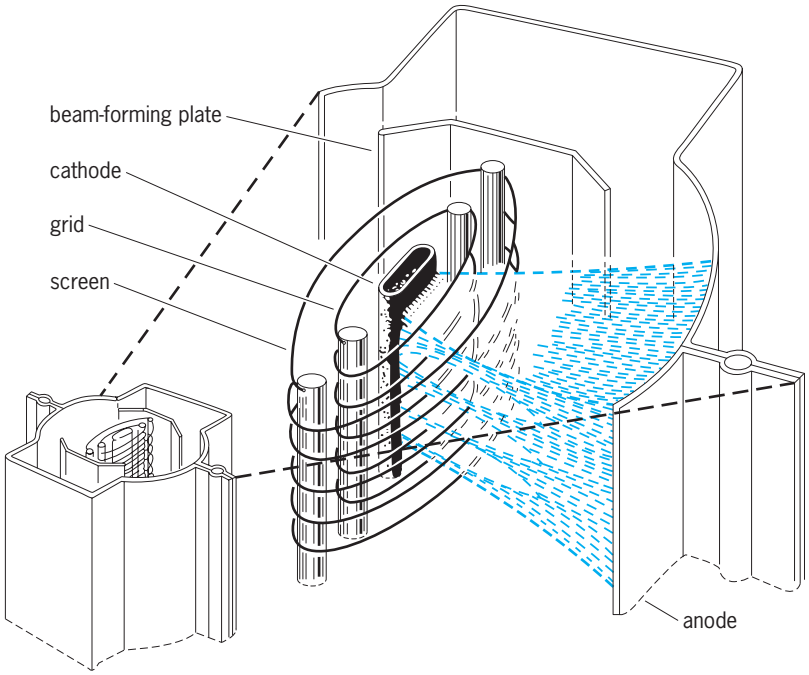


Fig. 1. Cutaway view of electrode arrangement in a beam-power tetrode. (*General Electric*)

the grid can collect no current. When the grid and anode are both positive, much more current flows and divides between the grid and anode.

By 1915, triode amplifiers and oscillators dominated the development of communications equipment. Gas-filled triodes, often called thyratrons and invented by Hull (1929), like gas-filled diodes, conduct much more current with much lower voltage drop than high-vacuum triodes because of space-charge neutralization by positive ions. The grid can start current flow and ionization, but cannot stop current flow until the ions are removed by reversing the polarity of the anode voltage.

Tetrodes and pentodes. Unfortunately, in triode amplifiers at high frequencies, the capacitance between the anode and grid electrodes, in combination with typical grid circuit reactances, can cause positive feedback, regeneration, or oscillation unless circuits that provide compensating negative feedback are used. For this reason W. Schottky (1919) invented the tetrode, which has a second or screen grid between the first or control grid and the anode. This grid was operated at a constant positive potential and effectively shielded the control grid from the anode. At large signal levels, it also created a problem by collecting secondary electrons emitted from the anode as a result of primary electron impacts when the instantaneous voltage on the anode was less than the screen grid voltage.

This problem was dealt with in two ways. G. Jobst and D. H. Tellegen (1926) introduced the pentode, which has a third very open suppressor grid between the screen grid and the anode. It was connected to the cathode. This created an electric field which returned secondary electrons to the anode. A more elegant solution to the secondary electron problem was provided in the beam-power tetrode (**Fig. 1**). In these tetrodes the anode was placed far enough from the screen grid that the charge of the electrons traveling between the screen grid and anode actually depressed the potential in the space between the screen and anode enough to return secondary electrons to the anode. This solution was analyzed by C. E. Fay, A. L. Samuel, and W. Schockley (1938). Beam-power tetrodes were popular in high-power transmitters and audio amplifiers because, in such applications, screen-grid-current interception and consequent power dissipation are problems. They are still popular in amateur radio equipment and high-power amplifiers used by musicians.

Replacement with solid-state devices. During 1950–1970, advances in solid-state diode and transistor technology, together with the development of small- and large-scale integrated circuits, resulted in the replacement of most low-power tubes with solid-state devices. There was really no contest because solid-state devices operate at much lower power input and at a much lower impedance level (lower voltage and high current) and hence provide higher bandwidth as well as efficiency. *See* INTEGRATED CIRCUITS; SEMICONDUCTOR DIODE; TRANSISTOR.

High-power tubes. However, for applications in which power serves a useful purpose (for example, transmitters), particularly at high frequencies, the outcome has been quite different. Because no electronic amplifier is 100% efficient, a high-power amplifier must dissipate power as heat. If the amplifier can run hot, it will be small and lightweight. If the temperature is limited to a low value, as it is when using most transistors, the amplifier will be large and heavy. Most of this size and weight will be in the metal heat sinks attached to the transistors.

