LECTURE NOTES IN ELECTRICAL ENGINEERING | 66

Amara Amara
Thomas Ea
Marc Belleville
Editors

Emerging Technologies and Circuits

Springer

Emerging Technologies and Circuits

Amara Amara • Thomas Ea • Marc Belleville
Editors

Emerging Technologies and Circuits

Springer

Editors
Amara Amara
Institut Supérieur d'Electronique de Paris (I.S.E.P)
21 rue d'Assas
75006 Paris
France
amara.amara@isep.fr

Thomas Ea
Institut Supérieur d'Electronique de Paris (I.S.E.P)
28 rue Notre Dame des Champs
75006 Paris
France
thomas.ea@isep.fr

Marc Belleville
CEA, LETI-MINATEC
17 rue des martyrs
38054 Grenoble Cedex 9
France
marc.belleville@cea.fr

ISBN 978-90-481-9378-3 e-ISBN 978-90-481-9379-0
DOI 10.1007/978-90-481-9379-0
Springer Dordrecht New York Heidelberg London

Library of Congress Control Number: 2010936131

Foreword

With the semiconductor market growth, new Integrated Circuit designs are pushing the limit of the technology and in some cases, require specific fine-tuning of certain process modules in manufacturing. Thus the communities of design and technology are increasingly intertwined. The issues that require close interactions and collaboration for trade-off and optimization across the design/device/process fields are addressed in this book. It contains a set of outstanding papers, keynote and tutorials presented during 3 days at the International Conference on Integrated Circuit Design and Technology (ICICDT) held in June 2008 in Minatec, Grenoble.

The selected papers are spread over five chapters covering various aspects of emerging technologies and devices, advanced circuit design, reliability, variability issues and solutions, advanced memories and analog and mixed signals. All these papers are focusing on design and technology interactions and comply with the scope of the conference.

Contents

Part I
Introduction

Synergy Between Design and Technology: A Key Factor in the Evolving Microelectronic Landscape

Michel Brillouët

1 Introduction

Microelectronics' impressive success can be attributed to three major factors:

- The transition for analogue to digital circuits as Gordon Moore predicted as early as in 1965 in his visionary paper [1]
- A virtuous innovation circle which fuelled the exponentional growth in revenue of this industry
- The decoupling of process and design flows with clear interfaces and sign-offs

So, why an enhanced synergy between design and technology is becoming now so important? This paper will address this question in the domains of the classical and equivalent scalings (the so-called 'More Moore' domain), of the integrated solutions (or 'More-than-Moore') and of the emerging research devices and architectures (or 'Beyond CMOS'). Finally it will discuss the interplay between this need and the present evolution of the microelectronic landscape.

2 Classical Scaling

2.1 Lithography

Patterning is a key enabling technology allowing reducing steadily the minimal dimension in an integrated circuit. However as we entered the sub-0.18 μm regime the critical dimension became lower than the wavelength of the exposure light.

M. Brillouët (✉)
CEA-LETI, Minatec, 17, rue des Martyrs, 38054 Grenoble Cedex 9, France
e-mail: michel.brillouet@cea.fr

A. Amara et al. (eds.), *Emerging Technologies and Circuits*,
Lecture Notes in Electrical Engineering 66, DOI 10.1007/978-90-481-9379-0_1,

What had predicted to be impossible proved to be manageable – namely printing sub-wavelength features – at the expense of more complicated pattern enhancement techniques. As shown in Fig. 1, additional sub-resolution patterns need to be added to have the designed feature be printed – without this correction the pattern may simply disappear from the wafer – and with the correct shape and size. However it turns out that the computation necessary for defining the added features is combinatorial and thus extremely CPU-time consuming. It results in huge data files and very long and complex mask writing increasing the cost exponentionally with the circuit complexity.

The question arises whether having the pattern printed as designed is absolutely necessity. In some cases – e.g. a contact hole – the pattern should having an appropriate size is needed, in other cases – e.g. the gate to isolation overlap or for double patterning – the relative alignment of two patterns on two different layers is critical, in some other cases – e.g. empty spaces filled with 'dummy' patterns – only the presence of the pattern is needed. One may thus expect that in the future the design intent – i.e. the functionality and criticality of the pattern – is fed in the OPC process, as it is done in mechanical engineering, in order to have 'just enough' correction and to reduce the overall cost of the process: more interaction is thus needed between the design community and the lithography engineer beyond what is presently done through OPC.

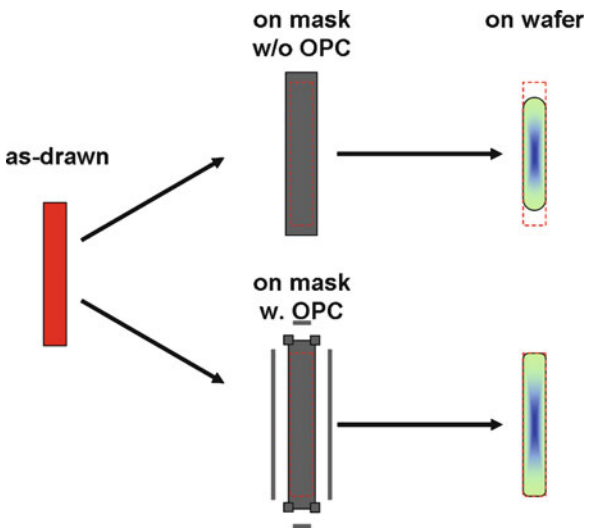


Fig. 1 Without optical proximity correction (OPC) a designed feature will not print on the wafer as expected. A pre-accentuation technique is needed to transfer the pattern as designed on the circuit

2.2 Interconnections

As the integrated circuits are growing in complexity more IP blocks are reused and assembled into more complex functions. It was assumed for some time that interconnecting those blocks was a functionally-neutral process. As N.S. Nagaraj [2] showed through many examples it may not be true anymore.

By adding more metallization layers it was necessary to planarize more the surface of the wafer in order to relax the burden on the lithography and etching processes: that had been obtained by applying Chemical Mechanical Polishing (CMP) at each metallization level. This technique induces however long-range waviness which is pattern-dependant. It does translate in the fact that the placement and orientation of the IP block will have an impact on the electrical characteristics of the block especially in analog applications. This can be mitigated by adding dummy patterns between the active features of the blocks, presenting thus a more uniform pattern density in the CMP: the drawback is that these added features may enhance cross-talk between active blocks. A deep understanding of the interconnection technology is thus needed to avoid that carefully characterized IP blocks don't behave as expected in the final product.

Routing interconnection lines over or near pre-characterized blocks may also induce marginality or failure in the integrated circuit as signal carried through these lines adds noise in the underlying blocks. The issue of signal integrity in complex circuits is however daunting: using distributed RLCG models one faces complex multi-scale physics of dynamic electromagnetic phenomena, with no simple return path of current in the wiring network, non-regular layout and exponentionally growing computation time. As stated in the ITRS [3], "reusable cores might require characterization of specific noise or power attributes ("field of use" or "assumed design context") that are not normally specified". One more time a deep understanding of the interplay between the expected functionality and the physical implementation calls for a stronger synergy between design and process.

Even simple ideas may prove to be painful to implement. For years wiring was drawn only in the x and y directions in the so-called Manhattan routing. In allowing wires in all 45° directions one can expect a reduction of 20% in interconnection length and 30% in the number of vias. However the whole supply chain from design tools and IP block redesign to mask and wafer manufacturing and control has to be adapted. A real assessment of the benefit of such a technique has to be made in real products through a strong cooperation between product designers and manufacturing people.

2.3 Yield Enhancement

Improving the yield of complex integrated circuits is a field where the synergy between design and technology is the most needed and apparent. Pinpointing physical

defects in a circuit is like looking for a needle in a hay stack. Only in combining test data, critical area analysis to predict potential vulnerable layouts and advanced physical characterization techniques one may hope to find the defect responsible for the circuit failure [4]. Through this analysis one may potentially mitigate further occurrence of this defectivity by improving either the manufacturing process and/or by optimizing the layout – e.g. in adding redundant via holes or by spreading out wires if there is enough room to do that way.

However many yield detractors can't assign to a specific type of physical defect and this trend is exacerbated in the most advanced technology nodes [5]. Actually deterministic yield limiters results from an inaccurate modelling of the process – device – product interactions which results in functional failures or marginalities in some blocks. Using worse case design methodology will induce too much penalty and a statistical design approach will save some margin in the final circuit. Regular layouts, transistor architectures less prone to variability of electrical parameters – e.g. by using undoped channel in SOI-like structures – can also mitigate the impact of these effects, but more research is needed, especially for the most advanced CMOS technologies, to address these issues.

In summary the gain expected in promoting more synergy between design and technology in classically scaled-down CMOS devices is driven by the need to guarantee manufacturing worthy products.

3 Equivalent Scaling

Just scaling the feature size is not enough anymore to pursue the performance increase at the transistor level. New materials – like high k dielectrics and metal gate, Ge or III–V compounds – and new technologies – e.g. strain engineering – need to be introduced to keep the historical pace. This general trend is called equivalent scaling in that it has the same effect as scaling dimensions without so much detrimental effects.

Especially important is to keep the electrostatic control of the gate on the canal while reducing the channel dimension: new transistor architectures are proposed like SOI or FinFET. This is however not design-neutral: new models and libraries are needed which take into account the specific effects of these new structures – e.g. the floating body behavior and the memory effect in PDSOI circuits . The huge effort needed to develop new libraries may be a real showstopper for introducing new MOS-like devices in complex design.

New design styles may have to be introduced. For example in the case of the FinFET the width of the transistor is quantized adding specific constraints in digital and analog bocks [6]. By a careful analysis of the different blocks needed in a library and taking into account the limitation in pitch of the most advanced lithographic technique, it brings to the conclusion that FinFET may be less dense than FDSOI for many IP blocks.

In conclusion by assessing new performance boosters from an IP design perspective one may be able to judge the real benefit of specific technological approaches.

4 Integrated Solutions

Scaling the digital part, how important it is, is not always enough to bring added value in the final product. Integrating in a single package with the scaled CMOS differentiated technologies, like interfaces between the digital blocks and the outside world which is fundamentally analog, is a challenge which the microelectronic industry faces more and more.

3D integration is in that respect an active research field. It may provide much higher performances for a digital system in reducing the footprint of the whole system and the number of I/Os and in shortening the inter-block wires. It may also be an answer to the present limitation of interconnecting heterogeneous parts in a single package. At the same time the resulting process can be rather complex and mismatch in wiring pitch and die size limit strongly the available options.

Though this 3D approach was demonstrated many times in laboratories, it is still a strong focus of many conferences. But this is more than just developing new specific manufacturing processes like high aspect ratio and dense Through Silicon Vias (TSV) or bonding of ultra-thin wafers. The benefit of 3D integration for a given application depends strongly on the design approach, including system partitioning, needed redundant periphery circuitry and 3D specific rules. C. Ababei has shown [7] that moving a memory array into three different levels may results – if not optimized – in a mere 8% enhancement in speed and 6% in energy per data access rather than the expected 40+% improvement. ASIC may bring added constraints like spreading blocks or adding thermal vias in order to mitigate the thermal load, managing electromagnetic interferences of the different layers as well as insuring the functionality of the thinned dies before and after 3D integration.

In summary a careful design of heterogeneous technologies integrated in a single package is a key enabler for adding more value in a given system.

5 Emerging Research Devices and Architectures

5.1 Bottom-Up Approach

ITRS [3] provides regularly a thorough discussion of the potential of emerging research devices which could complement or replace the existing CMOS digital gate. The analysis can be summarized as follow:

- Is the new device useful? Does it bring an interesting new functionality to the system?

- Is the device usable – e.g. does it allow room temperature operation –?
- Is this device really needed – i.e. significantly better than the "ultimately scaled" CMOS gate?
- Is the development effort worthy, having in mind the huge legacy and deep knowledge accumulated regarding the scaled-down MOS transistor? Is the new device improvable – e.g. scalable –? Is it versatile enough to address a wide range of applications?

Answering these questions brought the microelectronic community to the conclusion so far that for a digital switch there was no real alternative to the scaled down CMOS. Regarding very dense memories the statement is less definite and some proposals could prove their viability.

5.2 Top-Down Approach

This "technology-push" bottom-up approach is rather disappointing, but there is another way to look at the question of how emerging devices could enhance the system functionality. Starting from the system side, is there any innovative architecture which could benefit from the unique properties of these emerging devices? This "application-pull" top-down approach is somehow newer and fewer answers exist as a systematic methodology for addressing the question is missing. One of the reasons for this situation is that many architectural solutions can be derived from system specifications and these solutions, which lie in a design space resulting from many tradeoffs between technologies and resource constraints – e.g. area/volume, timing/speed, power/energy, memory size, error rate, etc. –, can translate into the use of many more devices and blocks.

J. M. Rabaey has outlined the growing importance of error rate as a specific design resource to be traded off with e.g. the circuit complexity and/or the power consumption (Fig. 2). Indeed in a deeply scaled technology the devices are expected to be less reliable. In this approach [8], a specific computation is performed by a

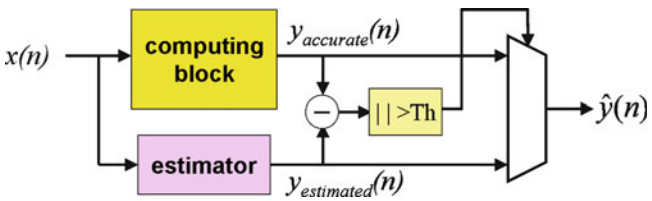


Fig. 2 In a deeply scaled-down technology gates are less reliable and the results of complex computation may be far from exact. By comparing the result of the full computation through a computing block with the estimation made by a much simpler and more reliable estimator, the reliability of complex computations can be enhanced. However error rate has to be traded off with the circuit complexity

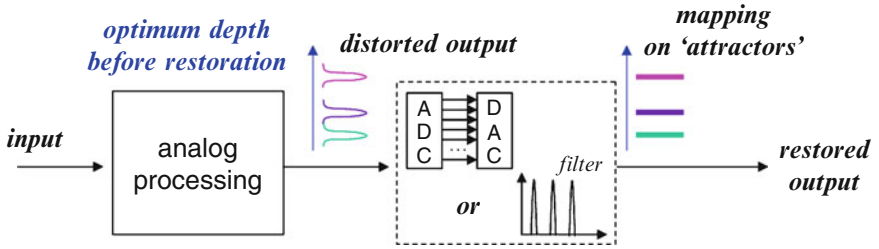


Fig. 3 Using analog processing with a limited computation depth may be a more efficient way to process data than a pure "brute-force" digital computation

complex computing block. In parallel the same input data are processed giving an estimate of the result through a less complex block where the gates are more conservative and thus more reliable. A circuit compares the results of the exact computation and its estimation: if the outputs of the two blocks are near enough the computation is assumed to be exact; if not the data can be processed again.

The information processing may also be more efficiently performed using analog blocks for a limited depth of computation [9]. The signal is then restored by "mapping" the distorted and noisy output on "attractors", i.e. well defined signal levels (Fig. 3). This approach may appear unlikely to succeed: it is however this principle which is applied when adding repeaters in an interconnection line.

5.3 "Morphic" Systems

The latest version of ITRS [3] outlined an attractive approach called "morphic" system: "morphic systems refer to architectures adapted to effectively address a particular problem set, often gaining their inspiration from biological or scientific computational paradigms." It should be outlined that the information processing should be still generic enough to have a chance to displace the present computational paradigm.

A wide set of such "morphic" systems are "self-relaxing" systems (Fig. 4). In such an approach a physical setup, initially in a stable state, is brought in an "excited" state by forcing inputs as new boundary conditions. By some more or less controlled mechanism the system relaxes to the "ground state" which represents the expected result. This result is read by looking at some outputs. Unfortunately this approach is most often less attractive than expected. If one has many different relaxation paths the computation time may become rapidly unpredictable. Furthermore the system can be stuck into metastable states which will provide wrong results. There are ways around these limitations, like applying a complex clocking scheme (e.g. in Quantum-dot Cellular Automata or QCA), adding noise (e.g. in probabilistic CMOS), implementing simulated annealing, etc. Unfortunately these approaches add a significant

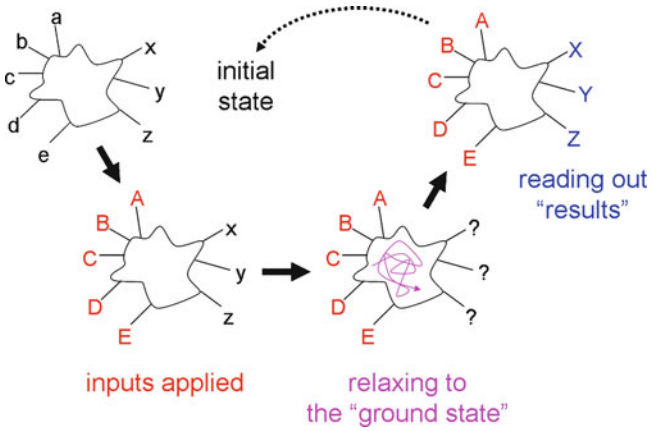


Fig. 4 A "self-relaxing" system is a physical setup which perform some processing on inputs defined as forced boundary conditions. This processing is obtained by letting the system relax in its "ground state" and the result is read at the outputs

complexity to the system and in some case (e.g. QCA [10]) are unlikely to beat even the present CMOS technology in complex computations.

In summary assessing emerging devices and architectures through the tentative design of some function of medium complexity ensures a reality check which avoid unrealistic claims and hypes.

6 What Business Model?

Having described few examples where the synergy between the design and technology communities is beneficial it is worth checking if the present evolution of the microelectronic landscape will continue to favors this synergy.

For decades the turnover of the microelectronic industry has grown 15–17% per year. This has been obtained through a technology push where new technologies created new markets which generated more revenue and margin, allowing more research effort to generate new ideas (Fig. 5). This virtuous circle is the real innovation engine which fuelled the exponential growth of this industry.

Unfortunately despite the optimistic statement of Gordon Moore –that "No exponential is forever, but we can delay 'forever' " [11] – there are many signs that the microelectronic industry undergoes major chances at an accelerated pace. Since the mid 1990s the CAGR is merely in the range of 5–8% and the microelectronic industry is more and more driven by financial considerations than by the "technology push" it used to be. Moreover the R&D cost tends to grow faster than the industrial revenue [12] owing to the formidable technical challenges we have to face in scaling further down the minimum feature sizes.

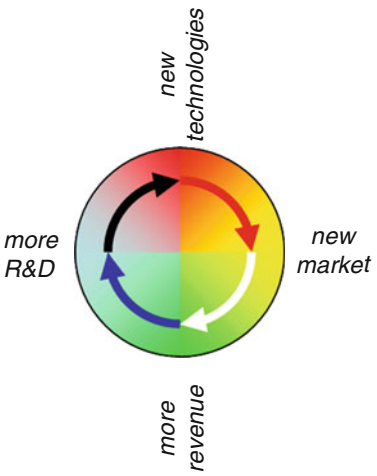


Fig. 5 The steady technological progress created new markets which generated more margin for increased R&D funding. This virtuous circle was the motor of the sustained exponential growth of the microelectronic industry

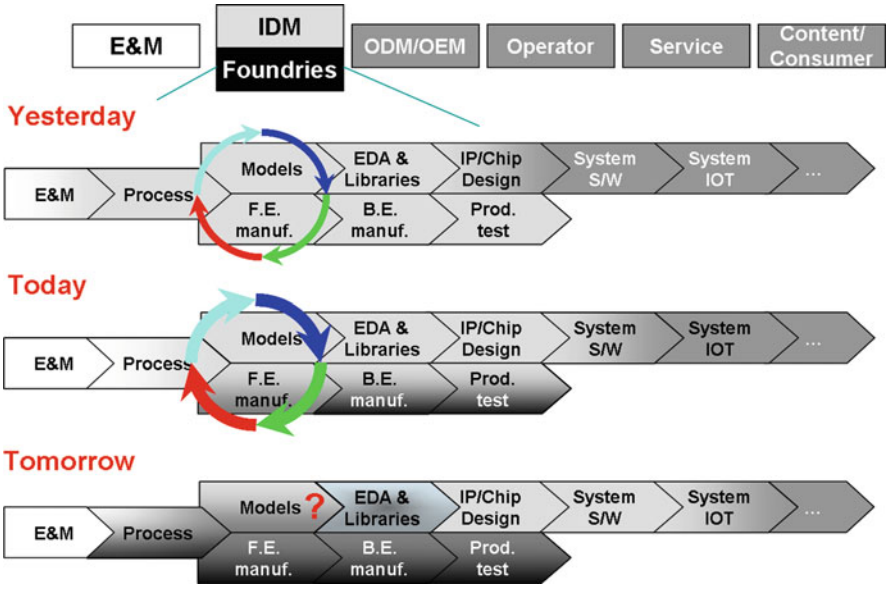


Fig. 6 Over time the integrated device manufacturers (IDMs) outsource more and more the hardware part of ICs, while moving up the supply chain towards richer software offering in their products. It brings some question mark about the way the interaction between process, design and manufacturing will take place in the future (From [13])

This trend was particularly clear when major ASIC companies announced in January 2007 that either they gave up the development of advanced digital CMOS by themselves, relying of foundries, or that they joined major consortia to further develop their expertise in the domain. The supply chain – especially for these applications requiring ASICs – seems to evolve along the following lines (Fig. 6):

- In the past – till the late 1980s – integrated device manufacturers (IDMs) addressed a significant part of the IC manufacturing, sometimes developing new equipments and materials, mastering the process development and production along with the design of integrated circuits
- More recently – basically in the 1990s – IDMs outsourced the development of equipment and materials to specialty companies, while most of the back-end part of manufacturing process (i.e. packaging and test) was subcontracted. Part of the IC manufacturing was also subcontracted to foundries which provided over that period more and more advanced processes. At the same time the IDMs moved up the value chain integrating more and more embedded software in their products.
- From the mid 2000s it appears that the foundry – fabless/fablite model gain acceptance, owing to the huge investment needed to develop and manufacture the most advanced CMOS processes. At the same time IDM's move towards more system-level software inclusion in their product along with some consideration towards system inter-operability (IOT).

What will be then the future for the synergy between the design and process communities, looking at this increasing fragmentation and specialization in the supply chain? For the biggest companies which can still afford and/or need to keep process development and manufacturing in house, this synergy will probably be stronger than ever. For the major players which choose the foundry / fablite model, new ways of cooperation need to be put in place. Interacting only through device modeling and design kits will not be enough to take full advantage of the newly developed processes: new supplier – customer relationships should be invented for staying competitive outside of specialty markets.

7 Conclusion

Synergy between design and technology is more and more needed to take full profit of the advanced devices and processes and this has to be done at the earliest stage of the R&D in order to build competitive advantages in the world competition. More specifically this interaction is beneficial in the classical CMOS scaling to ensure manufacturability of products and in the equivalent scaling to make a better use of innovative devices. Integrating system solutions in package will be more and more driven by the applications which will pull the development of new technological approaches. Finally this interaction between circuit design and emerging research devices and architectures is strongly recommended as a reality check for new ideas in order to prevent hypes and dead ends. It remains to be seen if the present de-verticalization of the microelectronics industry will allow developing new models of synergy between the two communities.

References

1. G.E. Moore, Cramming more components onto integrated circuits. Electronics **38**(8) (April 19, 1965); reproduced in Proc. IEEE, 86(1), 82–85 (1998)
2. N.S. Nagaraj, Impact of interconnect technology scaling on SOC design methodologies, *2005 IEEE International Interconnect Technology Conference*, pp. 71–73, 2005
3. *International Technology Roadmap for Semiconductors, 2007 edition*; http://www.itrs.net
4. F. Lorut, M. Lamy, S. Fabre, M. de la Bardonnie, K. Ly, R. Ross, C. Wyon, L. F. T. Kwakman, An effective failure analysis strategy for the introduction of 90 and 65 nm CMOS technology nodes, *2005 International Symposium for Testing and Failure Analysis*, paper # 19-1, 2005
5. A. J. Strojwas, Tutorial on design for manufacturability for physical design, *2005 International Symposium on Physical Design*, 2005; the presentation is available at http://www.sigda.org/ispd2005/ispd05/papers/2005/ispd05/htmfiles/ispdtoc.htm
6. E.J. Nowak, Turning silicon on his edge. IEEE Circuit Devices Magaz **20**(1), 20–31 (2004)
7. C. Ababei, Y. Feng, B. Goplen, H. Mogal, T. Zhang, K. Bazargan, S.S. Sapatnekar, Placement and routing in 3D integrated circuits. IEEE Des Test Comput **22**, 520–529 (2005)
8. J. M. Rabaey, Curing the ailments of nanometer CMOS through self-healing and resiliency, *2006 IEEE SoC Conference*, 2006
9. R. Sarpeshkar, Analog versus digital: extrapolating from electronics to neurobiology. Neural Comput **10**(7), 1601–1638 (1998)
10. K. Nikolić, D. Berzon, M. Forshaw, Relative performance of three nanoscale devices—CMOS, RTDs and QCAs—against a standard computing task. Nanotechnology **12**, 38–43 (2001)
11. G. Moore, No exponential is forever, but "forever" can be delayed! *2003 IEEE International Solid-State Circuits Conference*, paper # 1.1, 2003
12. D. Hutchenson, The R&D crisis. VLSI Research Doc:600201, Jan 28, 2005
13. H. Eul, Semiconductor industry in transition. Presented at ISS Europe 2008

Part II
Emerging Technologies and Circuits

New State Variable Opportunities Beyond CMOS: A System Perspective

Victor V. Zhirnov, Ralph K. Cavin, and George I. Bourianoff

1 Introduction

HERE is an international effort underway to discover a replacement for the CMOS transistor which will, within about one decade, not submit to further feature size scaling. There are many different candidates to replace the CMOS FET, but according to ITRS [1], none of them appear at this time to offer functional properties that are universally superior to the extremely scaled FET.

In seeking new device opportunities, it is important to understand not only the device characteristics, but also how connected systems of these devices might be used to perform complex logic functions. A taxonomy is offered for information processing devices in this paper. Performance estimates at the limits of physical scaling are described for those devices whose physics of operation is governed by the creation and control of energy barriers and these estimates are employed to begin investigation of the role of computer architecture on achievable limits for energy-efficient operation of the system. In Fig. 1 below, implications of the choice of state variable on system performance at various levels of abstraction are depicted. Metrics for binary switch operation include size, L_{sw}, switching time, t_{sw}, and switching energy, E_{sw}. Communication can range from local device-to-device connectivity to on/off chip connectivity by wired or wireless means. In any case, we show in the sequel that the energy required for communication is proportional to the number N_{car} of information carriers employed. System capability in Fig. 1 refers to the performance μ achievable by the system given the device and communication technologies with which it is implemented. The measure of

V.V. Zhirnov (✉) and R.K. Cavin
Semiconductor Research Corporation, 1101, Slater Rd, Durham, NC 27703, USA
e-mail: zhirnov@ src.org

G.I. Bourianoff
Intel Corporation, Austin, TX 78746, USA
e-mail: georgei.bourianoff@intel.com

A. Amara et al. (eds.), *Emerging Technologies and Circuits*,
Lecture Notes in Electrical Engineering 66, DOI 10.1007/978-90-481-9379-0_2,

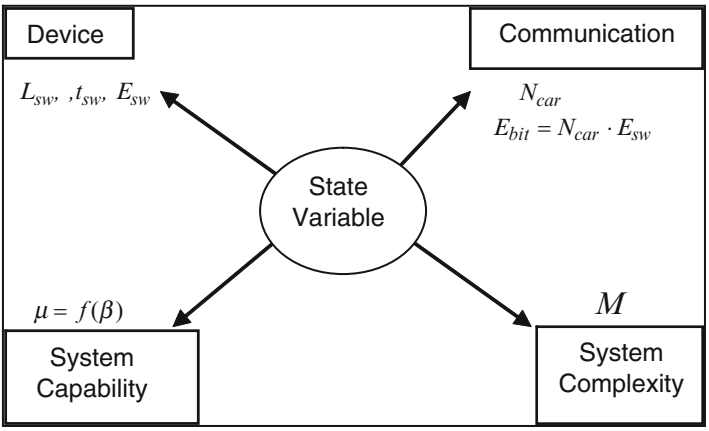


Fig. 1 State variable and different facets of information processing system

technology capability is binary information throughput $\boldsymbol{\beta}$. Finally, many system applications are narrowly focused, and novel device functionality may enable system scaling as measured by $\boldsymbol{M}$, i.e. the device count per function.

2 System Capability

One indicator of the ultimate performance of an information processor, realized as an interconnected system of binary switches, is the *maximum binary throughput (BIT)*; that is the maximum number of binary transitions per unit time per unit area. It is the product of the number of devices M with the clock frequency of the microprocessor f:

$$\beta = Mf \tag{1}$$

The computational performance of microprocessors μ is often measured in (millions) of instructions per second (MIPS) that can be executed against a standard set of benchmarks. It seems reasonable that there would be a strong correlation between microprocessor performance increases and corresponding improvements in integrated circuit technology, which are reflected in β. Performance data, μ, for a family of Intel microprocessors ranging from the 8080 to the Pentium 4 is plotted on a log-log scale against the technology metric, β in Fig. 2. The correlation is evident and to a good approximation,

$$\mu = f(\beta) = k\beta^{p} \tag{2}$$

For the selected class of microprocessors, $\mathrm{k} = 10^{-7}$ and $p = 0.6$.

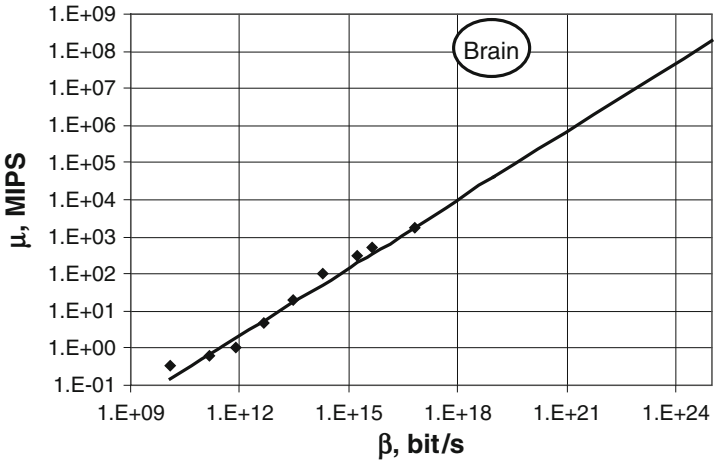


Fig. 2 Empirical relation between binary throughput and computational performance for several generations of Intel processors [2–4]

Note also from Fig. 2, that the human brain performance projection lies above the projected microprocessor line, giving rise to the hope that there may exist alternate technologies and computing architectures offering higher performance at much lower levels of energy consumption. Moreover, (1) suggests that there may be some fundamental relationship between the 'work done' by the processor in the sense of Turing and the underlying physics of the implementation technology. This is called the *Turing-Heisenberg rapprochement* and it is the purpose of this paper to offer a possible pathway to resolving this hypothesis.

Several alternative information processing technologies are being evaluated as possible adjuncts to CMOS. An important question is how these technologies will impact computing architectures, and more to the point, what is the functional relation between capabilities of these technologies and computer performance? In the next section, we briefly outline some of the candidate technologies in the context of the properties of a generic device.

3 Alternate Device Options

The basic computational element in digital information processing systems is binary switch (Fig. 3). In its most fundamental form, it consists of:

1. Two states 0 and 1 (state variables), which are equally attainable and distinguishable
2. A means to control the change of the state (WRITE)
3. A means to read the state
4. A means to communicate with other binary switches (TALK)

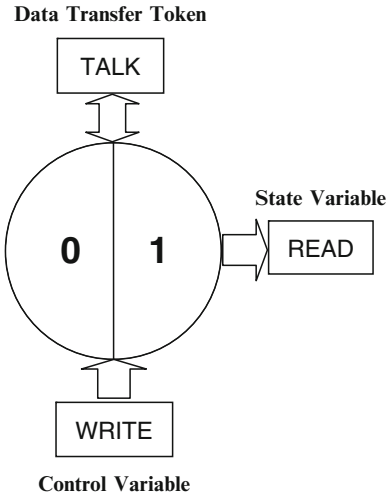


Fig. 3 A generic information processing device

The state representation, controls, and the outputs (READ and TALK operations) are all physical entities such as particles, quasi-particles, collections of particles, etc. and in the sequel, we call them *tokens*. Each token has a set of physical attributes associated with it, e.g., charge, mass, spin, etc. and it is the physical interaction between the token attributes within the device structure that determines the operation of the device. In most cases, an attribute can assume several values and for this reason, we call them *variables*.

When implementing circuits from devices, it is usually beneficial from an energy and space standpoint to choose tokens and their associated variables such that transformations are not required between input, output, and control variables. There are many ideas for future information processing devices for logic and we have mapped many of these devices into the taxonomy described above in Table 1 below.

4 Device and Interconnect Abstractions

4.1 Device Scaling Physical Limits

The purpose of this section is to lay the groundwork for estimating the performance limits of an information processor beginning with fundamental device physics. In order to do this, it is necessary to relate the device and interconnect system properties such as switching energy and times, number of electrons, etc., to the physical layout of the processor. It will be argued that the layout geometry can, in the limit, be viewed as an assembly of small *square tiles*, to each of which is associated size, energy and travel time parameters derived from basic physics. In all

Table 1 Taxonomy for candidate information processing devices

Device	Information token	Control variable	State variable	Data transfer token
FET – Novel Materials (2I-V's, carbon-based, etc.)	Electron	Charge	Charge	Electron
SpinFET	Electron	Charge and spin	Charge	Electron
Spin-torque	Electron	Spin	Spin	Electron
Spin-wave	Electron	Spin waves	Spin	Electron Photon
Tunneling transistor	Electron	Charge	Charge	Electron
Molecular transistor	Electron or atoms	Charge	Charge	Electron
NEMS	Atoms	Charge	Charge	Electron
Atomic switch/ electrochemical metallization	Atoms	Charge	Charge	Electron
Memristor	Atoms	Charge	Charge	Electron
Magnetic cellular automata	FM domain	Magnetic dipole	Spin	FM domain
Moving domain wall	FM domain	Magnetic dipole	Spin	FM domain
Multi-ferroic tunnel junction	FM domain	Spin	Charge	Electron
Optical or plasmonics	Atoms or electrons	Charge	Optical Density	Photons
Exciton	Excitons	Photons	Charge	Excitons (or photons)
Thermal transistor	Phonons	Thermal energy	Temperature	Phonons
Phase change	Atoms	Thermal energy	Charge	Charge
Quantum interference devices	Electron	Charge	Charge	Electron

cases, in order to minimize energy consumption estimates, the device and interconnect systems will be assumed to operate at the threshold of failure.

As the authors have argued in a series of papers [5–7] all known binary devices, regardless of the physics of their operation, can be represented by the generic barrier model of Fig. 4.

The energy barrier is needed to preserve a binary state in the presence of classic (thermal) and quantum (tunneling) errors (noise). The barrier properties, namely barrier height, E_b, and barrier width, a, determine the lower bound on the operational energy and size of binary device. The minimum Eb can be estimated from the Boltzmann probability of thermally-induced over-barrier transitions:

$$\Pi_{err} = \exp\left(-\frac{E_b}{k_B T}\right) \tag{3a}$$

Requiring $\Pi_{err} \leq 0.5$ (error probability less than 50%), from (3a), we obtain the minimum E_b:

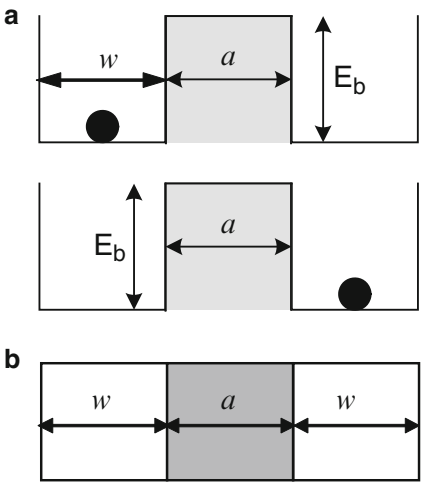


Fig. 4 A barrier model for a binary switch: (**a**) Schematic energy diagram, (**b**) Generic floorplan

$$E_b^{\min} = k_B T \ln 2 \sim 10^{-21} J(T = 300K) \tag{3b}$$

The minimum barrier width should be sufficient to suppress tunneling, and it can be estimated from the Heisenberg coordinate-momentum relation:

$$\Delta x \, \Delta p \geq \frac{\hbar}{2} \tag{4}$$

Tunneling is significant when $a \sim \Delta x$. In this case, a_H is the *Heisenberg distinguishability length* for "classic to quantum transition". Specifically,

$$\Delta x \geq \frac{\hbar}{2\sqrt{2mE_b}} = a_H \tag{5}$$

From (3b), obtain the Heisenberg-Boltzman limit for electron-based devices ($m = m_e$, the electron mass):

$$a_{HB} \sim \frac{\hbar}{2\sqrt{2m_e k_B T \ln 2}} \sim 1nm \tag{6}$$

Relations (3–6) govern the operations of all binary devices operating in equilibrium with the thermal environment, independent of their physical realization.

4.2 Device Density: Tiling Considerations

The binary switch one-barrier-and–two-wells energy diagram of Fig. 4a also suggests a generic topology of the ultimately scaled device, which is shown in

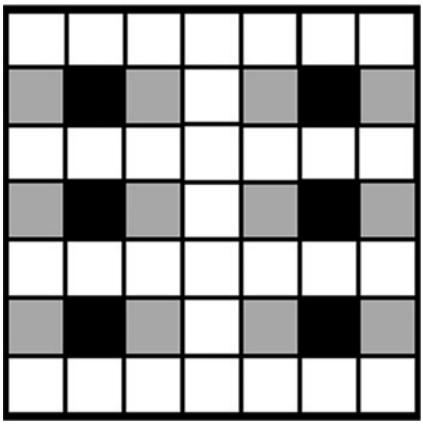


Fig. 5 Most compact device layout

Fig. 4b. Note that, while the smallest barrier size is limited by tunneling, the smallest well size, w, is limited by the quantum confinement. Interestingly, they both can be approximated by (5) and (6), and in the limiting case, $w_{min} = a_{HB} = a$. Thus, the two-dimensional floorplan of a smallest possible binary switch is $3a \times a$ rectangle consisting of three *square tiles* of the same size a. The most compact layout for an array of devices must allow at least one tile between each device for insulation as shown in Fig. 5. The tiling representation of binary switches allows one to calculate the maximum theoretical packing density of binary switches on a 2D plane [12]:

$$n_{\max} = \frac{1}{8a^2} \tag{7}$$

Substituting (6) in (7) the maximum device density $n_{\max} \sim 10^{13}\,\mathrm{cm}^{-2}$ is obtained.

4.3 Minimum Device Switching Time

The next pertinent question is the minimum switching time, which we refer to as *"Heisenberg time"* τ_H. This can be derived from the Heisenberg relation for time and energy:

$$\Delta E \Delta t \geq \frac{\hbar}{2} \tag{8}$$

or

$$\Delta t_{\min} \cong \frac{\hbar}{2\Delta E} = \tau_H \tag{9}$$

Equation 9 is an estimate for the maximum speed of dynamic evolution or the minimum passage time. It represents the zero-length approximation for the switching speed.

For the Boltzmann's limit, $E_b = k_B T ln2$, obtain

$$\tau_{HB} = \frac{\hbar}{2k_B T \ln 2} \approx 2 \cdot 10^{-14} s \tag{10}$$

Note that in physical systems, the Heisenberg speed limit is approachable only in 'dimensionless' systems, i.e. where the material particle is moving a distance not exceeding the Heisenberg length. One example of such system is electron transitions in atoms. It is straightforward to show that if the travel distance L is larger than a_H, the minimum travel time is increased as

$$\tau \sim \frac{L}{a_H} \cdot \tau_H \tag{11}$$

4.4 Energy Per Bit Operation

For a more accurate estimate of the lower boundary of energy per bit or switching energy, E_{sw}, again consider the device abstraction in Fig. 4a. The minimum energy needed to suppress the barrier (e.g. by charging gate capacitor) is equal to the barrier height E_b. To restore the barrier (e. g. by discharging gate capacitance) the expenditure of the minimum energy of E_b is also required. Thus the minimum energy of a full switching cycle is at least $2E_b$. Next, in order to enable rapid and reliable transition of an electron from state '0' to state '1', an asymmetry in energy between the two wells needs to be created (not shown in Fig. 4a). It is easy to show based on the distinguishability arguments that minimum energy difference ΔE_w between two wells is also given by Eq. 3b, and more generally $\Delta E_w = E_b$. If the number of electrons per switching transition is N_e, the total switching energy is

$$E_{SW} = 2E_b + N_e \Delta E_w = (N_e + 2)E_b \tag{12a}$$

For example, if $N = 1$ and $E_b = k_B T \ln 2$ then

$$E_{SW_{\min}} = 3k_B T \ln 2 \tag{12b}$$

Each device consists of three tiles and thus the switching energy per tile in a device is

$$\varepsilon_d \approx k_B T \tag{12c}$$

4.5 Interconnect Costs for Electronic Circuits

Devices must communicate to support computation, and there is an energy cost associated with communication. In charge based devices (e.g. CMOS) this implies that when the electron passes from state '0' to state '1' in one binary device (sending), it needs to control several downstream devices (receiving). The barriers (gates) of these receiving devices are electrically coupled to at least one well of the first device (Fig. 6). To model this, the length of the sending well is extended to accommodate the gates of receiving devices. In practice, this extension is achieved by interconnect systems (Fig. 6b). Note that the interconnect system can be represented as a combination of the square tiles as it is shown in Fig. 6b. It is straightforward to show from both topology and physics considerations that in the limiting case, the size of the *interconnect tile* is equal to the *device tile a*.

We now assess how reliably the charge in the extended well of the sending device **A** in Fig. 2a controls the receiving device **B** or **C**. Assume that one electron is needed to control the barrier of the receiving device, and one electron passes from one well to another in 0–1 switching ($N = 1$) of the sending device. After device **A** switched to '1' state, there is one electron on the right-hand well of **A**.

The electron can freely move along the line of length L and the probability to find this electron only in gate **B** or only in gate **C** is given by

$$\Pi_B = \Pi_C = \frac{a}{L} = \frac{1}{k} \tag{13}$$

(Note that $L/a = k$, the number of tiles needed to form an interconnect system of length L).

To increase the probability of successful control, the number of electrons, N_e, in the interconnect line needs to be increased and the resulting probability of placing an electron on gate B as:

$$\Pi_B = \Pi_C = 1 - \left(1 - \frac{a}{L}\right)^{N_e} \tag{14}$$

The solution of (11) for N_e is

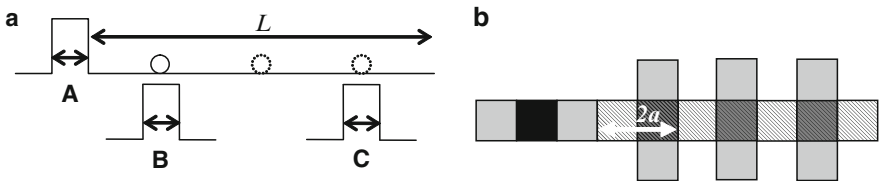


Fig. 6 Device abstraction of connected binary switches: (**a**) Barrier model, (**b**) Generic floorplan

Table 2 The number of electrons Ne for communication between two binary switches for probabilities of success 0.5 and 0.99

L/a = k	Ne (Π = 0.5)	Ne (Π = 0.99)
2	1	7
10	7	44
100	69	459
1000	693	4603

$$N_e = \log_{1-\frac{a}{L}}(1-\Pi) = \frac{\ln(1-\Pi)}{\ln\left(1-\frac{a}{L}\right)} = \log_{1-\frac{1}{k}}(1-\Pi) = \frac{\ln(1-\Pi)}{\ln\left(1-\frac{1}{k}\right)} \tag{15}$$

The number of electrons N_e needed for communication between two binary switches connected by a wire of length L is given in Table 2 for several probabilities of success. It can be seen from the Table 2 that communication at distance is a very costly process.

4.6 Fan-Out Costs

For logic operation, a binary switch needs to control at least two other binary switches. The probability that N_e electrons in the interconnect line of device **A** will be found on the gates of *both* **B** and **C** is:

$$\Pi_{B_{AND}C} = \Pi_B \cdot \Pi_C = \Pi_2 = \left(1-\left(1-\frac{a}{L}\right)^{N_e}\right)^2 \tag{16}$$

In general,

$$\Pi_F = \left(1-\left(1-\frac{a}{L}\right)^{N_e}\right)^F \tag{17}$$

where F is number of receiving devices, or fan-out.