The tubes and transistors discussed so far act as valves that control the flow of a current to a load. The potential energy of the current is derived from a direct-current power source. There is another, rather different class of electron tubes, most of which are referred to as microwave tubes, in which electrons are accelerated to a velocity at which they have a kinetic energy that is equivalent to the full voltage of the power supply that was used to accelerate them. If these electrons are bunched periodically in time, they can be made to give up their energy to the electric field in a gap or gaps in a very high frequency or microwave circuit. Microwave tubes include the inductive output tube, the klystron, traveling-wave tubes, crossed-field devices, and cyclotron-resonance devices.

There are no solid-state analogs to such tubes because the current flow in solid-state devices is too collisional. In electron tubes the electron density is very low, about 10^9–10^{11} electrons per cubic centimeter. At these densities, electrons do not collide with one another. Instead they move coherently and acquire a kinetic energy equal to the time integral of the scalar product of the electric field force and the velocity. As a result, in higher-voltage devices these electrons travel very rapidly (0.3 the velocity of light at 30 kV). Gap dimensions can be

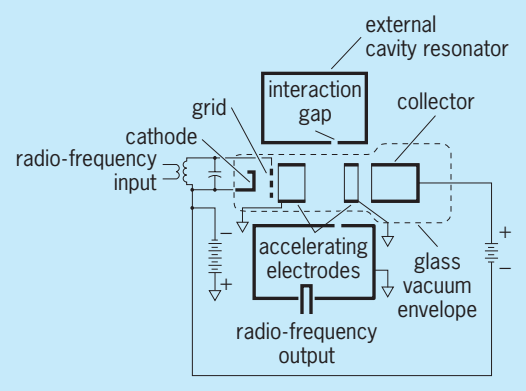


Fig. 2. Inductive output amplifier, showing separation of electron energy extraction from electron collection. (*After R. S. Symons, Tubes: Still vital after all these years, IEEE Spectrum, 35(4):52–63, April 1998*)

large for small transit-angle gaps. In a semiconductor, there are about 10^{19} carriers per cubic centimeter which have a great deal of thermal motion upon which is superimposed a low drift velocity proportional to the electric field. The carriers frequently collide with, and give up their energy to, the crystal lattice.

Inductive output tubes. In the inductive output tube (**Fig. 2**), invented by A. V. Haeff (1939), an electron beam is amplitude-modulated with a grid and then accelerated through a hole in the first accelerating electrode to form the high-velocity beam of electrons that passes through a gap in the center conductor of the coaxial external cavity resonator and into the collector. Inductive output tubes are used in many television transmitters operating between 470 and 900 MHz.

Klystrons. The klystron (**Fig. 3**), invented by R. Varian and S. Varian (1939), has a similar output cavity and collector, but has a beam which is first accelerated in a diode electron gun and then velocity-modulated in another reentrant cavity gap. Fast electrons overtake slowed electrons and yield an intensity-modulated beam by the time the electrons reach the output cavity. Additional cavities may be interposed between the input and output cavities (Fig. 3) to provide very high gain (often as high as 60 dB). *See* KLYSTRON.

Traveling-wave tubes. In traveling-wave tubes (**Fig. 4**), invented by R. Kompfner (1946), a high-velocity electron beam is velocity-modulated by, and gives up its energy to, periodically loaded or helical waveguides which slow the electromagnetic wave to a velocity nearly equal to that of the electron beam. Again very high gain is possible. *See* TRAVELING-WAVE TUBE.

Properties of linear-beam tubes. In all these linear-beam tubes, the electron beams are usually formed by highly convergent electron guns designed using techniques first devised by J. R. Pierce (1940). Computer simulation is now used to refine such designs. Steady magnetic fields provided by either electromagnets or permanent magnets are used to guide the electrons through the radio-frequency circuits. When the electrons leave these focusing fields, they repel each other due to their electric charge, and in addition, they cross radial magnetic field lines which also deflect them outward. The electron collectors on linear-beam tubes may have collection areas as much