Solving (17) for N_e obtain

$$N_e = \frac{\ln\left(1-\sqrt[F]{\Pi}\right)}{\ln\left(1-\frac{1}{k}\right)} \tag{18}$$

From simple geometrical considerations illustrated in Fig 4b, the minimum interconnect length in 2D topology is

$$L_{\min} = 2aF \tag{19}$$

Table 3 Fan-out costs

F	Ne ($\Pi = 0.5$)	Ne ($\Pi = 0.99$)
2	5	19
3	9	32
4	14	45

and, thus the a/L term in (15) becomes ½F for minimum interconnect lengths.

Table 3 presents the number of electrons needed to guarantee the specified reliability for the minimum interconnect length given by (19).

4.7 Energy Per Tile

According to (12a) the switching energy per connected switch in the Boltzmann's limit is

$$E_{SW} = (N_e + 2)k_B T \ln 2 \tag{20}$$

Using (18–20), one can calculate the switching energy per interconnect tile. Figure 7 displays the energy per tile for different interconnect lengths (measured in the number of tiles) for different fan-outs. It can be seen that in the limit of long interconnects, the energy per tile is constant and it varies from about 0.7 k_BT/tile to 1.8 k_BT/tile for F = 1–4.

Now, assuming that F is distributed between different fragments of a circuit in equal proportions, obtain average energy per interconnect tile to be:

$$\langle \varepsilon_i \rangle_{long} = 1.33 k_B T (k \to \infty) \tag{21a}$$

and

$$\langle \varepsilon_i \rangle_{short} = 1.18 k_B T \, (k \to k_{min}) \tag{21b}$$

Comparing (12c), (21a) and (21b), obtain:

$$\varepsilon_d \approx \langle \varepsilon_i \rangle \approx k_B T = \langle \varepsilon \rangle_{tile} \tag{21c}$$

i.e. in the limit, the average energy per functional tile of both devices and interconnect is approximately the same.

4.8 Digital Circuit Abstraction

An arbitrary circuit of interconnected binary switches can be presented as a 2D plane of densely packed device tiles (e.g. as in Fig. 5) and another 2D plane of

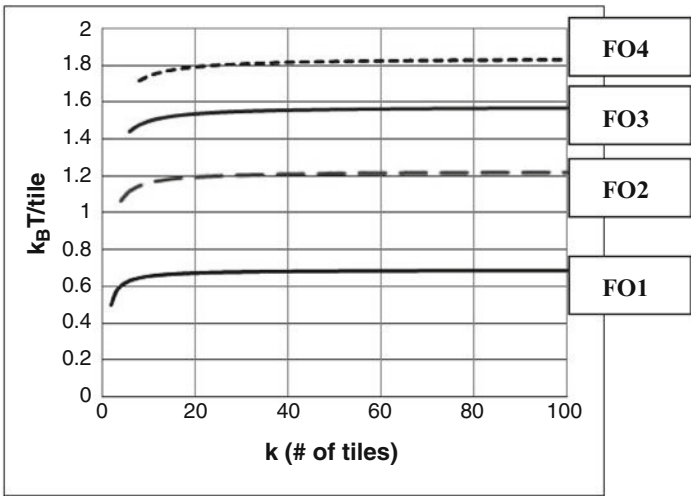


Fig. 7 Interconnect energy per tile

interconnect tiles. According to the above analysis, the circuit parameters such as switching energy and speed scale with the number of the functional tiles k:

$$E_{sw} = k \cdot \langle \varepsilon \rangle_{tile} \tag{22a}$$

$$t_{sw} = k \cdot \tau_H \tag{22b}$$

To estimate the minimum number of interconnect tiles per device, assume that for each of three tiles of the device, at least one contacting interconnect tile (3 total) is needed and one connecting interconnect tile (3 total) is needed. This results in six interconnect tiles per binary switch (average interconnect length $\langle L \rangle = 6a$), Thus the total number of tiles per device, $k = 3 + 6 = 9$, and from (22a) and (22b) obtain for a limiting case:

$$E_{sw} = 9k_BT = 3.73 \cdot 10^{-20} \frac{J}{device} \tag{23a}$$

$$t_{sw} = 9\tau_{HB} \approx 0.2ps \tag{23b}$$

Now, for a circuit of M binary switches obtain:

$$E_{sw}(M) = \frac{9}{2} M \cdot k_BT \tag{24a}$$

$$t_{sw}(M) = 9M \cdot \tau_{HB} \tag{24b}$$

(The factor of ½ in (24a) reflects the fact that not all n devices switch simultaneously. A 50% activity is assumed).

5 Energy/Performance Analysis of a Minimal Turing Machine

There is a growing need to increase the performance of microprocessors per unit of energy expended to better support mobile applications and to reduce the overall demand on energy systems by desktop applications. Ultimately, this leads to the question of the role of computer architecture in not only providing high performance but also in simultaneously reducing energy consumption. We know of no theoretical results that characterize the maximum computational efficiency of architectures, for example, in the spirit of the bound on efficiency of heat engines obtained by Carnot. In the following, we offer a framework for an approach to obtaining such a bound. Careful consideration must be given to defining the measure of "useful work" done by the processor since this ultimately is crucial to the definition of the energy-efficiency of the processor. One step in this direction is to investigate the system scaling limits to determine the minimum physical dimensions a universal automaton can occupy. Each automaton contains a certain number of discrete elements (e.g. transistors, resistors, diodes, etc.). The internal complexity of the system (i.e. the number of discrete elements $\boldsymbol{M}$) defines the system capability. As von Neumann put it [8], "*if one constructs the automaton (A) correctly, then any additional requirements about the automaton can be handled by sufficiently elaborated instructions. This is only true if A is sufficiently complicated, if it has reached a certain minimum of complexity*". In other words, a system cell must surpass a certain internal complexity threshold if it is to perform arbitrarily complex tasks by virtue of elaborate software instructions. We call the number $\boldsymbol{M_{min}} = \boldsymbol{\nu}$ the 'von Neumann threshold', which is the smallest complexity of the system to emulate general-purpose computing.

To estimate the von Neumann threshold, consider a 1-bit general purpose computer, which is referred to as the *Minimal Turing Machine* (MTM), and which contains an arithmetic-logic unit (ALU) and sufficient memory. Further, we analyze the operation of MTM based on the results from the digital circuit abstraction obtained in previous section.

5.1 Minimum Instruction Set 1-Bit Arithmetic Logic Unit

The 1-bit ALU (Fig. 8) input consists of two 1-bit operands **X** and **Y**, and an additional 1-bit *carry* input $\mathbf{C}_0$. The ALU has 1-bit result represented by the output **Z** and 1-bit carry output $\mathbf{C}_1$.

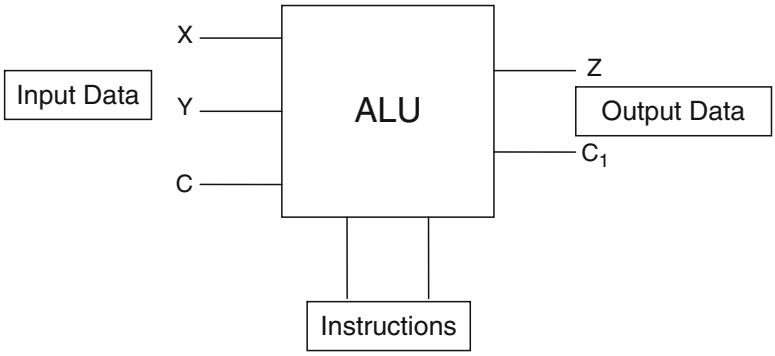


Fig. 8 1-bit arithmetic logic unit

The minimal ALU does only four operations on two 1-bit binary numbers **X** and **Y**:

Operation 1: **X** AND **Y**
Operation 2: **X** OR **Y**
Operation 3: (**X** + **Y**)
Operation 4: (**X** + (NOT **Y**))

This is a minimal set, and all other operations can be implemented by combinations of these four basic operations.

The command language consists of four words, which requires 2-bit word length ($w = \log_2 4$): 00, 01, 10, 11.

To implement these four operations, the ALU contains four independent logic circuits: AND, OR, and two full adders. In addition, there is a NOT gate (inverter) to implement NOT **X** in Operation 4. Also the carry outputs of the two full adders are connected to the C_1 terminal via an OR gate. Finally, a selector switch is needed to choose the value (out of four) on the output Z terminal according to the implemented instruction. Thus a 4-to-1 multiplexer is needed. Figure 9 shows the ALU schematics showing all functional units.

The element count [10, 11] for each block of the ALU is shown in Fig. 9. The basic ALU would require a total of about 98 elements (e.g. transistors).

5.2 Energetics of the Minimal ALU

The average energy per ALU operation can be estimated from (24a) obtain for $M = 98$:

$$E_{ALU} = \frac{9}{2} \cdot 98 \cdot k_B T \approx 440 k_B T \tag{25}$$

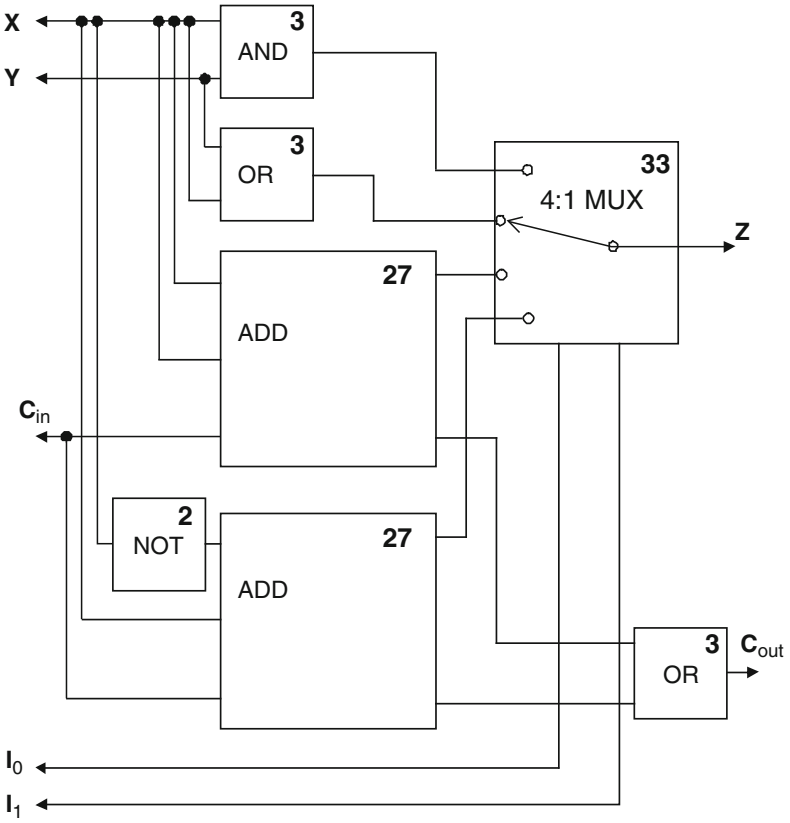


Fig. 9 1-bit ALU schematics showing all functional units and their element count

It is instructive to analyze the energy efficiency η of the minimal ALU, which is defined as

$$\eta = \frac{E_{op}}{E_{ALU}} \tag{26}$$

where E_{op} is the energy required to implement a given operation by a stand-alone circuit (e.g. AND gate).

For example, the AND operation requires a minimum $M = 3$ devices, and thus

$$E_{AND} = \frac{9}{2} \cdot 3 \cdot k_B T \approx 14 k_B T \tag{27a}$$

$$\eta_{AND} = \frac{E_{AND}}{E_{ALU}} = \frac{14}{440} \approx 3\% \tag{28b}$$

On the other hand, for an adder $M = 27$, thus $E_{ADD} \approx 122 k_B T$ and $\eta_{ADD} \approx 28\%$. These efficiencies are reminiscent of that of the Carnot cycle for heat engines.

The reason for the low ALU efficiency is that in each operation, all four units execute even though only one output is used. One might hypothesize that the overall efficiency of the ALU could be increased by utilizing a different architecture, such that the inputs are de-parallelized. One way to implement such an approach would be to only provide inputs to the instruction that is to be executed. In order to do this, two additional 4-to-1 de-multiplexers (1:4 DMUX) would be needed at the inputs to the ALU. Each DMUX circuit would require $M \sim 33$ devices. The energy efficiency of the AND/OR instruction would remain in the range of 3% since the DMUX circuits would require activation. Similarly, the efficiency of the ADD circuits would be in the range of 21% when DMUX overhead is considered. Hence this architectural variation does not seem to be much better from and energy efficiency point of view than the first architecture discussed and therefore it will be retained as the baseline in subsequent computations.

5.3 Minimal ALU Timing

The average delay in the minimal ALU can be estimated from (24b) with n now beeing the number of cascades between input and output by executing each operation (the critical path. The delay depends on the particular operation, with estimated range from about $45\tau_{HB}$ to $99\tau_{HB}$. In the following consideration an average delay s assumed.

$$\langle t_{ALU} \rangle \sim 100\tau_{HB} \sim 2ps$$

5.4 Minimal Turing Machine: Device Count

A functional machine requires significant surrounding infrastructure to support computation including latches for inputs and outputs, and mechanisms to retrieve and apply instructions. As shown in Fig. 10, the 1-bit CPU would require at least six external switches and five memory registers to manage the input and output data. These external components serve to increase the transistor count to approximately $M_{CPU} = 134$ and the energy dissipation to $\sim 600\ k_B T$/operation.

Finally, to complete the machine we need to add external memory (the Turing Machine tape) to run the instructions. The minimum memory needed to run one instruction can be estimated as follows: For each of the four ALU operations, three instruction cycles are needed (each cycle correspond to one step of the tape movement):

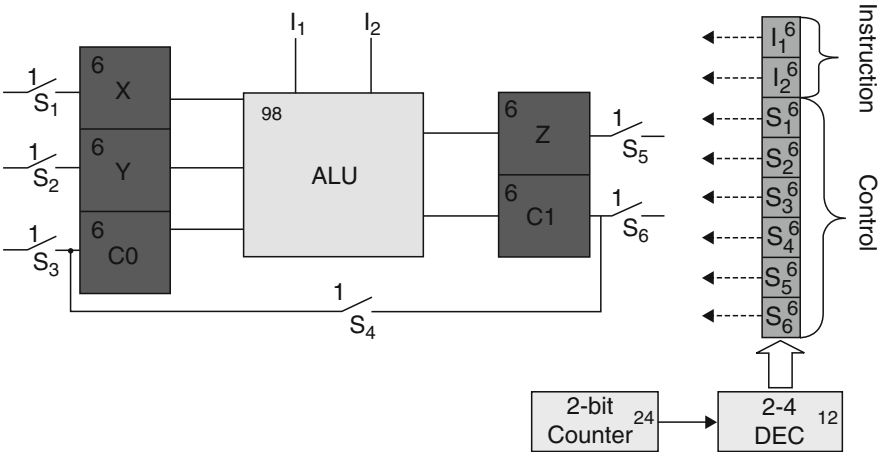


Fig. 10 Minimal turing machine showing all functional units and their element count

Cycle 1: Input
Cycle 2: Operation
Cycle 3: Output

Each of the three cycles is controlled by a 8-bit fragment of the tape (as shown in Fig. 10). Thus, 24 bits of memory are needed to complete an operation. Assuming six devices per memory bit (SRAM implementation), obtain 144 devices for the minimal external memory.

To implement the steps of tape movement, a program counter is needed. For just three steps of the movement, it would consist of 2-bit counter (24 devices) and a 2-to-4 decoder (12+ devices).

Summarizing, the total element count in the Minimal Turing Machine is:

$$M_{MTM} = 134(CPU) + 144(Tape) + 36(ProgramCounter) = 314$$

This number is consistent with von Neumann estimate, suggesting that the minimum circuit complexity required to implement general-purpose computing is of the order of a few hundred devices [9].

5.5 *Minimal Turing Machine: A Numerical Summary*

According to (7), the minimum area A, occupied by a circuit of M devices is:

$$A = 8Ma^2 \tag{29}$$

Thus, the area of MTM ($M = 314$) implemented with the limiting devices ($a = a_{HB} \sim 1$ nm) is:

$$A_{MTM} = 8 \cdot 314 \cdot a^2 \approx 2,500a^2 = 50a \times 50a \sim 50nm \times 50nm$$

The operational energy of one cycle of the MTM from (24a) is:

$$E_{MTM_1} = \frac{9}{2} \cdot 98 \cdot k_B T \approx 1400 k_B T \approx 6 \cdot 10^{-18} \frac{J}{cycle}$$

or, per full three-cycle operation,

$$E_{op} = 3E_{MTM_1} \approx 2 \cdot 10^{-17} \frac{J}{op}$$

The binary throughput β (Eq. 1) of the Minimal Turing Machine is:

$$\beta_{MTM} = \frac{M}{t_{op}} = \frac{314}{2ps} \approx 10^{14} \frac{bit}{s}$$

The computational performance μ (Eq. 2):

$$\mu_{MTM} = \frac{1}{3t_{op}} \approx 10^5 MIPS$$

6 Conclusions

The thesis of this paper is that as one seeks to define new device and interconnect systems that would either replace or supplement CMOS, it is important to evaluate these proposed new technologies in the context of their performance in an architectural framework. An energy barrier model is offered that is believed to be applicable for a broad class of devices and from which basic physics can be used to estimate device performance limits. The interconnect models used in this paper derive their inspiration from electron-based interconnect systems, and if other interconnect schemes are proposed, then performance limits similar to those offered in this paper would need to be derived to develop estimates for circuit-level performance. It is ultimately concluded that a good approximation for the energy utilized by the technology (both interconnects and devices) is about $k_B T$ per tile, where the tile dimensions are defined in terms of the Boltzmann-Heisenberg scaling limits for devices. The approach taken to connecting device and interconnect limits to the ultimate performance of a processor is to choose the simplest functional unit that

has sufficient device count (von Neumann complexity) to serve as a general purpose computational element (Turing Machine). The application of the derived fundamental performance limits of circuits to this architecture allows the estimation of the achievable performance of the processor as a function of technology capability. While it is shown that remarkable performance is attainable with technologies whose operational properties submit to the models derived in the paper, it does not appear that the energy efficiencies of which they are capable of matching those of biological systems.

References

1. Semiconductor Industry Association (SIA), *International Technology Roadmap for Semiconductors* (International SEMATECH, Austin, TX, 2007)
2. *The Intel Microprocessor Quick Reference Guide* and *TSCP Benchmark Scores*
3. W. Gitt, Information – the 3rd fundamental quantity. Siemens Rev **56**, 36 (1989)
4. H. Moravec, When will computer hardware match the human brain? J. Evol. Technol. **1**, 1 (1998)
5. V.V. Zhirnov, R.K. Cavin, J.A. Hutchby, G.I. Bourianoff, Limits to binary logic switch scaling – A Gedanken Model. Proc. IEEE **91**, 1934 (2003)
6. R.K. Cavin, V.V. Zhirnov, J.A. Hutchby, G.I. Bourianoff, Energy barriers, demons, and minimum energy operation of electronic devices. Fluctuation Noise Lett. **5**, C29 (2005)
7. R.K. Cavin, V.V. Zhirnov, D.J.C. Herr, A. Avila, J.A. Hutchby, Research directions and challenges in nanoelectronics. J. Nanopart. Res. **8**, 841 (2006)
8. J. von Neumann, *Theory of Self-Reproducing Automata* (University of Illinois Press, Urbana, IL, 1966)
9. J. von Neumann, *The Computer and the Brain* (Yale University Press, New Haven/London, 1959)
10. R.L. Geiger, P.E. Allen, N.R. Strader, *VLSI Design Techniques for Analog and Digital Circuits* (McGraw-Hill, New York, 1990)
11. J.M. Rabaey, *Digital Integrated Circuits* (Prentice-Hall, Upper Saddle River, NJ, 1996)
12. J.W. Joyner, Limits on device packing density as 2-D tiling problem, unpublished

A Simple Compact Model to Analyze the Impact of Ballistic and Quasi-Ballistic Transport on Ring Oscillator Performance

S. Martinie, D. Munteanu, G. Le Carval, and J.L. Autran

1 Introduction

In nanoscale MOSFETs with decananometer channel lengths, relaxation times of free carriers indicate that the drain current will have an intermediate character between drift-diffusion and ballistic/quasi-ballistic transport [1]. Then, ballistic and quasi-ballistic transport regimes have to be considered in the modeling of ultra-short Double-Gate devices with an accurate description. Since the conventional Drift-Diffusion model (usually considered as a standard simulation level for devices) fails at describing ballistic transport, new specific models have to be developed for this regime.

The highest value of the source-to-drain current which can be delivered by a given MOSFET architecture corresponds to the pure ballistic current limit. Carrier transport in the channel is considered to be ballistic when carriers travel from the source to the drain regions without encountering a scattering event. This may be possible if the feature size of the device becomes smaller than the carrier mean free path [2]. If the carrier transport is purely ballistic in the channel, modeling the device behavior reduces to the description of the carrier transmission over and through the source-to-drain potential barrier. The source-to-drain current is usually given by the difference of the

S. Martinie (✉)
CEA-LETI MINATEC, France
IM2NP-CNRS, France
e-mail: jean-luc.autran@univ-provence.fr

D. Munteanu
IM2NP-CNRS, France

G. Le Carval,
CEA-LETI MINATEC, France

J.L. Autran
IM2NP-CNRS, France
Institut Universitaire de France (IUF), Paris, France

A. Amara et al. (eds.), *Emerging Technologies and Circuits*,
Lecture Notes in Electrical Engineering 66, DOI 10.1007/978-90-481-9379-0_3,

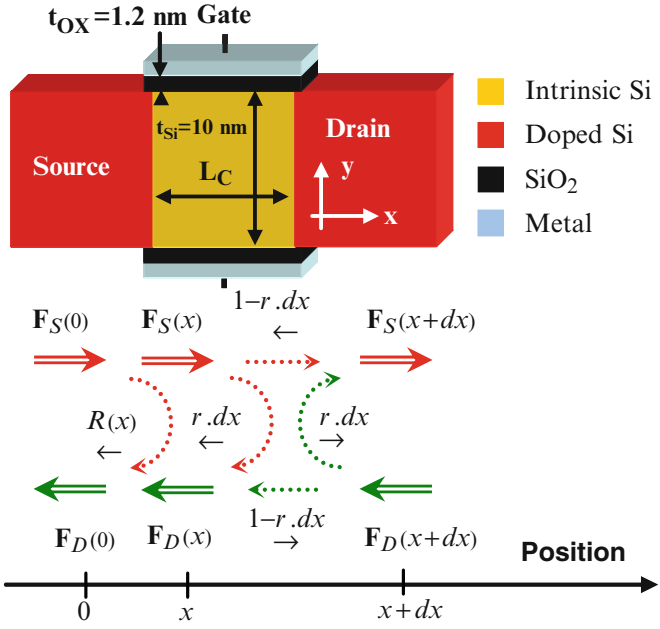


Fig. 1 Schematics of the DGMOS device used in this work and definition of the main geometrical, electrical and flux parameters. R and r are respectively the backscattering coefficient and the scattering probabilities in the presence of an electric field. The symbols "→" and "←" of scattering probabilities represent the carrier's velocity components that are parallel or anti-parallel, resp., to the electric field direction

fluxes injected from the source and from the drain at the location (i.e. the abscissa) of the maximum of the energy barrier [3, 4].

When the channel length is increased, the current decreases from its maximum value due to scattering effects. Above a certain channel length, backscattering may affect the transport, which therefore cannot be considered as purely ballistic anymore. In other terms, carrier transport makes a transition from the ballistic to quasi-ballistic or drift-diffusive regime with the longer channel lengths. Reflections of carriers after crossing the barrier may take place within a certain length. In this case, the transport is quasi-ballistic and these reflections must be included in the modeling of the ballistic current [5].

Future Double-Gate (DGMOS) MOSFETs (Fig.1), designed with channel lengths in the decananometer scale, are expected to be more ballistic than diffusive. Then, it is fundamental to understand the physics of the ballistic transport and to develop compact models that assure the transition from drift-diffusion to the ballistic regime in a unique description. A few compact models of ballistic transport have been published for nanoscale DG MOSFETs [3–7]. However, very little work

has been done on the transition from the ballistic regime to drift diffusion [8–10]. Several analytical models based on the Drift-Diffusion formalism demonstrate that is possible to introduce the diffusive transport in compact modeling. However, when the channel length approaches the mean free path, the mobility definition cannot more strictly explain the electronic transport occurring in the device. In this case the one-flux theory (also called "McKelvey theory" [11, 12]) should be considered; this approach can be applied to systems where the channel length is of the order of or smaller than the mean free path [13].

The pioneering work of Lundstrom et al. [14, 15] on the scattering theory has shown the strong impact of the ballistic transport on the MOSFETs operation and especially on the expression of the drain current. Lundstrom et al. developed physically-based compact models for DGMOS based on the one-flux theory and describing the device operation in ballistic or quasi-ballistic regimes. The main parameter of this approach is the backscattering coefficient, which expresses the ballistic and the quasi-ballistic transport. These works [14, 15] also demonstrate the usefulness of the one-flux theory in qualitatively describing quasi-ballistic transport in compact modeling.

The evaluation of the ballistic transport impact on the device performance is currently of great interest and solutions are envisaged to enhance the amount of ballistic transport in device operation. However, in order to confirm the necessity of these works on ballistic transport, it is nowadays essential to directly evaluate the impact of ballistic and quasi-ballistic transport at the circuit level through simulation of several circuit demonstrators. The implementation of compact models in Verilog-A environment offers the opportunity to describe as accurately as possible the physics of transport and to analyze its impact on various circuit elements.

In this work, we demonstrate the feasibility of a simulation study of ballistic/quasi-ballistic transport at circuit level and we highlight the impact of this advanced transport on CMOS inverter commutation and oscillation frequency of ring oscillators. Our model is based on the Lundstrom's approach [15] and considers a new expression of the backscattering coefficient given by the flux method. The model takes also into account short channel effects and takes into account the effects of different scattering processes through a dynamical mean free path. Using this model, CMOS inverters and ring oscillators have been simulated to highlight the impact of ballistic and quasi-ballistic transport on static and transient performance. This chapter is organized as follows: in the first part we explain the origin of backscattering coefficient and we introduce the drain current model for quasi-ballistic transport. In the second part, the relation between physics of quasi-ballistic transport and its impact on circuit performance is thoroughly analyzed. For this purpose several small circuits (CMOS inverters and ring oscillators) are simulated using the quasi-ballistic transport model (developed in the first part) which is implemented in Verilog-A environment. The commutation time and the oscillation frequency of circuits considering the quasi-ballistic regime are compared to the conventional case of circuits working in the drift-diffusion regime.

2 Analytical Model

2.1 Backscattering View

The starting point of the model development is the McKelvey's [11, 12] flux method, illustrated on Fig. 1. The two fluxes $\mathbf{F}_{S(x)}$ and $\mathbf{F}_{D(x+dx)}$ incident on a semiconductor slab with thickness dx, transmit or reflect with the backscattering probabilities per length r, contributing to the outward fluxes $\mathbf{F}_{S(x+dx)}$ and $\mathbf{F}_{D(x)}$; which can be described by the following equations:

$$\mathbf{F}_S(x+dx) = \mathbf{F}_S(x).(1 - \underset{\leftarrow}{r}.dx) + \mathbf{F}_D(x+dx).(\underset{\rightarrow}{r}.dx) \tag{1}$$

$$\mathbf{F}_D(x)/\mathbf{F}_S(x) = \underset{\leftarrow}{R}(x) \tag{2}$$

$$\mathbf{F}_D(x) = \mathbf{F}_D(x+dx).(1 - \underset{\rightarrow}{r}.dx) + \mathbf{F}_S(x).(\underset{\leftarrow}{r}.dx) \tag{3}$$

Developing this system of equation and simplifying at the first order we obtain [4]:

$$\frac{d\mathbf{F}_D(x)}{dx} = \underset{\rightarrow}{r}.\mathbf{F}_D(x) - \underset{\leftarrow}{r}.\mathbf{F}_S(x) \tag{4}$$

$$\frac{d\mathbf{F}_S(x)}{dx} = \underset{\rightarrow}{r}.\mathbf{F}_D(x) - \underset{\leftarrow}{r}.\mathbf{F}_S(x) \tag{5}$$

Adding Eqs. 4 and 5 and considering $\mathbf{F} = \mathbf{F}_D - \mathbf{F}_S$ and $n = (\mathbf{F}_D + \mathbf{F}_S)/v_{th}$ (where v_{th} is the thermal velocity) we obtain:

$$\mathbf{F} = \left(\frac{\underset{\leftarrow}{r} - \underset{\rightarrow}{r}}{\underset{\rightarrow}{r} + \underset{\leftarrow}{r}}\right).v_{th}.n + \frac{v_{th}}{\underset{\rightarrow}{r} + \underset{\leftarrow}{r}}.\frac{dn}{dx} \tag{6}$$

By analogy to the classical drift-diffusion approach, we obtain the Einstein relation:

$$\frac{D_n}{\mu_n} = \frac{k.T}{q} \Rightarrow \frac{E}{\underset{\leftarrow}{r} - \underset{\rightarrow}{r}} \equiv \frac{k.T}{q} \tag{7}$$

where D_n and μ_n are respectively the diffusion coefficient and the mobility. We use here the expression of scattering probabilities proposed in [12] (with the assumption of non-degenerate gas):

$$\underset{\leftarrow}{r} = \lambda^{-1};\ \underset{\rightarrow}{r} = \lambda^{-1} - \frac{q.E}{k.T} \tag{8}$$

where λ is the mean free path, k is the Boltzmann constant, q is the electron charge, T is the lattice temperature and E is the electric field. Supposing that the electric field is constant and applying the same initial condition expose in [11], the form of backscattering coefficient become:

$$\underset{\leftarrow}{R} = \frac{\underset{\leftarrow}{r}}{\sqrt{\underset{\leftrightarrow}{r}^2 - \underset{\rightarrow}{r} \cdot \underset{\leftarrow}{r}}.\mathbf{coth}\left(x.\sqrt{\underset{\leftrightarrow}{r}^2 - \underset{\rightarrow}{r} \cdot \underset{\leftarrow}{r}}\right) + \underset{\leftrightarrow}{r}}; \quad \underset{\leftrightarrow}{r} = \frac{\underset{\rightarrow}{r} + \underset{\leftarrow}{r}}{2} \tag{9}$$

In the assumption of a linear potential profile in the channel where $E = V_{DS}/L$:

$$\underset{\leftarrow}{R} = \frac{\lambda^{-1}}{\frac{1}{2}.L_{kT}^{-1}.\left(1 + \mathbf{coth}\left(\frac{x}{2}.L_{kT}^{-1}\right)\right) + \lambda^{-1}} \tag{10a}$$

where L_{kT} is the distance over which the channel potential drops by kT/q compared to the peak value of the source to channel barrier:

$$L_{kT} = L.\frac{k.T}{q.V_{DS}} \tag{10b}$$

Physically, L_{kT} represents the critical distance over which scattering events modify the current; this characteristic length depends on both source-to-drain drop voltage and gate length [14]. We obtain the classical form of the backscattering coefficient [14] in linear (low V_{DS}) and saturated (high V_{DS}) region respectively:

$$R = \frac{L}{L + \lambda} \tag{11a}$$

$$R = \frac{L_{kT}}{L_{kT} + \lambda} \tag{11b}$$

As explained in [2], the analytical formulation of the "kT-layer" given by (10a) cannot correctly take into account all scattering effects due to the dependence on both the bias condition and mean free path [16]. In fact this problem is recurrent in compact modelling and comes from the self-consistently solving of the Poisson and transport equations which is not possible to integrate in an analytical modelling approach.

As expected, Fig. 2 illustrates that the backscattering coefficient decreases when the V_{DS} increases with a direct impact on the current value, as detailed in paragraph 3. However, Eq. 10 tends to the value $R = L/(\lambda + L)$ proposed in [17–19] in the absence of electric field (as illustrated in the inset of Fig. 2). Finally, we note that Eq. 10 is

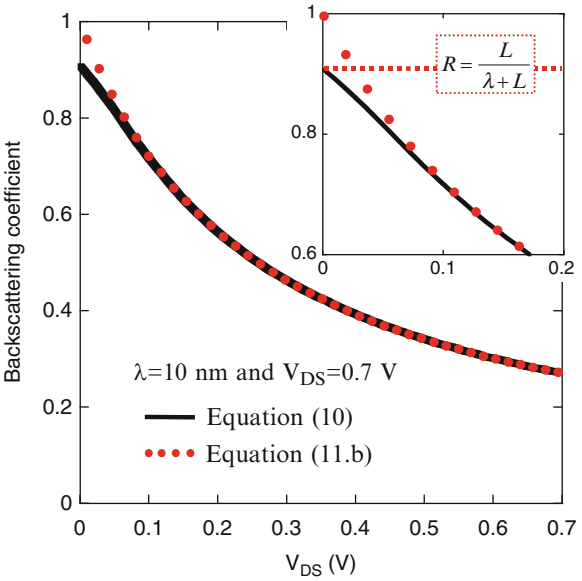


Fig. 2 Backscattering coefficient ($\lambda = 10$ nm) as a function of the channel length with VDS = 0.7 V for L = 100 nm

valid from low to high V_{DS} value without any interpolation function or fitting parameter between low and high field condition, as those proposed in [6].

2.2 Drain Current Modeling

In the assumption of a non-degenerate gas, we use the classical expression of the drain current given by the Lundstrom's approach [14] (based on the Natori's formulation [21]). This expression is obtained by multiplying the injection velocity and the charge determined classically from the oxide capacitance and the threshold voltage and taking into account the drain injection at low bias. This formula describes the ballistic and quasi-ballistic current through the backscattering formula presented in paragraph 2.1:

$$I_D = W.C_{ox}.(V_{GS} - V_T).v_{th}.\left(\frac{1-R}{1-R}\right).\left(\frac{1-e^{-qV_{DS}/k.T}}{1-\left(\frac{1-R}{1-R}\right)e^{-qV_{DS}/k.T}}\right) \quad (12)$$

where W is the gate width, C_{ox} is the gate oxide capacitance, V_{GS} is the gate to source voltage, V_{DS} is the drain to source voltage and V_T is the threshold voltage. To obtain an accurate model and describe all electrostatic effects, we have also

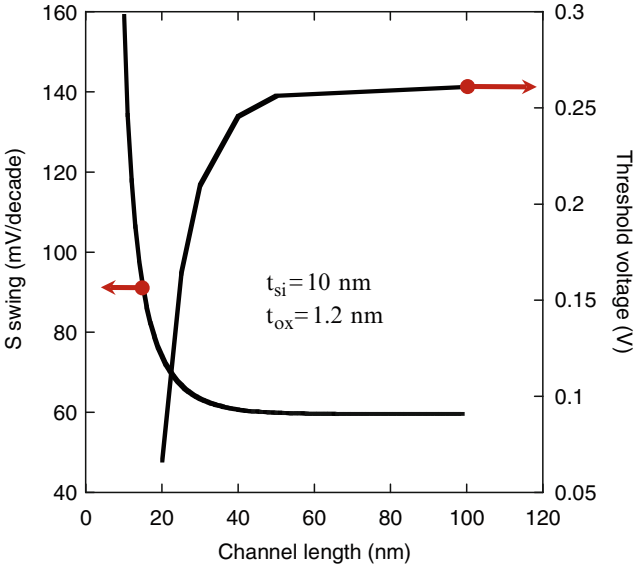


Fig. 3 Threshold voltage and the S swing parameter as a function of the channel length for tsi = 10 nm and tox = 1.2 nm

introduced SCEs using the Suzuki's model [22], respectively. Thus, V_T in Eq. 12 is modified by ΔV_T:

$$V_T = V_{th} - \Delta V_T \tag{13}$$

where V_{th} and ΔV_T are the long channel threshold voltage and its variation due to SCEs in the Suzuki's model [22]. We find the analytical expression of the surface potential $\Psi_S(x)$ by considering a parabolic dependence for the potential in the perpendicular direction to the Si/SiO_2 interface (y dependence on the Fig. 1). The development also considers the boundary conditions at the front and back interfaces and takes into account that the potential dependence in the y direction is symmetrical with respect to the middle of the silicon film. With these assumptions, an analytical form of surface potential is obtained from the solving of Poisson equation which takes into account only the depletion charge (the inversion charge is neglected at threshold). Finally, the analytical models [20] of ΔV_T and the swing S parameter (describing the short channel effects) are determined using the minimum value of the surface potential $\Psi_S(x)$ along the Si/SiO_2 interface.

As expected, the Fig. 3 illustrates the strong influence of short channel effects on V_T and S: when reducing the gate length the threshold voltage is reduced (which increases the I_{on} current), and the I_{off} current and the subthreshold swing are increased.

The above-threshold regime is linked to the subthreshold regime using an interpolation function based on the subthreshold swing S parameter also defined by Suzuki in [22]. This assures the perfect continuity of our model between on-state current (I_{on}) and off-state current.

Finally, we define a "dynamical mean free path" (dfp) [23] to include the scattering process with impurities (τ_{imp}) and phonon interactions (τ_{ph}). τ_{imp} and τ_{ph} are calculated as in [23]. This expression replaces λ in (10) to describe the quasi-ballistic transport:

$$dfp = v_{bal}.\tau_{tot}; \begin{cases} \tau_{tot}^{-1} = \tau_{imp}^{-1} + \tau_{ph}^{-1} \\ v_{bal} = \sqrt{\dfrac{2.\varepsilon_{bal}}{m*}}; \varepsilon_{bal} = \dfrac{3}{2}.k_B.T_L + q.V_{DS} \end{cases} \tag{14}$$

where m^* is the mass in direction of transport, v_{bal} the ballistic velocity, τ_{tot} the total scattering rate and ε_{bal} the carrier energy.

3 Simulation and Discussion

3.1 Ballistic and Quasi-Ballistic Transport

After implementation in Verilog-A environment, the model has been used to simulate the n-channel DGMOS structure schematically presented in Fig. 1. The source and drain regions are heavily doped (1×10^{20} cm^{-3}) and an intrinsic thin silicon channel is considered. The channel length varies from 10 to 200 nm; a gate oxide of 1.2 nm thick and a midgap metal gate are also considered.

It is well-known that the ballistic current is independent of the channel length [21] except when SCEs or Drain Induced Barrier Lowering (DIBL) effect appears. In order to clearly confirm this point, simulations have been performed for several length (20, 25, 30, 40, 50, 100 and 200 nm) and considering two types of transport (quasi-ballistic and ballistic; Fig. 4a). Note that for the ballistic case, the mean free path value has been chosen to be extremely large compared to the channel length. In contrast to the ballistic case, the quasi-ballistic transport has the same behaviour as that of diffusive transport and the form of the output characteristics depends on L_c.

Figure 4b shows the drain current versus the gate voltage characteristics for the simulated devices at $V_{DS} = 0.7$V. As expected, the ballistic and quasi-ballistic current shows a perfect continuity between the above and the subthreshold regime.

3.2 Circuit Level Simulations

In addition to the simulation of single device operation, we have simulated different small-circuit element as CMOS inverter and ring oscillator (Fig. 5) to show the impact of ballistic and quasi-ballistic transport at the circuit level. In this approach,

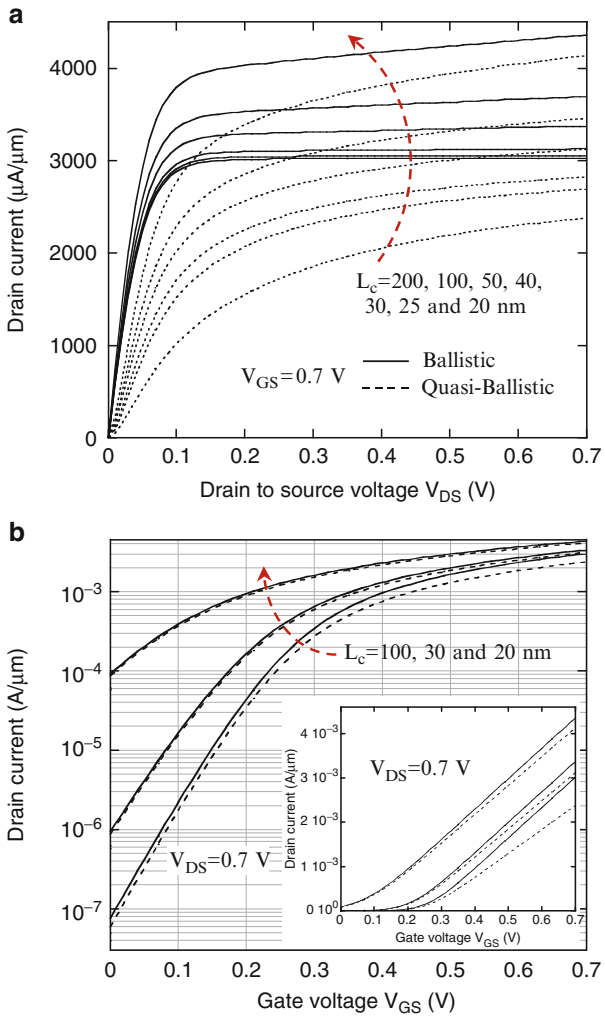
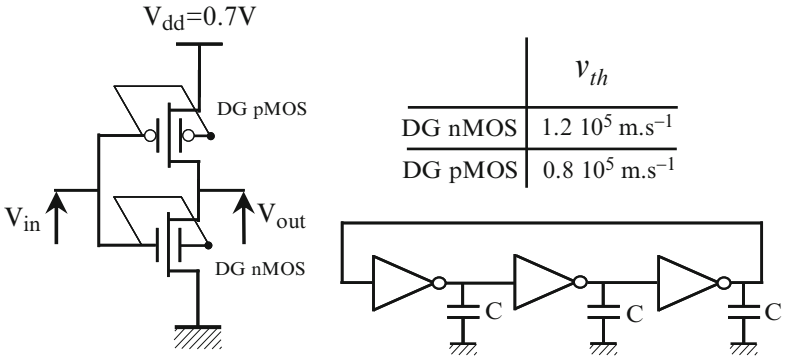


Fig. 4 Drain current versus VDS (**a**) for Lc = 200, 100, 50, 40, 30, 25 and 20 nm and VGS (**b**) for Lc = 100, 30 and 20 nm. Solid line for ballistic transport and dashed line for quasi-ballistic transport

we suppose that the transport description (for ballistic and quasi-ballistic case) for holes is identical to that of electrons, with uniquely changing the thermal velocity value in non-degenerate conditions [24].

The output voltage (V_{out}) of the CMOS inverter switches more sharply from the "1" state to the "0" state in the ballistic case than in quasi-ballistic transport (Fig. 6), and this is independent of the current level. In fact the commutation of the CMOS inverter depends on the limit between the linear and the saturation regions, which controls the switch between transistors. When SCEs occur, the transition



	v_{th}
DG nMOS	1.2 10^5 m.s^{-1}
DG pMOS	0.8 10^5 m.s^{-1}

Fig. 5 Schematics of the CMOS inverter circuit and three stage ring oscillator with a charge capacitance C

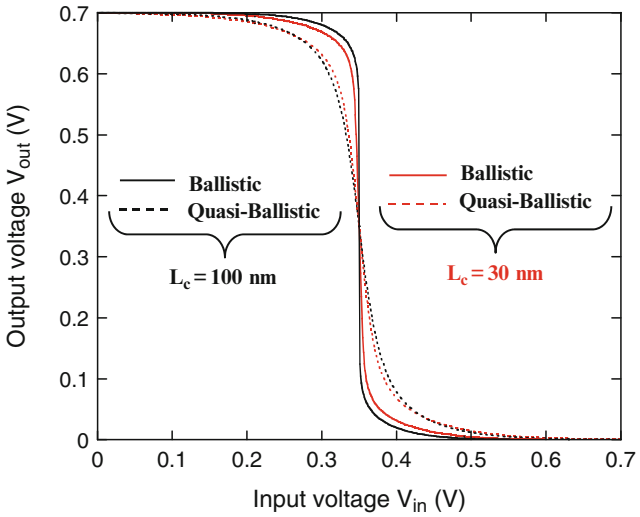


Fig. 6 Vout versus Vin in a CMOS inverter for Lc = 100 nm and 30 nm in the ballistic and quasi-ballistic case

between linear and saturate regime is modified, and the switch from the "1" state to the "0" state is less sharp. In the quasi-ballistic case, the abruptness of the CMOS characteristic is strongly deteriorated. These results prove that the ballistic transport improves the commutation and the static performances of the CMOS inverter.