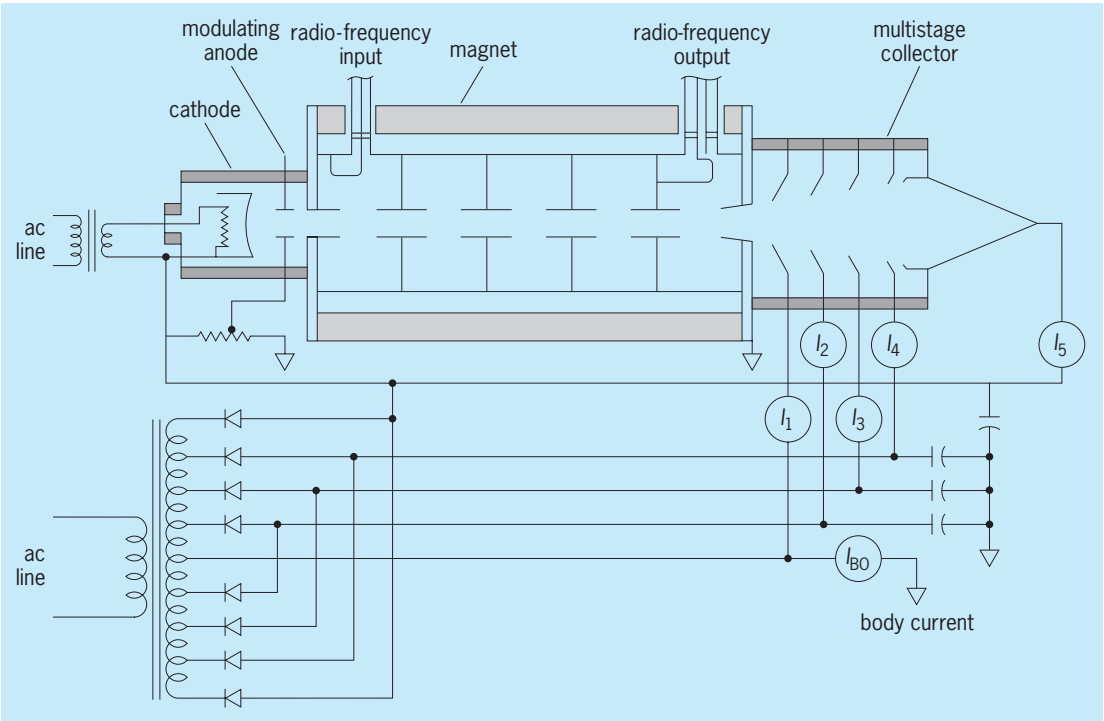


Fig. 3. High-efficiency, depressed-collector-klystron amplifier. (*After D. Christiansen, ed., Electronic Engineers' Handbook, 4th ed., McGraw-Hill, 1996*)

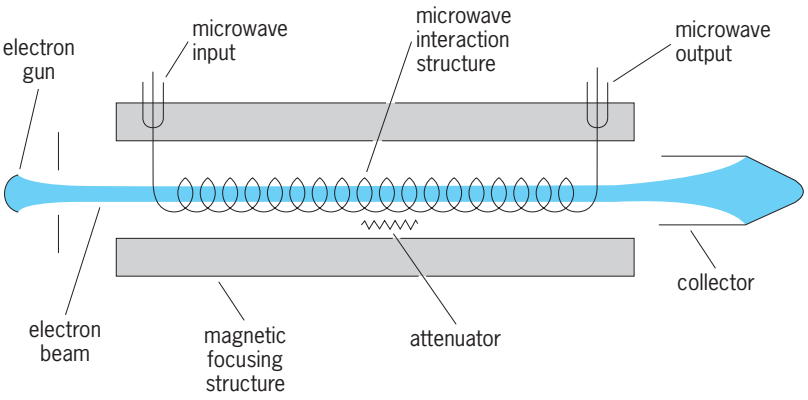


Fig. 4. Basic elements of a typical traveling-wave tube. (*After D. Christiansen, ed., Electronic Engineers' Handbook, 4th ed., McGraw-Hill, 1996*)

as 10,000 times the cross-sectional area of the electron beam, and continuous power as high as 1 MW has been achieved at frequencies as high as 10,000 MHz from such tubes. On klystrons and traveling-wave tubes, sometimes several collectors at different potentials are used (Fig. 3). These multistage depressed collectors, suggested by C. V. Litton (1940), can sort electrons by energy. Electrons that gave only small amounts of energy to the electric field in the radio-frequency circuit are collected on lower-potential electrodes. Tubes using these collectors are more efficient. R. Symons (1993) realized that an input-output tube with a multistage depressed collector could have the unique property of providing almost constant efficiency over a large range of signal amplitudes because both the current and the average collection voltage of the electrons could be made to go up and down with signal amplitude. Inductive-output tubes, klystrons, and traveling-wave tubes are used in television broadcasting, satellite communications systems, radar, scientific accelerators, medical accelerators used for cancer therapy, and military countermeasures equipment.