Figure 7 shows the oscillation frequency as a function of the charge capacitance for two channel lengths: 100 and 30 nm. In a first time, we remark the reduction of the oscillation frequency when the charge capacitance increases, due to variation of

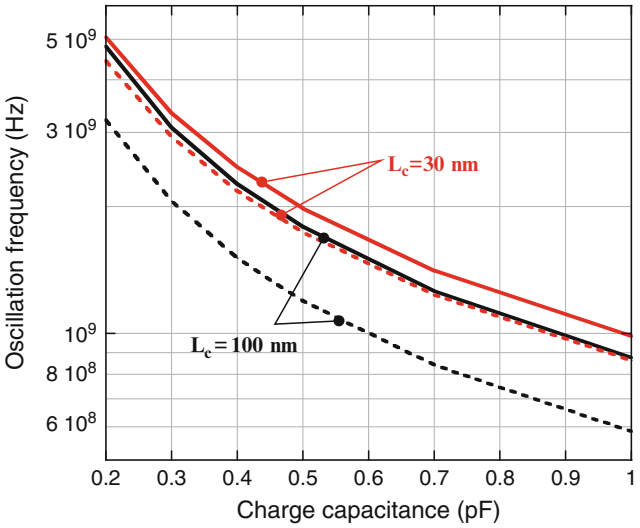


Fig. 7 Oscillation frequency versus charge capacitance for Lc = 100 nm and 30 nm in the ballistic and quasi-ballistic transport. Solid line for ballistic transport and dashed line for quasi-ballistic transport

the propagation time through the inverters. In a second time, we note the strong influence of SCEs that increase the current value and reduce the difference between the oscillation frequency in quasi-ballistic and ballistic transport. These results show that the oscillation frequency is directly influenced by the type of transport and the electrostatic conditions.

3.3 Impact of Quasi-Ballistic Transport on Circuit Performances

To clearly highlight the impact of quasi-ballistic transport and to analyse the qualitative relation between the mean free path and the oscillation frequency we performed simulations a fixed value of the *dfp*: 40, 30, 20 and 10 nm. Moreover, to focus on the influence of carrier transport, we also perform simulations in the particular case where I_{on} is artificially maintained constant at the value obtained in the ballistic case (Fig. 8). In this way, we can only analyze the effect of quasi-ballistic transport on ring oscillator frequency, independently from the I_{on} increase when λ increases.

Figure 9 shows the frequency oscillation of the ring oscillator for $L_c = 100$ and 20 nm for different mean free path values with or without the same I_{on}. The oscillation frequency decreases with the reduction in the mean free path for both cases. In the case of a variable I_{on}, the variation in the oscillation frequency with λ is more important than the case of a constant I_{on}. This is due to both the occurrence of

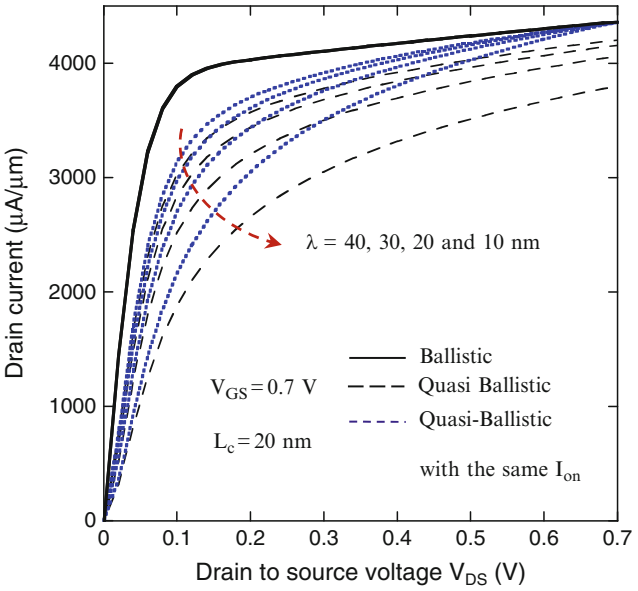


Fig. 8 Drain current versus VDS for Lc = 20 nm with and without same Ion. Solid line for ballistic transport and dashed line for quasi-ballistic transport

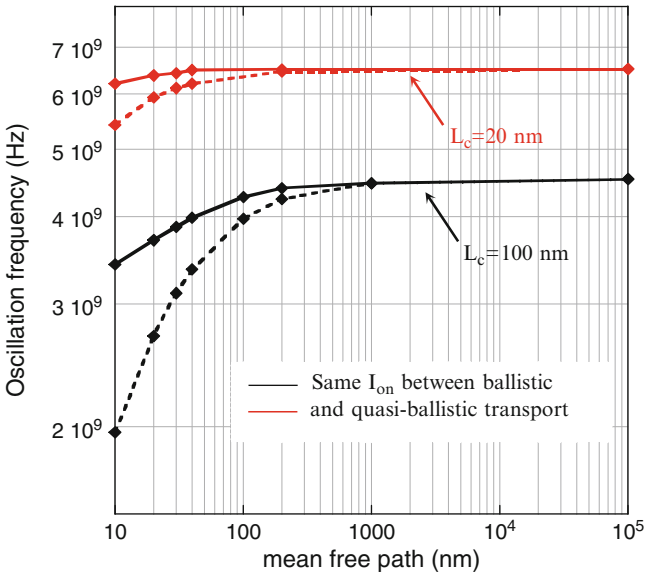


Fig. 9 Oscillation frequency versus mean free path for Lc = 100 and 20 nm in the quasi-ballistic case with or without same Ion

quasi-ballistic transport (for $\lambda < 1{,}000$ nm) and to the variation of I_{on}. When I_{on} is maintained constant (i.e. the effect of the I_{on} increase is removed), the influence of quasi-ballistic transport on the oscillation frequency is clearly highlighted (Fig. 9).

The oscillation frequency becomes maximal when the $\lambda >> L_c$, that corresponds to a ballistic transport. These results confirm that the physics of transport has a marked impact on the circuit performances, as already stated in the previous paragraph.

4 Conclusion

In this work, a compact model for DGMOS taking into account ballistic and quasi-ballistic transport has been proposed and implemented in Verilog-A environment. Short channel effects and model continuity (via the introduction of an interpolation function to link the above and the subthreshold voltage) have been included to obtain a complete description of the drain current. The dynamical mean free path definition was considered to describe scattering processes with impurities and phonons. Finally, the model has been used to simulate two different small-circuits (CMOS inverter and ring oscillators) and show the significant impact of ballistic/quasi-ballistic transport on the commutation of CMOS inverter and oscillation frequency of ring oscillator. Our simulation results prove that the ballistic transport improves the commutation and the static performances of the CMOS inverter, and increases the oscillation frequency of ring oscillators. This work demonstrates the feasibility of a simulation study of ballistic/quasi-ballistic transport at circuit level and highlights the direct relation between the type of transport and static or transient performances of small-circuits.

However, including the access resistance in the compact model will probably change our conclusions, especially for the simulation of ring oscillators where the oscillation frequency is strongly impacted by the value of the I_{on} current. Moreover this model can be optimized [23–26] to include quantum mechanical effect, other scattering mechanisms and a better description of the "kT-layer". Finally, our model can be successfully extended to different architectures as Silicon nanowire devices, where the assumptions of the one-flux theory are even more justified because the transport is purely one-dimensional.

References

1. B. Iñiguez, T.A. Fjeldly, A. Lázaro, F. Danneville, M.J. Deen, Compact-modeling solutions for nanoscale double-gate and gate-all-around MOSFETs. IEEE Trans. Electron Devices **53** (9), 2128–2142 (Sept 2006)
2. M. Lundstrom, *Fundamentals of Carrier Transport*, 2nd edn. (Cambridge University Press, Cambridge, 2000)

3. D. Jiménez, J.J. Sáenz, B. Iñíguez, J. Suñé, L.F. Marsal, J. Pallarès, A unified compact model for the ballistic quantum wire and quantum well MOSFET. J. Appl. Phys. **94**(2), 1061–1068 (Jul 15, 2003)
4. D. Jiménez, J.J. Sáenz, B. Iñíguez, J. Suñé, L.F. Marsal, J. Pallarès, Modeling of nanoscale gate-all-around MOSFETs. IEEE Electron Device Lett **25**(5), 314–316 (May 2004)
5. A. Rahman, M.S. Lundstrom, A compact scattering model for the nanoscale double gate MOSFET. IEEE Trans. Electron Devices **49**(3), 481–489 (Mar 2002)
6. H.A. Hamid, B. Iñíguez, D. Jiménez, L.F. Marsal, J. Pallarès, A simple model of the nanoscale double gate MOSFET based on the flux method. Phys Stat Sol. C **2**(8), 3086–3089 (May 2005)
7. J.L. Autran, D. Munteanu, O. Tintori, E. Decarre, A.M. Ionescu, An analytical subthreshold current model for ballistic quantum-wire double gate MOS transistors. Mol. Simul. **31**(2/3), 179–183 (Feb 15, 2005)
8. G. Mugnaini, G. Iannaccone, Physics-based compact model of nanoscale MOSFETs—Part II: Effects of degeneracy on transport. IEEE Trans. Electron Devices **52**(8), 1802–1806 (Aug 2005)
9. G. Mugnaini, G. Iannaccone, Physics-based compact model of nanoscale MOSFETs – Part I: Transition from drift-diffusion to ballistic transport. IEEE Trans. Electron Devices **52**(8), 1795–1801 (Aug 2005)
10. G. Mugnaini, G. Iannaccone, Analytical model for nanowire and nanotube transistors covering both dissipative and ballistic transport, in *Proceedings of the European Solid-State Device Research Conference (ESSDERC)*, Grenoble, France, Sept 2005, pp. 213–216
11. J.P. McKelvey, J.C. Balogh, Flux method for the analysis of transport problems in semiconductors in the presence of electric fields. Phys. Rev. **137**(A5), A1555–A1561 (Mar 1965)
12. J.-H. Rhew, M.S. Lundstrom, Drift-diffusion equation for ballistic transport in nanoscale metal-oxide-semiconductor field effects transistor. J. Appl. Phys. **92**(9), 5196–5202 (Nov 2002)
13. J.P. McKelvey, R.L. Longini, T.R. Brody, Alternative approach to the solution of added carrier transport problems in semiconductors. Phys. Rev. **123**(1), 2736–2743 (July 1961)
14. M. Lundstrom, Z. Ren, Essential physics of carrier transport in nanoscale MOSFETs. IEEE Trans. Electron Devices **49**(1), 131–141 (Jan 2002)
15. M. Lundstrom, Elementary scattering theory of the Si MOSFET. IEEE Trans. Electron Device Lett **18**(7), 361–363 (Jul 1997)
16. S. Martinie, G. Le Carval, D. Munteanu, S. Soliveres, J.L. Autran, Impact of ballistic and quasi-ballistic transport on performances of Double-Gate MOSFET-based circuits. IEEE Trans. Electron Devices **55**(9), 2443–2453 (Sept 2008)
17. S. Martinie, D. Munteanu, G. Le Carval, J.L. Autran, New unified analytical model of backscattering coefficient from low to high field conditions in quasi-ballistic transport, IEEE Electron Device Lett., in press 29(12), 1392–1394 (Dec 2008)
18. S. Datta, *Electronic transport in mesoscopic system* (Cambridge University Press, Cambridge, 1997)
19. S. Martinie, D. Munteanu, G. Le Carval, J.L. Autran, A simple compact model to analyze the impact of ballistic and quasi-ballistic transport on ring oscillator performance, in *Proceedings of the Integrated Circuit Design and Technology conference (IcIcDt)*, Grenoble, France, June 2008, pp. 273–276
20. S. Martinie, D. Munteanu, G. Le Carval, J.L. Autran, A new unified compact model for quasi-ballistic transport: Application to the analysis of circuit performances of a double-gate architecture, in *Proceedings of the Simulation of Semiconductor Process and Devices Conference (SISPAD)*, Hakone, Japan, Sept 2008, pp. 377–380
21. K. Natori, Ballistic metal-oxide-semiconductor field effect transistor. J. Appl. Phys. **76**(8), 4879–4890 (Oct 1994)
22. K. Suzuki, Y. Tosaka, T. Sugii, Analytical threshold voltage for short channel n+-p+ double-gate SOI MOSFETs. IEEE Trans. Electron Devices **43**(5), 732–738 (May 1996)

23. E. Fuchs, P. Dollfus, S. Barraud, D. Villanueva, F. Salvetti, T. Skotniki, A new bascattering model giving a description of the quasi-ballistic transport in Nano-MOSFET. IEEE Trans. Electron Devices **52**(10), 2280–2289 (Oct 2005)
24. F. Assad, Z. Ren, S. Datta, M. Lundstrom, P. Bendix, Performance limits of silicon MOSFET's, *IEEE IEDM Tech. Dig*, 1999.
25. D. Munteanu, J.L. Autran, S. Harrison, K. Nehari, O. Tintori, T. Skotniki, Compact model of the quantum short-channel threshold voltage in symmetric Double-Gate MOSFET. Mol. Simul. **31**(12), 831–837 (Oct 2005)
26. V. Barral, T. Poiroux, F. Andrieu, C. Buj-Dufournet, O. Faynot, T. Ernst, L. Brevard, C. Fenouillet-Beranger, D. Lafond, J.M. Hartmann, V. Vidal, F. Allain, N. Daval, I. Cayrefourcq, L. Tosti, D. Munteanu, J.L. Autran, S. Deleonibus, Strained FDSOI CMOS technology scalability down to 2.5nm film thickness and 18nm gate length with a TiN/HfO_2 gate stack. IEDM Tech. Dig., pp. 61, 2007

Part III
Advanced Devices and Circuits

Low-Voltage Scaled 6T FinFET SRAM Cells

N. Collaert, K. von Arnim, R. Rooyackers, T. Vandeweyer, A. Mercha, B. Parvais, L. Witters, A. Nackaerts, E. Altamirano Sanchez, M. Demand, A. Hikavyy, S. Demuynck, K. Devriendt, F. Bauer, I. Ferain, A. Veloso, K. De Meyer, S. Biesemans, and M. Jurczak

1 Introduction

Planar bulk devices suffer from high random dopant fluctuations (RDF) when scaled down to sub-32 nm technology nodes. This is considered as a major roadblock for the integration of these devices in high density 6T SRAM cells [1, 2]. The increasing variation of transistor parameters like V_T, I_{ON}, I_{OFF}, etc., can result in a large variability in performance and power. The possibility of leaving the channels undoped and their excellent immunity against Short Channel Effects (SCE) favors the use of FinFET-based multi-gate devices [3] for these technology nodes.

In [4] the potential of FinFET for Large Scale Integration (LSI) was already shown using relaxed device dimensions. A ring oscillator delay of 13.9 ps/stage was reported at $V_{DD} = 1$ V and $I_{OFF} = 1.9$ nA/stage. It clearly demonstrated that FinFET circuits are especially interesting for low-power and low-voltage applications. However, in [4] the transistor dimensions were still quite relaxed. When aggressively scaled down, the variation in fin width rather than RDF can become a major

N. Collaert (✉), R. Rooyackers, T. Vandeweyer, A. Mercha, B. Parvais, L. Witters, E. Altamirano Sanchez, M. Demand, A. Hikavyy, S. Demuynck, K. Devriendt,
A. Veloso, S. Biesemans, and M. Jurczak
IMEC, Leuven, Belgium
e-mail: collaert@imec.be

K. von Arnim and F. Bauer
Infineon Technologies AG, Neubiberg, Germany

A. Nackaerts
NXP-TSMC Research Center, Heverlee, Belgium

I. Ferain and
K. De Meyer
K.U.Leuven, ESAT-INSYS, Heverlee, Belgium

A. Amara et al. (eds.), *Emerging Technologies and Circuits*,
Lecture Notes in Electrical Engineering 66, DOI 10.1007/978-90-481-9379-0_4,

source of device variations, especially when this transistor dimension is scaled down to 10 nm and below.

In this work, we will demonstrate, through the evaluation of ring oscillators and SRAM cells, that FinFET is an excellent candidate for sub-32 nm low-voltage applications. The aggressively scaled SRAM cells with $L_G = 30$ nm and $W_{FIN} = 10$ nm show excellent V_{DD} scalability down to 0.6 V with high Static Noise Margins (SNM). Next, it will be shown that by operating these cells at lower V_{DD} a low σSNM can be achieved.

2 Device Fabrication

A schematic presentation of the process flow is shown in Fig. 1. The starting substrate is a Silicon-On-Isolator (SOI) wafer with 65 nm Si film and 145 nm buried oxide (BOX). Fin widths down to 10 nm were fabricated using 193 nm lithography and aggressive trimming.

To pattern fins, wide active areas and S/D regions, a dedicated Optical Proximity Correction (OPC) was applied to reduce the W_{FIN} variations between dense and sparse areas as well as between inner and outer fins in multi-fin devices. The channels were left undoped and no Si reflow was used to reduce the sidewall roughness. For the gate dielectric, different splits were considered as shown in Table 1. The latter table also contains the extracted CET and V_T values for 20 nm wide fins. The extracted J_G-CET values can be found in Fig. 2.

For all devices 5 nm ALD TiN, capped with 100 nm poly, was used as gate electrode. In this way, almost symmetric V_T's are achieved (Fig. 3). The latter figure also shows the improvement in V_T roll-off when scaling down the fin width from 20 to 15 nm.

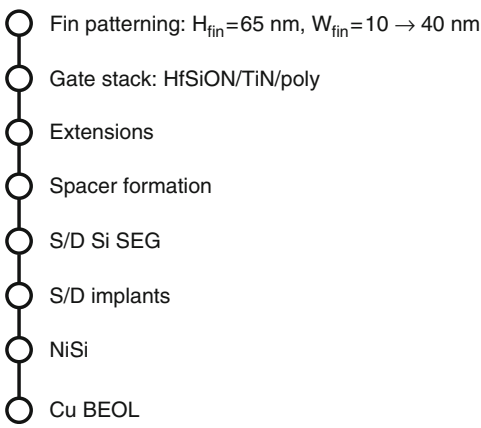


Fig. 1 Schematic presentation of the process flow

Table 1 Gate stacks used in this study; CET and V_T have been extracted for $L_G = 45$ nm and $W_{FIN} = 20$ nm

	MOCVD HfSiO	Nitridation	CET (nm)	V_{TN}[V]/V_{TP}[V]
gate1	40% HfSiO	NH_3 @ 800°C	2.1	0.22/−0.13
gate2	40% HfSiO	DPN	2	0.21/−0.04
gate3	40% HfSiO	NH_3 @ 700°C	2.4	0.34/0.04
gate4	80% HfSiO	NH_3 @ 800°C	1.9	0.34/−0.12

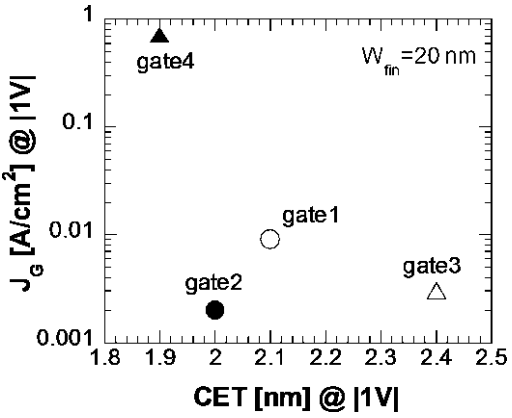


Fig. 2 J_G-CET for the different gate stacks; the parameters were extracted on devices with $W_{FIN} = 20$ nm

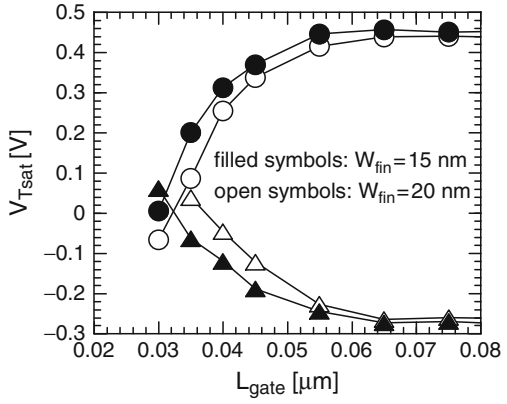


Fig. 3 V_T as function of L_G for nMOS and pMOS with different W_{FIN}

Next, as was implanted for the nMOS extensions and B for the pMOS extensions. A PECVD spacer with 35 nm width was formed and 30 nm undoped Selective-Epitaxial-Growth (SEG) was done to raise the source/drain areas thereby decreasing the parasitic resistance R_{SD}.

After the highly doped (HDD) source/drain implantations, NiSi was used for salicidation. The rest of the processing includes a standard Cu Back-End-Of –Line

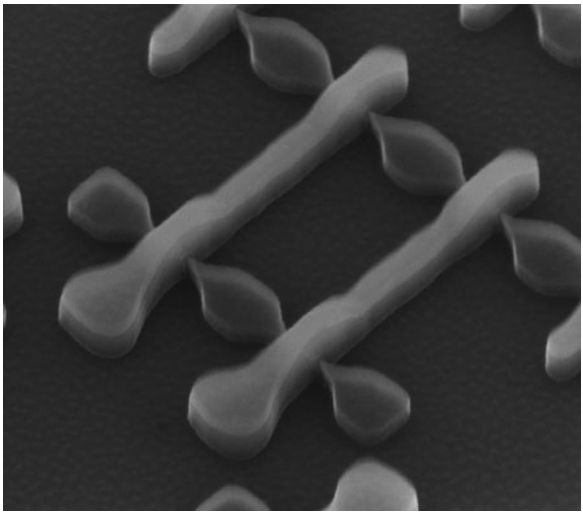


Fig. 4 Tilted SEM picture of an SRAM cell after gate patterning

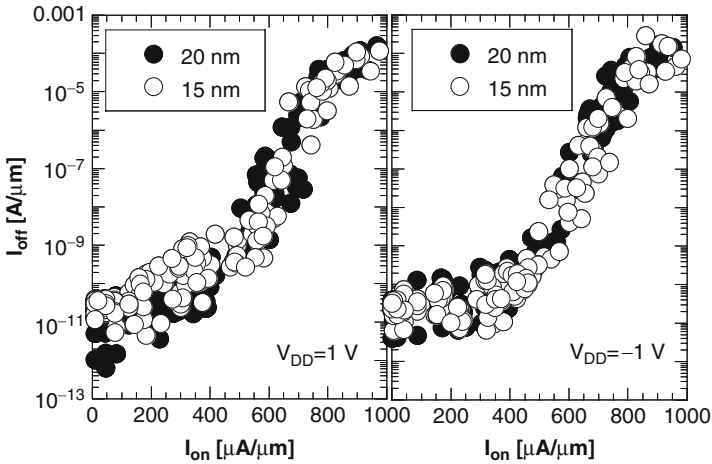


Fig. 5 I_{ON}-I_{OFF} curves of nMOS (*left figure*) and pMOS (*right figure*) with $W_{FIN} = 15$ and 20 nm

(BEOL) going up to metal1 (M1). Figure 4 shows a tilted SEM picture of a SRAM cell after gate patterning.

Figure 5 shows the I_{ON}-I_{OFF} performance of the fabricated nMOS and pMOS. The devices show excellent scalability and high performance. Figure 6 shows the typical transfer characteristics of devices with $L_G = 40$ nm and $W_{FIN} = 20$ nm. Both nMOS and pMOS show low leakage and high drive current.

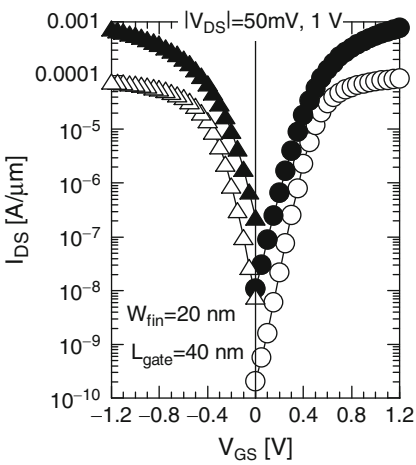


Fig. 6 Typical I_{DS}-V_{GS} curves of nMOS and pMOS with $W_{FIN} = 20$ nm and $L_G = 40$ nm at $|V_{DS}|$ =50 mV and 1 V

3 Results and Discussion

3.1 Ring Oscillators

The impact of the different ratios between the number of pMOS and nMOS fins P_{FIN}/N_{FIN} on the ring oscillator (RO) delay is shown in Fig. 7 for inverter RO's with fan-out 1 (FO1). The minimum RO delay is typically achieved at a ratio close to 1. This is linked to the strong pMOS FinFET performance as is also seen in Figs. 5 and 6. This balanced P_{FIN}/N_{FIN} ratio enables a more area-efficient library layout with high-performance NOR gates.

Figure 8 shows the inverter delay as function of the static power dissipation P_{STAT} for the different gate stacks. In this case, the fin width was scaled down to 20 nm and the gate length $L_G = 50$ nm. Due to the reduced SCE for gate4 a RO delay of 11 ps at 100 nW/stage can be obtained.

Further decrease of the RO delay with gate4 (HfSiO with 80% Hf), even below 10 ps, can be achieved by scaling down the L_G. This is shown in Fig. 9 where the inverter delay is shown for devices with different L_G. For $W_{FIN} = 20$ nm it is clear that the increased SCE, when scaling down the L_G from 50 to 40 nm, results in a 10 to 100 times higher off-state leakage while the inverter delay can be improved further, with values below 10 ps. Scaling down from $L_G = 40$ to 35 nm is not beneficial anymore for the inverter delay. Especially the pMOS devices suffer from high leakage, a degraded subthreshold swing and high Drain Induced Barrier Lowering (DIBL).

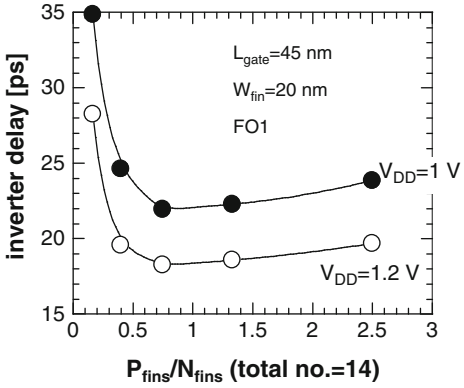


Fig. 7 RO delay as function of the W ratio between pMOS and nMOS the total number of fins is kept constant and equal to 14

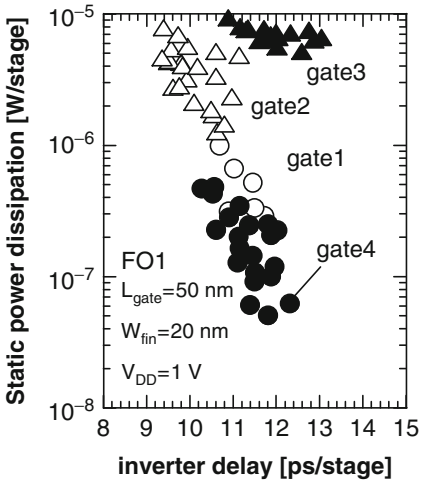


Fig. 8 Inverter delay as function of the static power dissipation P_{STAT}

3.2 SRAM Performance

Two types of SRAM cells have been evaluated: a SRAM cell with $L_G = 45$ nm and $W_{FIN} = 20$ nm (45/20) and a more aggressively scaled cell with $L_G = 30$ nm and $W_{FIN} = 10$ nm (30/10). For both cells, the PU (pull-up), PD (pull-down) and PG (pass-gate) transistors are single fin devices and the fin width and gate length are kept constant for all these transistors (β ratio = 1). Figure 10 shows the measured butterfly curves at different V_{DD}. Both cells show excellent cell stability down to 0.6 V.

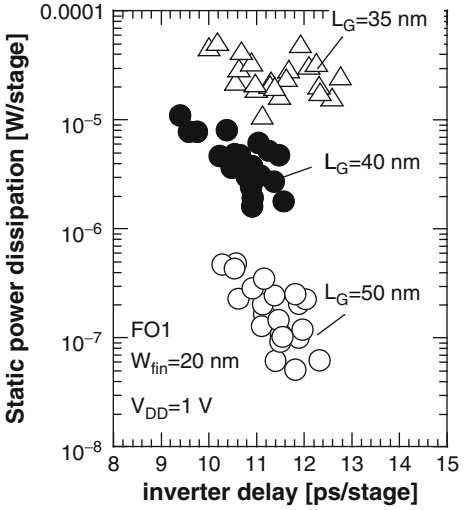


Fig. 9 Inverter delay for ring oscillators with $W_{FIN} = 20$ nm and different L_G; devices with the following gate dielectric are shown: HfSiO with 80% Hf (gate4)

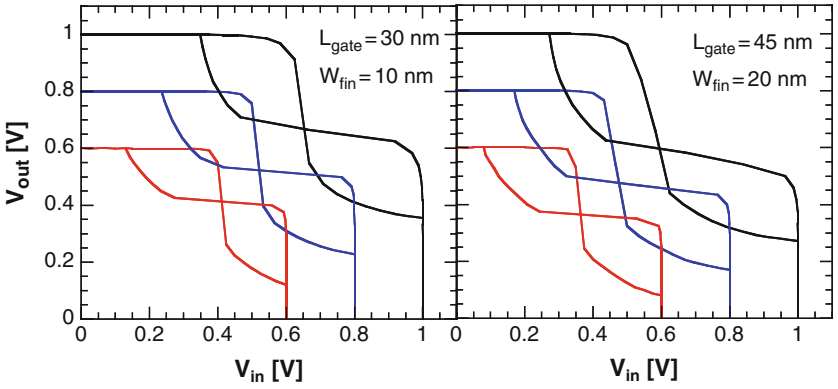


Fig. 10 Measured butterfly curves for two SRAM cells 30/10 (*left*) and 45/20 (*right*); gate dielectric with 80% Hf (gate4)

Figure 11 shows the mean value of the SNM as function of V_{DD} for the different gate stacks. Especially for the 30/10 cell, the better short channel behavior favors the use of gate4 (Table 1). In the latter case, the SNM of the 30/10 cell can be maintained at almost the same level as the 45/20 cell.

Figures 11 and 12 show that with gate3 a much degraded SNM is obtained for the 30/10 cells, whereas the SNM is still high for the 45/20 cell. The high CET of this gate stack is not favorable for keeping the SCE under control when the devices are scaled down. Again, the more leaky but stronger pMOS devices (PU) are responsible for the degraded SNM.

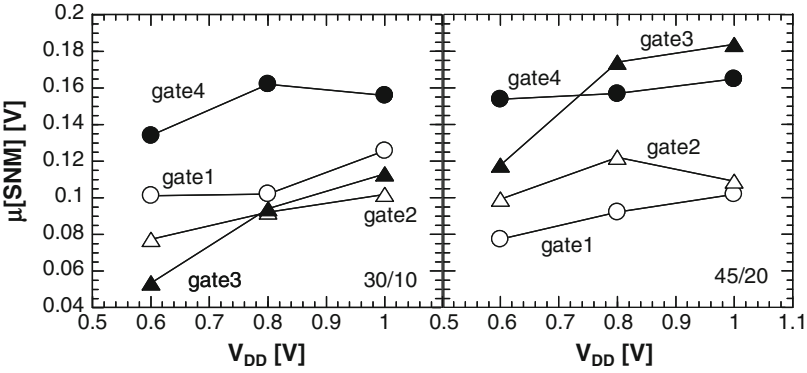


Fig. 11 Read static noise margin as function of VDD for a 30/10 cell (*left*) and 45/20 cell (*right*)

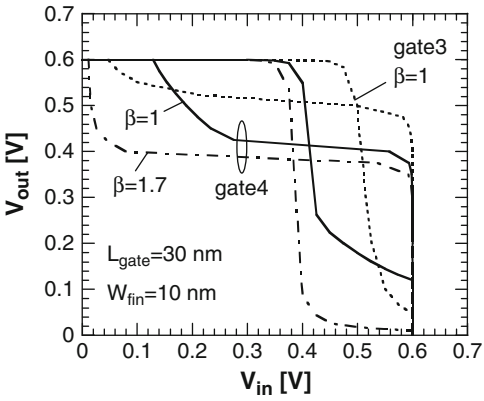


Fig. 12 Measured butterfly curves for a 30/10 cell comparing different gate stacks; the impact of the β ratio is shown at the same time for gate4

Increasing the β ratio of the cell increases the eye of the butterfly curve as shown in Fig. 12. Typically the β ratio is increased by increasing the width of the PD thereby making the PG weaker. In FinFET technology this is done by adding several fins in parallel. However, this decreases the cell density significantly. In this work, β was increased by increasing the L_G of the PG by either 10 nm (β=1.3) or 20 nm (β=1.7). This has little impact on the cell area, but is highly beneficial for the read SNM as can be seen in Fig. 12.

This is especially true at 1V where the read SNM of the 30/10 cell can be increased by almost 80%. One needs to note that when the V_{DD} is scaled down to 0.8 and 0.6 V, the improvement in SNM with increasing β is less significant. This also shows that when low-voltage operation is targeted SRAM cells with β=1 can be used with reasonably high SNM.

Even though the SNM is higher with increasingβ, especially at higher V_{DD}, the write-ability of the cell can be a concern. Already for β=1, this parameter suffers

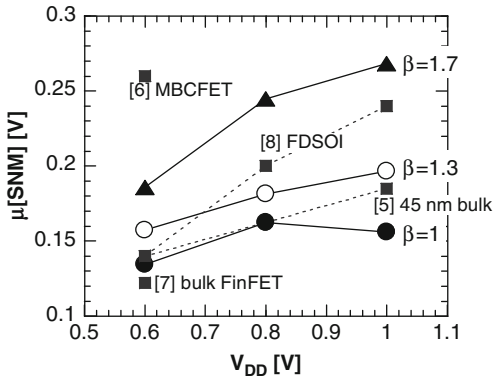


Fig. 13 Impact of the SRAM β ratio is shown for a 30/10 cell and gate stack4 is used

severely from the strong pMOS (Fig. 12) and is degraded even further for $\beta>1$ [9]. However, this high sensitivity of the SNM to L_G indicates that gate length tuning provides sufficient possibilities for the SRAM designer to build stable and writeable cells; even though the transistor widths are fixed by the fin height in SOI FinFET. In combination with write-assist techniques the improved mismatch performance and the better subthreshold slope allow for a continued voltage scaling, which is the key requirement for sub-32 nm low-power low-voltage design.

Finally, we have benchmarked our data to recent SRAM data in Fig. 13. It shows that our data is very competitive and among the best published so far.

3.3 Device and SRAM Variability

In this last section, we will have a look at the device and SRAM variability. The $\sigma(\Delta V_T)$ and $\sigma(\Delta K_0/K_0)$ (transistor gain) for the devices in the SRAM (PU, PD and PG) are shown in Fig. 14. The increased mismatch values for the 30/10 cells can be directly related to variations in both L_G and W_{FIN}. The choice of the gate stack is also crucial in order to set the correct V_T and lower the V_T mismatch in SRAM cells. Again, the most interesting gate stack to use consists of 80%Hf as gate dielectric.

The V_T and conductivity variations can lead to a severe reduction in SRAM SNM when considering large memory arrays. This is demonstrated in Fig. 15 *w*here the butterfly curves of several, in design identical, cells are shown. The reduction of SNM is clearly seen. The overall distribution of SNM is shown in Fig. 16.

This figure shows that as the V_{DD} is decreased the SNM deviation σSNM is reduced, even more so for a cell with higher β ratio. Although the SNM is slightly reduced at lower V_{DD}, the decreased σSNM indicates that it is beneficial to

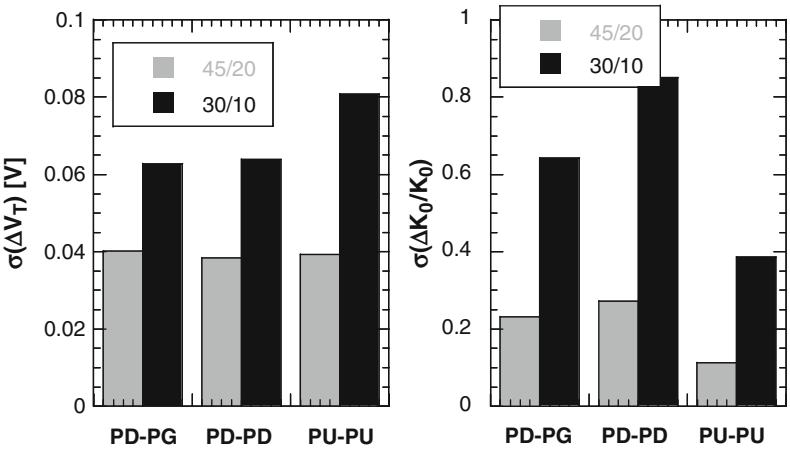


Fig. 14 σ(ΔV+) (*left figure*) and s(ΔK0/K0) (conductivity, *right figure*) are shown for the SRAM transistors in a 30/10 and 45/20 cell

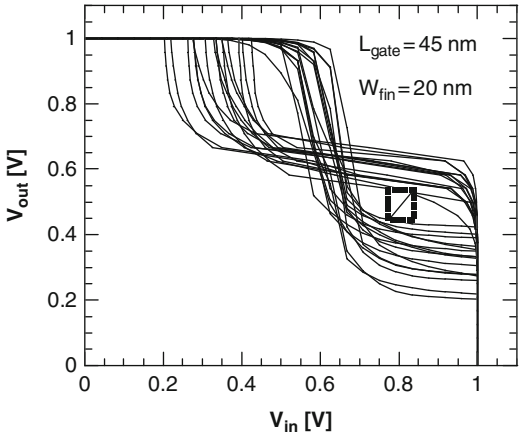


Fig. 15 Butterfly curves of 30 SRAM cells with $L_G = 45$ nm and $W_{FIN} = 20$ nm; the SRAM cells are measured at $V_{DD} = 1$ V

operate these cells at lower V_{DD}. A summary of the data can be found in Fig. 17 where the σSNM/SNM is given as function of the β cell ratio.

4 Conclusions

We have demonstrated scaled FinFET ring oscillators with W_{FIN}=20 nm and SRAM cells with L_G scaled down to 30 nm and W_{FIN}=10 nm. The cells show excellent V_{DD} scalability down to 0.6 V with a high static noise margin of 185 mV.

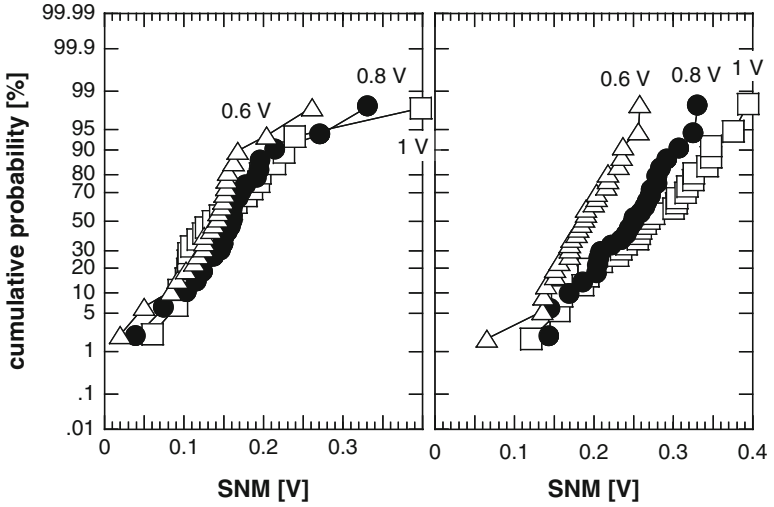


Fig. 16 SNM distribution for the 30/10 cells with different b ratios ($\beta = 1$ on the left side and $\beta = 1.7$ on the right side), measured at $V_{DD} = 0.6$, 0.8 and 1 V

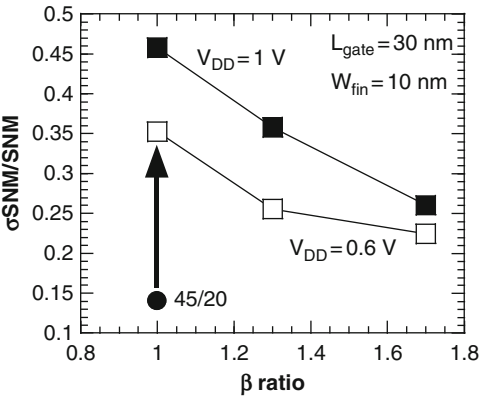


Fig. 17 σSNM/SNM as a function of the β ratio; the results of the 30/10 cells are compared at different V_{DD} and benchmarked to the results of the more relaxed 45/20 cell at $V_{DD} = 0.6$ V

The high cell stability can be achieved by choosing the correct gate stack. Assessment of the device and SRAM variability shows a much reduced σSNM at lower V_{DD}. The latter demonstrates that FinFET-based cells are excellent candidates for sub-32 nm low-voltage design.

Acknowledgment This work has been partially funded by the European PULLNANO Integrated Project (FP6 – IST-026828).

References

1. A. Asenov, Simulation of statistical variability in nano MOSFETs. *VLSI Symposium* (2007), pp 86–87
2. F. Boeuf, M. Sellier, A. Farcy, T. Skotnicki, Impact of layout, interconnects and variability on CMOS technology roadmap. *VLSI Symposium* (2007) pp 24–25
3. A. Dixit, K.G. Anil, E. Baravelli, P. Roussel, A. Mercha, C. Gustin, M. Bamal, E. Grossar, R. Rooyackers, E. Augendre, M. Jurczak, S. Biesemans, K. De Meyer, Impact of stochastic mismatch on measured SRAM performance of FinFETs with resist/spacer-defined fins: role of line-edge-roughness. *IEDM Technical Digest* (2006), pp 709–712
4. K. von Arnim, E. Augendre, C. Pacha, T. Schulz, K.T. San, F. Bauer, A. Nackaerts, R. Rooyackers, T. Vandeweyer, B. Degroote, N. Collaert, A. Dixit, R. Singanamalla, W. Xiong, A. Marshall, C.R. Cleavelin, K. Schrüfer, M. Jurczak, A low-power multi-gate FET CMOS technology with 13.9ps inverter delay, large-scale integrated high performance digital circuits and SRAM. *VLSI Symposium* (2007), pp 106–107
5. R. Morimoto, T. Kimura, Y. Okayama†, T. Hirai, H. Maeda, K. Oshima, R. Watanabe, H. Fukui, Y. Tsunoda, M. Togo, S. Kanai, S. Shino, T. Hoshino, K. Shimazaki, M. Nakazawa, K. Nakazawa, Y. Takasu, H. Yamasaki, H. Inokuma, S. Taniguchi†, T. Fujimaki, H. Yamada, S. Watanabe, S. Muramatsu, S. Iwasa, K. Nagaoka, S. Mimotogi, T. Iwamoto, H. Nii, Y. Sogo, K. Ohno, K. Yoshida, K. Sunouchi, M. Ikeda, M. Iwai, T. Kitano, H. Naruse, Y. Enomoto, K. Imai, S. Yamada, M. Saito, T. Kuwata, F. Matsuoka, N. Nagashima, Layout-design methodology of 0.246-fim2-embedded 6T-SRAM for 45-nm high-performance system LSIs. *VLSI Symposium* (2007), pp 28–29
6. Sung Min Kim, Eun Jung Yoon, Min Sang Kim, Sung Dae Suk, Ming Li, Lian Jun, Chang Woo Oh, Kyoung Hwan Yeo, Sung Hwan Kim, Sung Young Lee, Yong Lack Choi, Na-young Kim, Yun-young Yeoh, Hong-Bae Park, Chul Sung Kim, Hye-Min Kim, Dong-Chan Kim, Heung Sik Park, Hyung Do Kim, Young Mi Lee, Dong-Won Kim, Donggun Park, Byung-Il Ryu, TiN/HfSiOx gate stack multi-channel field effect transistor (McFET) for sub 55 nm SRAM application. *VLSI Symposium* (2006) pp 88–89
7. H. Kawasaki, K. Okano, A. Kaneko, A. Yagishita, T. Izumida, T. Kanemura, K. Kasai, T. Ishida, T. Sasaki, Y. Takeyama, N. Aoki, N. Ohtsuka, K. Suguro, K. Eguchi, Y. Tsunashima, S. Inaba, K. Ishimaru, H. Ishiuchi, Embedded bulk FinFET SRAM cell technology with planar FET peripheral circuit for *hp*32 nm node and beyond. *VLSI Symposium* (2006), pp 86–87
8. Hou-Yu Chen, Chang-Yun Chang, Chien-Chao Huang, Tang-Xuan Chung, Sheng-Da Liu, Jiunn-Ren Hwang, Yi-Hsuan Liu, Yu-Jun Chou, Hong-Jang Wu, King-Chang Shu, Chung-Kan Huang, Jan-Wen You, Jaw-Jung Shin, Chun-Kuang Chen, Chia-Hui Lin, Ju-Wang Hsu, Bao-Chin Perng, Pang-Yen Tsai, Chi-Chun Chen, Jyu-Horng Shieh, Han-Jan Tao, Shih-Chang Chen, Tsai-Sheng Gau, Fu-Liang Yang, Novel 20 nm hybrid SOI/bulk CMOS technology with 0.183 μm 2 6T-SRAM cell by immersion lithography. *VLSI Symposium* (2005), pp 16–17
9. F. Bauer, K. von Arnim, C. Pacha, T. Schulz, M. Fulde, A. Nackaerts, M. Jurczak, W. Xiong, K. T. San, C. -R. Cleavelin, K. Schruefer, G. Georgakos, D. Schmitt-Landsiedel, Layout options for stability tuning of SRAM cells in multi-gate-FET technologies. *ESSCIRC* (2007), pp 392–395

Independent-Double-Gate FINFET SRAM Cell for Drastic Leakage Current Reduction

Kazuhiko Endo, Shin-ichi O'uchi, Yuki Ishikawa, Yongxun Liu, Takashi Matsukawa, Kunihiro Sakamoto,
Meishoku Masahara, Junichi Tsukada, Kenichi Ishii, and Eiichi Suzuki

1 Introduction

The decreased feature size of metal-oxide-semiconductor (MOS) devices in ultra-large-scale-integrated circuits (ULSIs) requires the nano-scale complementary MOS (CMOS) fabrication technology. As silicon devices are scaled down to the nanometer regime, the device technology is facing to several difficulties. Standby power consumption in CMOS devices is now one of the most serious problem and becoming a limiting factor in MOSFET scaling [1]. Short channel effects (SCEs) such as threshold voltage (V_{th}) roll off and sub-threshold slope (S-factor) degradation causes significant increased in power consumption. Catastrophic increase in static power consumption due to shot channel effects (SCEs) becomes the serious problem in future VLSI circuits. Especially, the leakage current in the SRAM array is the most critical issue for a low-power SoC because it occupies the considerable part of LSIs.