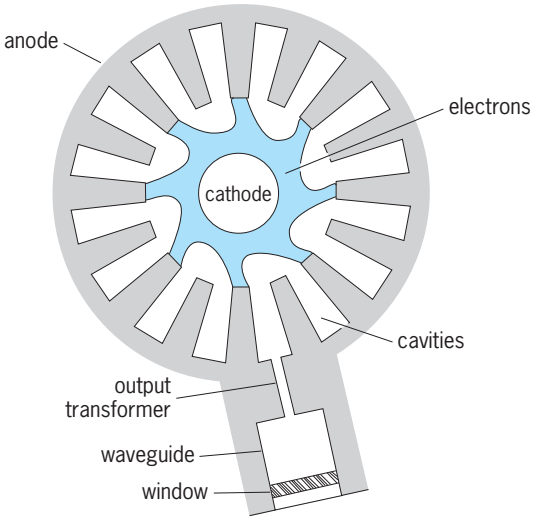


Fig. 5. Conventional microwave structure. (*After D. Christiansen, ed., Electronic Engineers' Handbook, 4th ed., McGraw-Hill, 1996*)

Magnetrons. In magnetrons (**Fig. 5**) and crossed-field amplifiers, electrons circulate about a cylindrical cathode in a radial static electric field and an axial magnetic field. Concentric with, and outside, the cathode is a periodically loaded transmission line that propagates a wave having circumferential electric field components that travel in synchronism with the rotating electron cloud. The electrons follow orbits that allow them to take energy from the radial static electric field and transfer it to the circumferential radio-frequency electric field of the wave on the circuit. Magnetrons are used in huge quantities in household microwave ovens. They and crossed-field amplifiers are also used in ground-based, shipboard, and airborne radars.

Cyclotron-resonance devices. Cyclotron-resonance devices including gyrotrons, gyroklystrons, and gyro-traveling-wave tubes again employ electrons that have been accelerated to the full energy provided by the electrical power supply. The beam is formed in a magnetic field so that it has a great deal of momentum perpendicular to the magnetic field, and the electrons follow helical paths. A radio-frequency electric field perpendicular to the axis of the electron trajectories will modulate the energy of the electrons and hence the relativistic mass and the cyclotron frequency. This azimuthal velocity modulation causes the electrons to draw into rodlike bunches that can give up their energy to a circuit supporting either the same alternating electric field that bunched them (in a gyrotron), or to an alternating electric field in another circuit (in a gyroklystron). Cyclotron-resonance devices are important because, up to a point, space-charge forces tend to create tighter bunches of electrons instead of driving electrons apart as they do in other linear beam tubes (inductive output tubes, klystrons, and traveling-wave tubes). For this reason, cyclotron-resonance devices can be built using very long circuits producing very weak electric fields, and as a result, having very low losses at very high frequencies. Efficient gyrotrons have been built at frequencies as high as several hundred gigahertz and have produced continuous power of hundreds of kilowatts. *See* GYROTRON; MICROWAVE TUBE.

Robert S. Symons

Bibliography. D. Christiansen (ed.), *Electronic Engineers' Handbook*, 4th ed., 1996; J. W. Gewartowski and H. A. Watson, *Principles of Electron Tubes*, 1965; A. S. Gilmour, Jr., *Microwave Tubes*, 1986; K. R. Spangenberg, *Vacuum Tubes*, 1948.

Electron wake

The pattern of electron density fluctuation and electromagnetic disturbance set up by the passage of a swift ion through condensed matter. In dense media that can sustain well-defined resonance oscillations at a frequency ω_0, wakes of periodic character will form behind swift charged particles having speed v. The periodicity in space, λ, the distance between

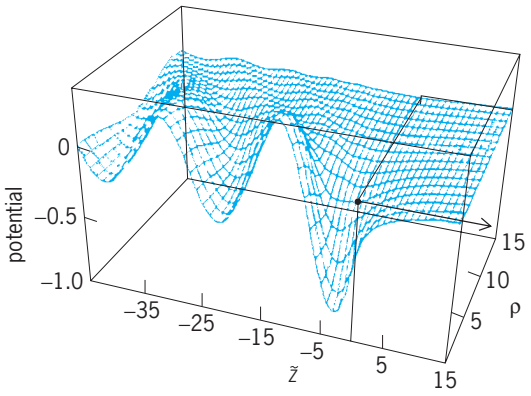


Fig. 1. Theoretical prediction of the scalar electric potential due to the electron wake of an ion penetrating a metallic medium having properties similar to those of aluminum. The ion has a speed three times the Bohr speed of 2.2 × 10^6 m/s. Coordinates $\bar{z}$ and ρ are measured from the ion position in directions parallel and perpendicular to the ion velocity, respectively, in units of the Bohr radius, a_0 = 0.0529 nm. The potential, which does not include that of the bare ion charge, is scaled by the ion charge and is given in units of the potential at a distance of a_0 from a proton. (*After P. M. Echenique, R. H. Ritchie, and W. Brandt, Spatial excitation patterns induced by swift ions in condensed matter, Phys. Rev., B20:2567–2580, 1979*)

troughs of the wake, is given by Eq (1). The oscil-

$$\lambda = \frac{2\pi v}{\omega_0} \tag{1}$$

lations (**Fig. 1**) trail behind the ion, move with the ion velocity, and have the frequency ω_0. In addition, close collisions between the ion and electrons of the medium cause electrons to recoil to form the analog of a bow wave ahead of the ion. *See* RESONANCE (QUANTUM MECHANICS).

In metals such as aluminum, the resonant frequency ω_0 is just the plasma frequency, ω_p, of the conduction electrons. The plasma frequency is given by Eq. (2), in SI units, where n_0, e, and m^* are the

$$\omega_p = \left(\frac{n_0 e^2}{\epsilon_0 m^*}\right)^{1/2} \tag{2}$$

equilibrium electron density and the charge and effective mass of the electron, respectively, and ϵ_0 is the permittivity of free space. Frictional effects, due to collisions with ions of the medium, give plasma oscillations a finite lifetime. Consequently, the wake is damped at distances of the order of 10 λ, which is typically about 20 nanometers behind the ion. Dynamical screening effects limit the wake to lateral distances on the order of v/ω_0 from the track of the particle. The maximum amplitude of the electric potential is given approximately by $Ze\omega_0/(4\epsilon_0 v)$, in SI units, where Ze is the ion charge. Electron wakes may be generated in media such as amorphous carbon and in insulators as well. *See* ELECTRICAL UNITS AND STANDARDS; FREE-ELECTRON THEORY OF METALS; PLASMA (PHYSICS).

The wake at the position of the guiding ion is of special significance. The electric field there times the ion charge represents the reaction of the medium to the ion and yields the stopping power of the medium for the ion, that is, the energy loss per unit path length.

Coulomb explosion. When molecular ions are injected into a solid with speeds greater than v_0 = 2.2 × 10^6 m/s, the so-called Bohr speed (the speed of an electron in the ground state of hydrogen according to the Bohr model), the valence electrons are stripped, leaving atomic ions to propagate as clusters of correlated charged particles through the medium. A dicluster is composed of two atomic ions traveling close together at nearly the same velocity. A wake is formed given by a (generally nonlinear) superposition of wakes due to the individual ions of the cluster. The dynamically modified Coulomb repulsion between its constituents causes the cluster, in effect, to explode. A pair of ions traveling with the same initial velocity, and created exactly abreast of one another, will recede rapidly from one another because of the Coulomb force acting on them, until polarization of the medium screens out the repulsive force at separations on the order of v/ω_0. At the same time, the center of mass of the pair will continue to move in the original direction. In most cases of interest, the initial speed is much greater than the relative speed of recession acquired by the pair. In typical experiments, the cluster-particle interaction probes the slope of the cluster wake potential near the origin, because the foils used in most experiments are so thin that the separation of the ions after traveling through the foil does not greatly exceed their initial separation. An important effect of the wake interaction is to cause the cluster to lose energy at a faster rate than would its isolated constituents traveling at the same speed, because the wake field of a given ion in a cluster acts, in most experiments, to retard the other ions of the cluster.

Measurements of the angular deflections and energy losses of protons resulting from diclusters formed from swift $(HeH)^+$ or $(OH)^+$ ions bombarding thin foils yield angular distributions (**Fig. 2**) that have a circular character due to the action of a Coulomb explosion. Such distributions generally have a large peaked region on the perimeter due to the trailing protons that are focused by the wake of the other ion in the Coulomb explosion. There is also a much smaller peak due to protons that lead the ion. Such experiments have been important in establishing the structure of molecular ions that were formerly not well known.

In other work with diclusters, the oscillatory character of the wake is vividly displayed in a two-foil experiment. A cluster enters the first foil and, after passing through a vacuum separating the carbon foils, enters the second one. The trailing ion experiences the wake force of the leading one. The dependence of the yield of secondary electrons from a final target on the distance between carbon foils shows the characteristic oscillatory behavior. *See* COULOMB EXPLOSION.

Nonlinear effects. It has been proposed that a moving positron (and possibly an electron) may become self-bound in the wake that it creates in matter. This composite entity might be termed a wakeon.