Fortunately, non-planar double-gate (DG) MOSFETs provide a potential solution for power consumption issues in ultra-large-scale integrated circuits (ULSIs) [2]. They have fundamental advantages of excellent short-channel effects (SCEs) immunity and high current drivability [2]. Among several types of DG MOSFETs, a fin-type DG-MOSFET (FinFET) has widely been investigated thanks to its process compatibility with the conventional planar MOSFET [3]. Some FinFET SRAM cells have been investigated for the scaled SRAM operation [4–6]. In usual three-terminal (3T) FinFETs, the gate electrodes are under the same potential and the threshold voltage (V_{th}) is fixed depending on the work-function of the gate material [7]. In contrast to the conventional 3T-FinFETs, V_{th}-controllable four-terminal (4T-) FinFETs have been proposed and demonstrated by separating the gate electrode using a

K. Endo (✉), S. O'uchi, Y. Ishikawa, Y. Liu, T. Matsukawa, K. Sakamoto,
M. Masahara, J. Tsukada, K. Ishii, and E. Suzuki
National Institute of AIST, Japan
e-mail: endo.k@aist.go.jp

A. Amara et al. (eds.), *Emerging Technologies and Circuits*,
Lecture Notes in Electrical Engineering 66, DOI 10.1007/978-90-481-9379-0_5,

chemical-mechanical-polishing (CMP) process or an etch-back process [8–10]. These FinFETs are shown in Fig. 1. For the future ultra-low power circuits design, the flexible control of the V_{th} will inevitably be required.

In this chapter, we investigate an experimental integration 4T-FinFETs by a newly developed fabrication process, and demonstrate power-controllable CMOS inverter and SRAM cell operations using the flexible-V_{th} 4T-FinFETs. The circuit diagram of the proposed 4T-FinFET SRAM is shown in Fig. 2 [12]. Each V_{th} control gate for the 4T-FinFET is connected to the corresponding control lines, V_{G2p} or V_{G2n}. These control lines are parallel with word lines (WLs) to realize a row-by-row Vth control for the SRAM array. Level shifters are used to supply V_{G2p} and V_{G2n} by converting a row decoder output signal. The V_{th} of each transistor in the

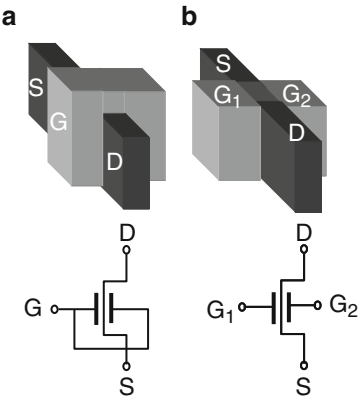


Fig. 1 Schematic illustration of the 3T and 4T FinFETs fabricated using a SOI substrate

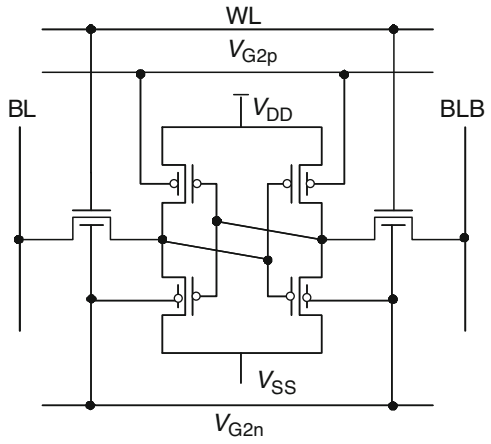


Fig. 2 Circuit diagram of the proposed 4T-FinFET SRAM

stand-by rows is increased by controlling V_{G2p} and V_{G2n} simultaneously to suppress the leakage current. On the other hand, when a certain row is accessed for read or write, V_{th} of each transistor is decreased to maintain high on-current.

2 Fabrication Procedure

We used lightly dope p-type (100)-oriented silicon-on-insulator (SOI) wafers; thus, the channel-orientations of the fabricated FinFETs were (110). Schematic device fabrication processes are shown in Fig. 3. A 50-nm-thick non-doped silicate glass (NSG) layer and the electron beam (EB) resist masks were formed to make hard masks on the wafer. To fabricate vertical Si-fins, the SOI layer was etched by a conventional reactive ion etching (RIE) using a Cl_2 inductively coupled plasma (ICP) as shown in Fig. 1a. After the Si-fin etching, a 3-nm-thick gate-oxide was formed at 850°C followed by the TiN and n^+ polycrystalline-Si (poly-Si) gate formation using EB lithography and RIE. After the gate electrode was formed, a shallow implantation into the extension of the source/drain (S/D) was performed. To distribute impurity atoms (BF_2 for pMOS and P for nMOS) uniformly into the vertical channel, 60° tilted implantation was carried out at an acceleration energy of 5 keV and a dose of 5×10^{13} cm^{-2} in each side [13]. A 1-nm-thick screening oxide was used to suppress the significant dopant loss [14]. A S/D implantation was

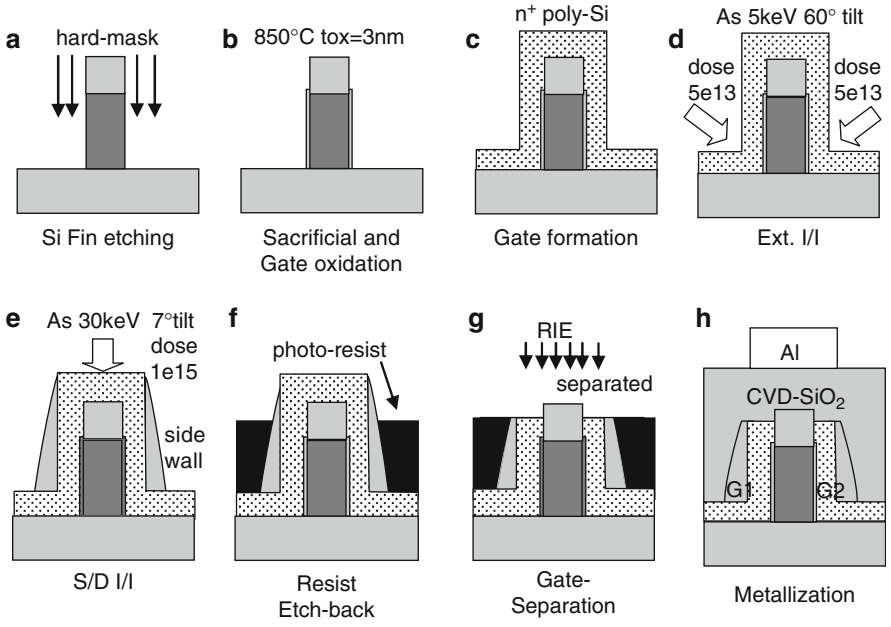


Fig. 3 Fabrication process flow for the 4T-FinFET. After the two-step ion implantation (d)–(e), the gate separation process as shown in (g) is carried out

performed at an acceleration energy of 15 keV and a dose of 1×10^{15} cm^{-2} after a 100-nm-thick gate-sidewall was formed by using CVD grown SiO_2. The acceleration energy was set to 15 keV to preserve the seed-crystal layer for the recrystalline annealing.

Then, the gate electrode for the 4T-FinFET was separated by using a newly developed resist etch-back process whereas the 3T-FinFET region was protected by a thicker photo-resist. Due to the three-dimensionally shaped Si-Fin, the thickness of the spin-coated EB resist (SAL-601) was thinner at the top of the Si-fin than that at the other planar portion. Consequently, the poly-Si gate at the top of the Si-fin was revealed by the partial ashing of the EB resist. After the poly-Si gate was revealed by thinning the EB resist, the poly-Si gate was separated using ICP-RIE with HBr based chemistry and the poly-Si gate over the Si-fin connected to the each side of the gate was completely removed. Finally, the S/D was activated at 900°C for 2 s and the devices were sintered at 450°C in 3% H_2 ambient after the metallization.

3 Results and Discussons

3.1 Fabrication Results

Figure 4a shows a scanning electron microscope (SEM) plan-view of the fabricated 3T-FinFET after the gate etching. The Si-channel was fully surrounded by the gate poly-Si at the center of the Si-fin. After the gate side-wall spacer formation using a chemical vapor deposited SiO_2 as shown in Fig. 4b, P or BF_2 ions were implanted into the source and drain region. After the ion implantation, the EB resist was spin-coated on the whole chip for the gate separation etching. The EB resist was then partially removed so that the top of the poly-Si for the 4T-FinFET region was selectively revealed as shown Fig. 4c. Figure 4d shows a SEM image of the 4T-FinFET after the gate separation etching. The poly-Si gate was successfully separated by the gate separation etching and the fin-top was revealed through a resist opening.

Figures 5 and 6 show the drain-current versus gate-voltage ($I_D - V_{G1}$) characteristics of the fabricated 3T- and 4T-FinFETs with the gate length (L_g) of 110 nm and the fin width (T_{Fin}) of 18 nm. The V_{th} can be flexibly controlled by introducing a bias voltage to the control electrode (G_2) of the 4T-FinFET as shown in Fig. 6. On the contrary, the 3T-FinFET provides excellent S-slope and drain induced barrier lowering (DIBL) value. To control the threshold voltage of the 4T-FinFET efficiently, we investigated a proper bias condition for the control gate (G_2). Figure 6a shows the $I_D - V_{G1}$ curves for the single drive (SD) mode where the V_{G2} is fixed to a certain value. In this case, V_{th} shift rate γ defined by $-\delta V_{th}/\delta V_{G2}$ increases with a reducing fin width (T_{Fin}) as summarized in Fig. 8a. This is because γ is expressed as $\gamma = 3T_{ox}/(3T_{ox}+T_{Fin})$ where T_{ox} is the gate-oxide thickness. This means that the

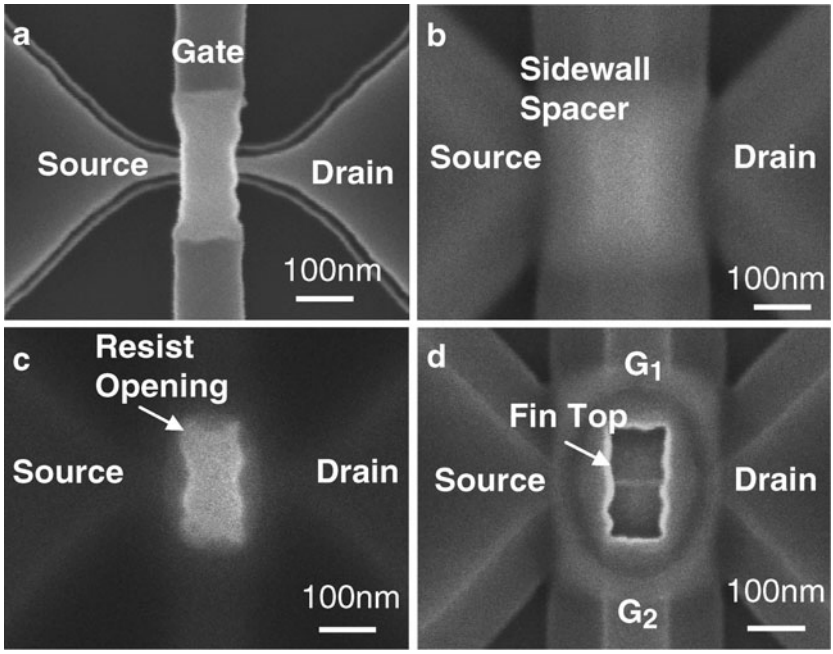


Fig. 4 SEM plan-view of the FinFETs (**a**) after the gate poly-Si etching, (**b**) after the side-wall spacer formation, (**c**) after the EB resist coating and partial ashing, (**d**) after the gate separation etching

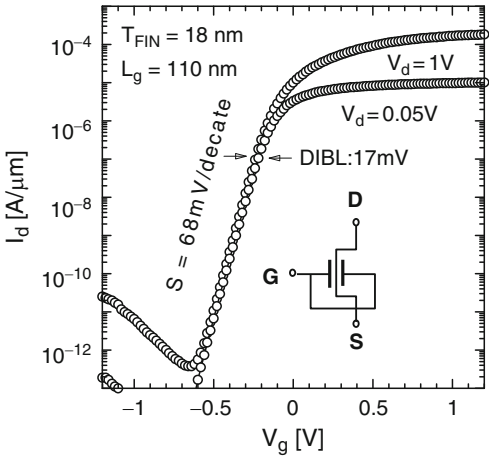


Fig. 5 $I_D - V_G$ characteristics of the 3T-FinFET with $T_{Fin} = 18$ nm and $L_G = 110$ nm

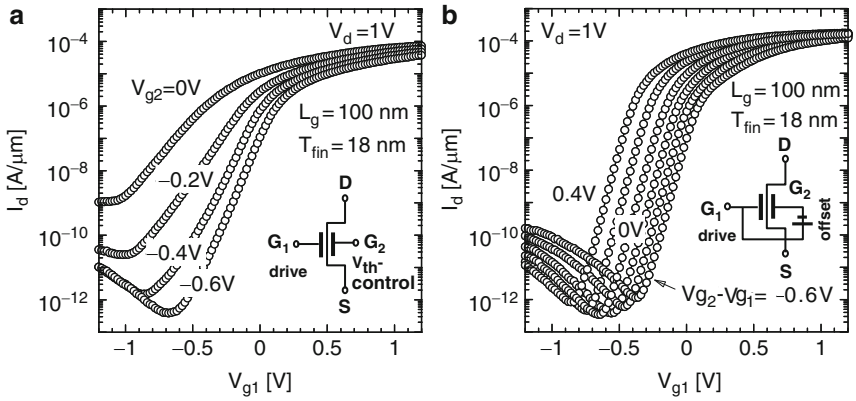


Fig. 6 $I_D - V_{G1}$ characteristics of (**a**) single drive (SD) and (**b**) synchronized double gate drive (DD) mode operation for the 4T-FinFET with $T_{Fin} = 18$ nm and $L_G = 110$ nm

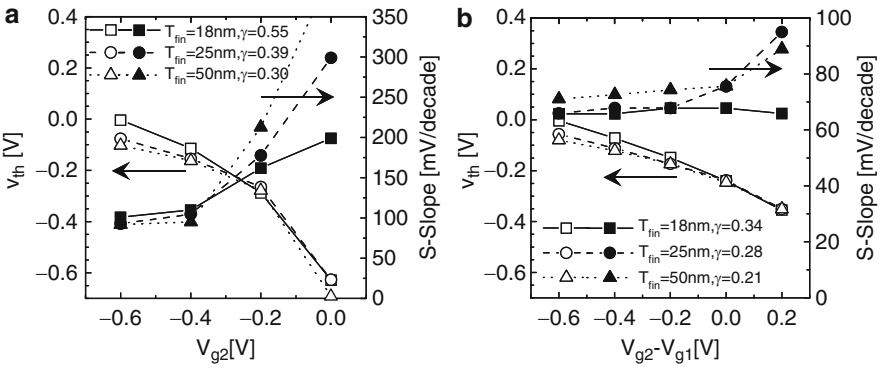


Fig. 7 Threshold voltage controllability and S-slope of (**a**) single drive (SD) mode and (**b**) synchronized double gate drive (DD) mode operation for the 4T-FinFETs with $L_G = 110$ nm and various T_{Fin}

4T-FinFET with a smaller T_{Fin} is effective for the V_{th} adjustment than that with a thicker T_{Fin}. This value is much higher than that obtained by the back-bias control for the conventional planer MOSFET [15]. This suggests that the V_{th} controllability of the 4T-FinFET is much efficient than that obtained by the body biasing method thanks to the smaller control gate of the 4T-FinFET. However, the s-slope as summarized in Fig. 7a is significantly deteriorated in this static biasing mode especially at $V_{G2} > -0.4$V because the second gate (G_2) changes from depletion to inversion state [16]. To control the V_{th} of the 4T-FinFET with a small s-slope, the both gates need to be synchronously biased. Figures 6b and 7b shows the results with the synchronously driving double gates (DD) operation where both V_{G2} and V_{G1} are simultaneously driven with an offset voltage. In this mode, the $I_D - V_{G1}$ characteristics can be controlled with no degradation in the S-slope especially for the device

with narrower T_{Fin}. This indicates that the DD mode operation provided by appropriate circuit geometry is promising for the future integration of 4T-FinFETs.

3.2 CMOS Inverter Operation

These 4T-FinFETs were integrated and the CMOS inverter composed by 4T-FinFETs was successfully fabricated as shown in Fig. 8. In this circuit, there are not only input and output terminals but also V_{th} control terminals for the both pMOS and nMOS devices. To control the characteristics of the CMOS inverter, both second-gates need to be controlled.

If one V_{G2} is fixed and the other is varied, the logical threshold voltage can be flexibly changed as shown in Fig. 9. On the contrary, by changing both V_{G2} for pMOS (V_{G2p}) and nMOS (V_{G2n}) synchronously towards the opposite direction, the short circuit current of the CMOS inverter can dynamically be controlled with keeping the logical threshold as shown in Fig. 10. This implies that the power-managed CMOS operation can be accomplished. The lower logical threshold voltage below 0.5 V is the issue for the further optimization. We found that the peak short circuit current can exponentially be reduced with decreasing V_{G2}. This indicates that the CMOS circuits are tunable for both high- and low-power operations, which strongly suggests the advantage of the power-managed CMOS circuit using 4T-FinFETs.

3.3 SRAM Operations

For the SRAM cell integration, we introduced a TiN metal-gate process to properly adjust the V_{th} [17]. Figure 11 shows the cross-sectional scanning transmission electron microscope (STEM) view of the co-fabricated 3T and 4T-FinFET.

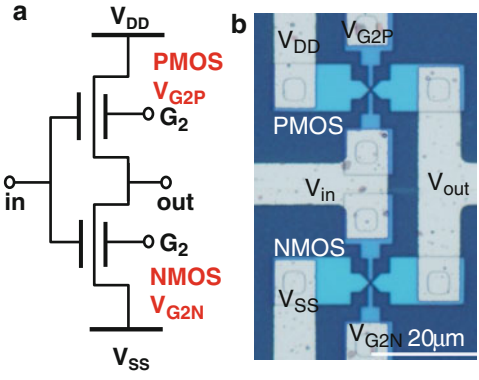


Fig. 8 (**a**) Circuit diagram and (**b**) optical microscopic view of the CMOS inverter composed by 4T-FinFETs

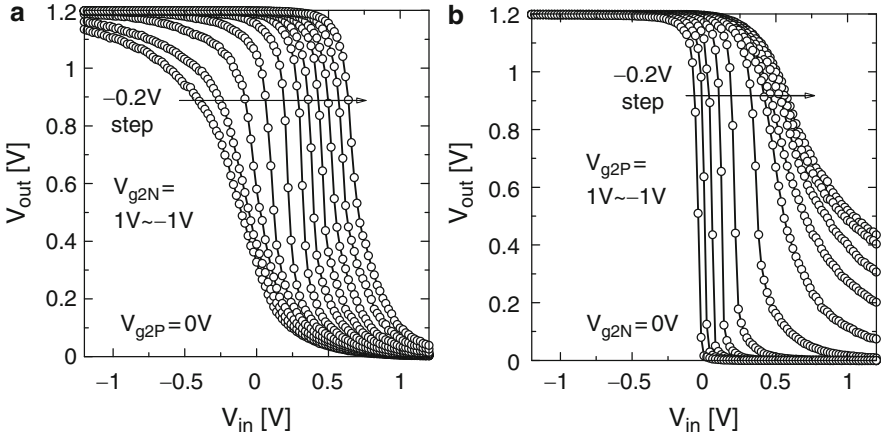


Fig. 9 Input–Output characteristics of the CMOS inverter composed by 4T-FinFETs with T_{Fin} = 50 nm and L_g = 110 nm; (**a**) V_{G2p} = 0V and V_{G2n} is varied, (**b**) V_{G2n} = 0V and V_{G2p} is varied

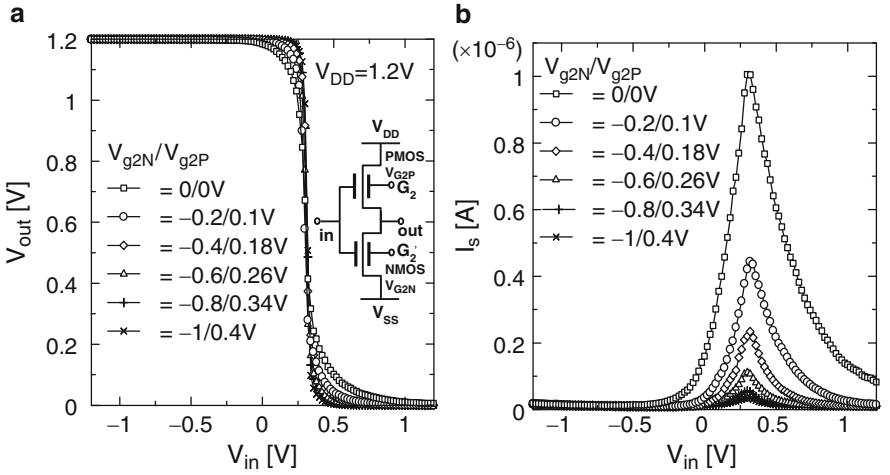


Fig. 10 (**a**) Input–output characteristics of the CMOS inverter composed by 4T-FinFETs (T_{Fin} = 50 nm and L_g = 110 nm) with various control-gate biasing conditions and (**b**) short circuit currents. Drastic reduction of the short circuit currents is demonstrated by adding the appropriate bias to the control gates (G_2)

The resist opening position was selectively etched during the gate-separation process and the gate-separated 4T-FinFET was successfully fabricated.

Figure 12 shows the drain-current versus gate-voltage ($I_D - V_{GI}$) characteristics of the fabricated 3T- and 4T-FinFETs with the gate length (L_G) of 110 nm and the fin width (T_{Fin}) of 50 nm. Almost symmetric $I_D - V_G$ characteristics are realized thanks to the mid-gap TiN metal-gate. The V_{th} can be flexibly controlled by

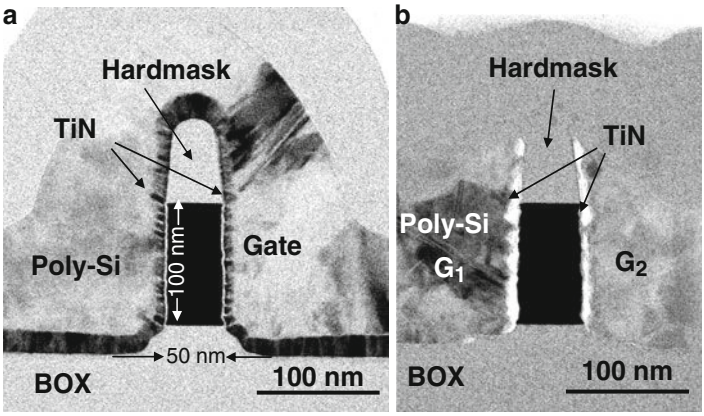


Fig. 11 Cross-sectional scanning TEM images of the 3T-FinFET (**a**) and the independent-gate 4T-FinFET (**b**)

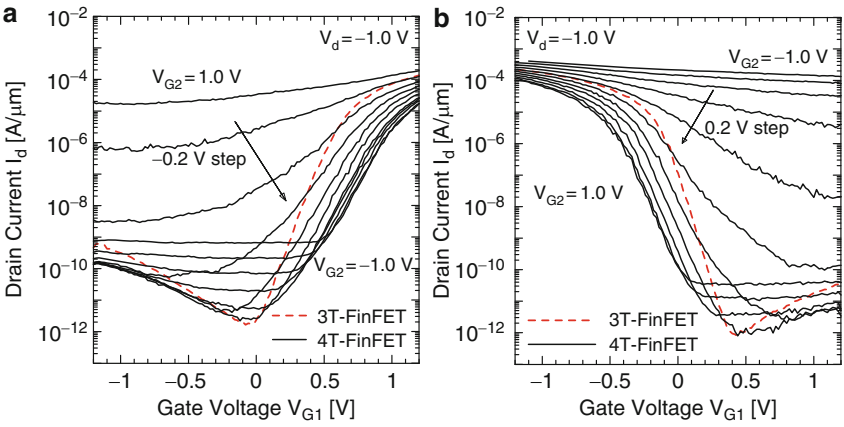


Fig. 12 $I_D - V_{G1}$ characteristics of the 3T and 4T-FinFET with $T_{Fin} = 50$ nm and $L_G = 110$ nm; (**a**) nMOS, (**b**) pMOS

introducing a bias voltage to the control electrode (G_2) of the 4T-FinFET. To suppress the I_{off} of the transistor without decreasing the drivability, the 4T-FinFET technology is effective because the on and off currents are flexibly controlled depending on the bias voltage of the control gate.

By integrating the 4T-FinFETs, the SRAM cell has successfully fabricated. Figure 13 shows the fabricated SRAM half cell with 4T-FinFET. In addition to the WL and BL, the cell has Vth control gates for the pass-gate, the pull-down, and the pull-up transistors. In an m-row × n-column SRAM sub-array, the average leakage current of one cell is calculated by the Eq. 1

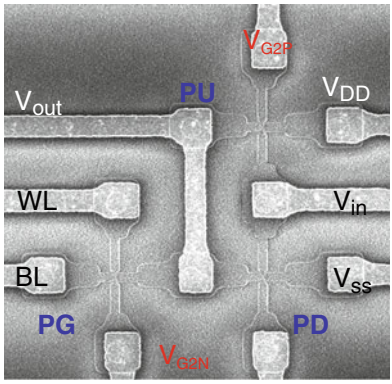


Fig. 13 SEM image of the 4T-FinFET SRAM half cell

Table 1 Experimental leakage currents

Circuitry	Leakage current (A/μm)
4T-FinFET (high-V_{th})	2.6×10^{-10}
4T-FinFET (low-V_{th})	1.6×10^{-5}
4T-FinFET SRAM (average)	6.3×10^{-8}
3T-FinFET SRAM	2.4×10^{-7}

$$I_{leak,average} = \frac{1}{m} I_{leak,lowVth} + \frac{m-1}{m} I_{leak,highVth} \tag{1}$$

We assume that the bias voltage for the V_{th} control is varied between −1V and 1V and calculate the average leakage current in a 256 × 256 SRAM array. Table 1 summarizes the leakage current in the 4T-FinFET SRAM and the conventional 3T-FinFET SRAM cells. Although the leakage current of the high-Vth 4T-FinFET is higher than that of the 3T-FinFET, the average SRAM leakage current is much lower for the 4T-FinFET case. If we reduce the leakage current of the high-Vth 4T-FinFET by introducing an asymmetric gate-oxide [18], the leakage current of the SRAM cell can be further suppressed.

Figure 14 shows the butterfly curves for the 4T-FinFET SRAM with various bias conditions for the V_{th} control gates. The short circuit currents in the flip-flop of the SRAM cell are also shown. To balance the drivability of the transistors, both of the V_{G2n} and V_{G2p} need to be controlled in the opposite direction as shown in Fig. 12. The butterfly curves of the 4T-FinFET SRAM do not change with the bias voltages because the drivability of the pass-gate to the pull-down transistor (beta ratio) is unchanged. Therefore, the static noise margin (SNM) of the SRAM keeps around 210 mV regardless the bias voltage.

Furthermore, the shot circuit current in the flip-flop can be flexibly controlled by the second gate bias voltage as shown in Fig. 14. This means that not only the static

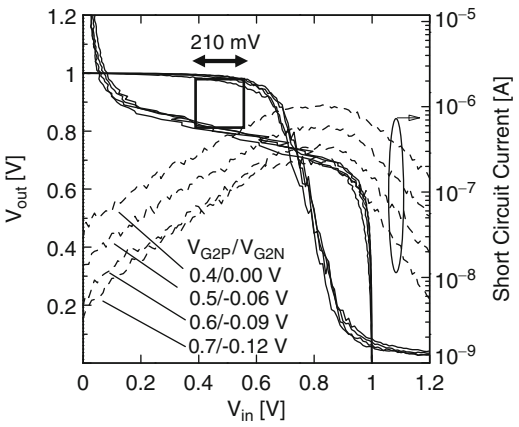


Fig. 14 Butterfly curves and short circuit currents for the 4T-FinFET SRAM with various biasing condition for the V_{th} control gates

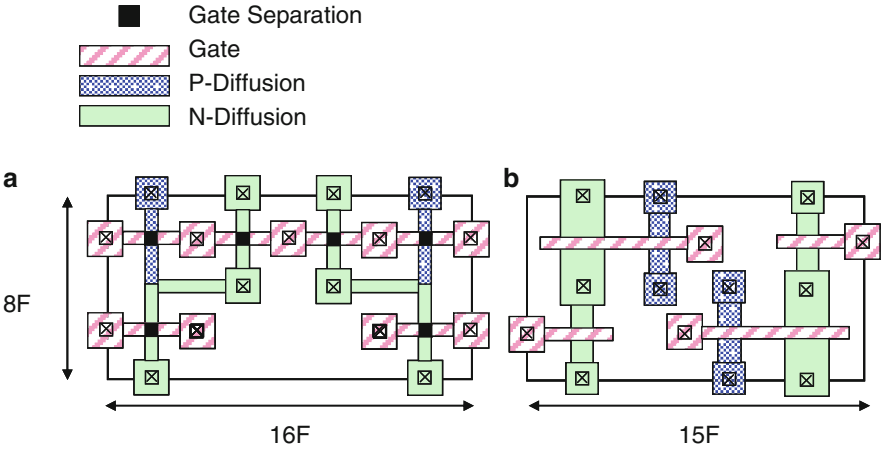


Fig. 15 Layout for the 4T-FinFET SRAM cell (a) and planar-bulk MOSFET SRAM cell. The cell areas are $128F^2$ and $120F^2$, respectively

leakage current, but also the dynamical power consumption can be controlled by the 4T-FinFET SRAM.

Figure 15 shows the proposed layout for the 4T-FinFET SRAM cell. The standard CMOS SRAM occupies an area of $120F^2$, while the 4T-FinFET SRAM occupies $128F^2$. Here, F stands for half of the first metal-layer wiring pitch. Thus, a slight area overhead of 7% compared to the standard CMOS SRAM is required to integrate the second gate contact for the 4T-FinFETs.

4 Conclusion

3T- and 4T-FinFETs have been successfully co-fabricated by utilizing the optimized CMOS fabrication processes. The fabricated 3T-FinFET shows excellent sub-threshold characteristics and DIBL whereas the 4T-FinFET provides efficient Vth controllability. These FinFETs are integrated into CMOS inverter circuits and SRAM cells. We demonstrated the reduction of not only the standby leakage current, but also the dynamic poser consumption by appropriately controlling the Vth of the 4T-FinFET. Although the leakage current of low-Vth 4T-FinFET is higher than that of the 3T-FinFET, the row-by-row Vth control can allow the average leakage current of the 4T-FinFET SRAM much lower than that of the 3T-FinFET SRAM. Thus, the fabricated 4T-FinFET SRAM is promising for the future scaled circuitry.

Acknowledgement The author would like to thank Ms. Yuki Ishikawa, Dr. Yongxun Liu, Dr. Takashi Matsukawa, Dr. Shin-ichi O'uch, Dr. Meishoku Masahara, Mr. Junichi Tsukada, Mr. Kenichi Ishii, Ms. Hiromi Yamauchi, and Dr. Eiichi Suzuki for their support and helpful discussions.This work was supported in part by the Innovation Research Project on Nanoelectronics Materials and Structures t from the METI.

References

1. D.J. Frank, Tech. Digest Int. Elec. Dev. Meeting, 643–646 (2002)
2. T. Sekigawa, Y. Hayashi, Solid State Electron. **27**, 827 (1984)
3. D. Hisamoto, W. Lee, J. Kedzierski, H. Takeuchi, K. Asano, C. Kuo, E. Anderson, T. King, J. Bokor, C. Hu, IEEE Trans. Electr. Dev. **47**, 2320 (2000)
4. H. Kawasaki, K. Okano, A. Kaneko, Y. Yagishita, T. Izumida, T. Kanemura, K. Kasai, T. Ishida, T. Sasaki, Y. Takeyama, N. Aoki, N. Ohtsuka, K. Suguri, K. Eguchi, Y. Tsunashima, S. Inaba, K. Ishimaru, H. Ishiuchi, *Digest of Technical Papers, IEEE VLSI Technology Symposium* (2006), p. 86–87
5. A. Dixit, K.G. Anil, E. Baravelli, P. Roussel, A. Mercha, C. Gustin, M. Bamal, E. Grossar, R. Rooyackers, E. Augendre, M. Jurczak, S. Biesemans, K. De Meyer, Tech. Digest Int. Elec. Dev. Meeting (2006), pp. 709–712
6. A. Bansal, S. Mukhopadhyay, K. Roy, IEEE Trans. Electr. Dev. **47**, 1409 (2007)
7. P. Francis, A. Terao, D. Flandre, F. Van de Wiele, IEEE Trans. Electr. Dev. **41**, 715 (1994)
8. Y. Liu, M. Masahara, K. Ishii, T. Sekigawa, H. Takashima, H. Yamauchi, E. Suzuki, IEEE Electr. Dev. Lett. **25**, 510 (2004)
9. D.M. Fried, J.S. Duster, K.T. Kornegay, IEEE Electr. Dev. Lett. **24**, 59 (2003)
10. L. Mathew, Y. Du, S. Kalpat, M. Sadd, M. Zavala, T. Stephens, R. Mora, R. Rai, S. Becker, C. Parker, D. Sing, R. Shimer, J. Sanez, A.V.-Y. Thean, L. Prabhu, M. Moosa, B.-Y. Nguyen, J. Mogab, G.Workman, A. Vandooren, Z. Shi, M. Chowdhury, W. Zhang, J. Fossum, *Digest of Technical Papers, IEEE VLSI Technology Symposium* (2005), pp. 200– 201
11. L. Mathew, Y. Du, A. Thean, M. Sadd, A. Vandooren, C. Parker, T. Stephens, R. Mora, R. Rai, M. Zavala, D. Sing, S. Kalpat, J. Hughes, R. Shimer, S. Jallepalli, G. Workman, W. Zhang, J. Fossom, B. White, B. Nguyen, J. Mogab, in *Proceedings of the IEEE SOI Conference* (2004), p. 187

12. S. O'uchi, M. Masahara, K. Endo, Y.X. Liu, T. Matsukawa, K. Sakamoto, T. Sekigawa, H. Koike, E. Suzuki, ICICE Trans. E91-C, 534 (2008).
13. K. Endo, M. Masahara, Y.X. Liu, T. Matsukawa, K. Ishii, E. Sugimata, H. Takashima, H. Yamauchi, E. Suzuki, Jpn. J. Appl. Phys **45**, 3097 (2006)
14. M. Koh-Masahara, K. Esuga, H. Furumoto, T. Shirahata, E. Seo, K. Shibahara, S. Yokoyama, M. Hirose, Jpn. J. Appl. Phys **38**, 2324 (1999)
15. M. Togo, T. Fukai, Y. Nakahara, S. Koyama, M. Makabe, E. Hasegawa, M. Nagase, T. Matsuda, K. Sakamoto, S. Fujiwara, Y. Goto, T. Yamamoto, T. Mogami, M. Ikeda, Y. Mamagata and K. Imai, Digest of Technical Papers, IEEE VLSI Technology Symposium, 2004, p. 88
16. M. Masahara, Y. Liu, K. Sakamoto, K. Endo, T. Matsukawa, K Ishii, T. Sekigawa, H. Yamauchi, H Tanoue, S. Kanemaru, H. Koike and E. Suzuki, IEEE Trans. Electron Devices 52, 2046 (2005).
17. Y.X. Liu, S. Kijima, E. Sugimata, M. Masahara, K. Endo, T. Matsukawa, K. Ishii, K. Sakamoto, T. Sekigawa, H. Yamauchi, Y. Takanashi, E. Suzuki, IEEE Trans. Nanotechnology **5**, 201 (2007)
18. M. Masahara, R. Surdeanu, L. Witters, G. Doornbos, V.H. Nguyen, G. Van den Bosch, C. Vrancken, K. Devriendt, F. Neuilly, E. Kunnen, M. Jurczak, S. Biesemans, IEEE Electron Device Lett. **28**, 217 (2007)

Metal Gate Effects on a 32 nm Metal Gate Resistor

Thuy Dao, Ik_Sung Lim, Larry Connell, Dina H. Triyoso, Youngbog Park, and Charlie Mackenzie

1 System – On – Chip Requirements

Driven by the anticipation of fewer components, hence lower cost and lower power consumption, System – on – Chip for wireless communication products has been proposed and worked on for more than 15 years. One of the most aggressive SOC integration roadmaps is for the single-chip cellular phone. Figure 1 shows a simple block diagram of the cellular phone chip-set.

In 2001, a typical cellular phone chip-set included digital CMOS, Flash, Power Management, a RF/Mixed Signal block, a Power Amplifier, and <200 passives. In 2007, a low cost cellular phone chip-set may have CMOS, Flash, a PA, and <20 passives. The integration of Analog/RF functions into digital logic requires a paradigm shift in RF radio design and architecture [1] and in the technology roadmap; historically, the analog / RF roadmap typically lagged the digital /base-band roadmap by 18 months to 2 years. Starting with the 45 nm node, both digital/ baseband and analog/RF technologies are being offered in about the same time frame by semiconductor foundries. In addition, trade-offs in RF performance are required when integrating the RF functions into a single CMOS logic chip. As CMOS technology continues to scale down into the nano regime to enable higher performance and smaller form factors, the SOC or increasing integration of RF/Analog functions with digital logic becomes attractive enough for many to consider accepting the performance trade-off for lower cost. However, scaling of

T. Dao and D.H. Triyoso
Freescale Semiconductor, Austin, Texas, USA

I. Lim (✉), Y. Park, and C. Mackenzie
Freescale Semiconductor, Tempe, Arizona, USA
e-mail: Dao Thuy: rmav4@freescale.com

L. Connell
Freescale Semiconductor, Lake Zurich, Illinois, USA

A. Amara et al. (eds.), *Emerging Technologies and Circuits*,
Lecture Notes in Electrical Engineering 66, DOI 10.1007/978-90-481-9379-0_6,

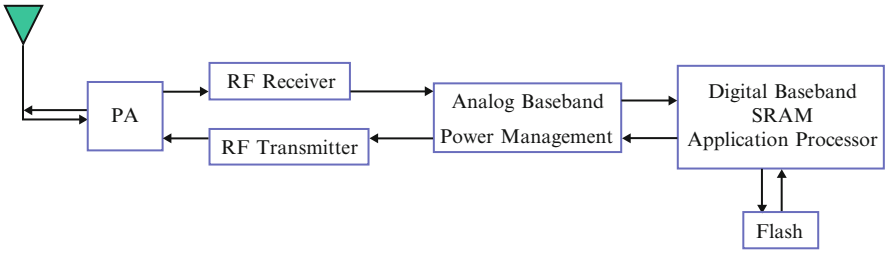


Fig. 1 Cell phone simple block diagram

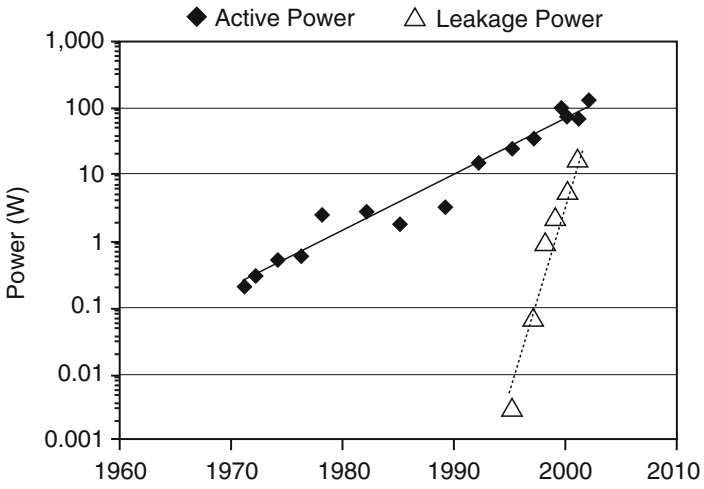


Fig. 2 Active versus standby power [1]

bulk CMOS technology with the use of traditional planar silicon gate and gate oxide dielectric is reaching the physical limitations for gate oxide thickness and gate leakage current has increased substantially. In the past 10 years, the standby power has increased at a faster rate than the active power [2], so much so that it is now a significant concern to chip designers.

Figure 2 shows the increase in the magnitude of standby power versus active power since 1970. At the same time, reducing standby power for longer battery life is increasingly important for portable products. The standby power or leakage power is approaching 40% of the total power dissipation required for today's high performance microprocessors [2]. In a transistor, the magnitude of the leakage power depends on the device channel length, channel width, gate dielectric thickness and material, substrate (SOI or bulk), and threshold voltage. The three major leakage components: junction leakage, subthreshold leakage, and gate leakage, with gate leakage being the main component, all contribute to the rise in leakage power. Gate Leakage is defined as leakage from gate to substrate,

gate/drain, gate/source, drain/gate overlap, and source/gate overlap. The gate leakage has became significant in the past few years due to technology scaling into the nano technology regime – narrower channel length and increasing overlap, thinner gate oxide thickness and increasing electric field across the oxide. The implementation of a high-K dielectric and metal gate, which is projected to be in manufacturing at 32 nm technology node, should help reduce gate leakage. However, the integration of the advanced gate stack materials has serious implications for transistor performance and reliability. Significant resources have been devoted to studying and solving these issues in order to introduce the high-K/metal gate stack into high volume production. However, there have not been many reports on the effect of high-K / metal gate on analog passive devices.

2 High-K/Metal Gate Effects on Passives

The semiconductor industry has expended much effort to search for the right high-k and metal gate materials to continue transistor scaling as well as to continue increasing capacitance density for MIM capacitors. At the 32 nm complementary metal oxide semiconductor (CMOS) technology node, the metal-oxide-semiconductor-field-effect-transistor (MOSFET) gate dielectric equivalent SiO_2 thickness is expected to be <13 Å. At this thickness, leakage current through the SiO_2 gate oxide becomes unacceptable. The search for alternative gate dielectric material to replace SiO_2 has been going on for many years. Recently the semiconductor industry has converged and identified hafnium-based dielectrics as the most promising replacement of SiO_2. Intel has implemented a Hf-based high-k dielectrics in its 45 nm Penryn chip [3]. The rest of the semiconductor industries plan to implement high-k/metal gate at CMOS 32 nm node and beyond. Figure 3 shows a typical high K/Metal gate stack.

Some of the challenges in integrating HfO_2 high-k dielectrics into CMOS platform include mobility degradation, fixed charge and threshold voltage instabilities. Improved properties are reported when HfO_2 thickness is reduced. However, ultra-thin HfO_2 only has a medium dielectric constant (k of 20 or less) which limits its scalability. Another approach to improve HfO_2 properties is by alloying it with other metal oxides such as La_2O_3, ZrO_2, and SiO_2 [4–11 for example]. Freescale

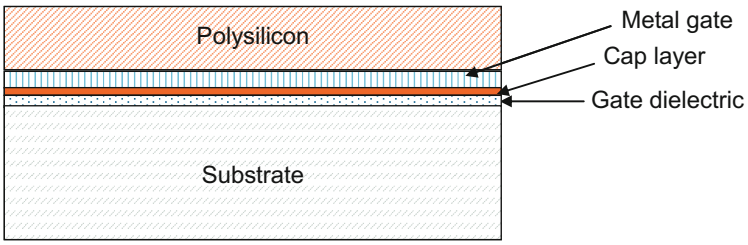


Fig. 3 High-K/metal gate stack

has reported improved HfO_2 properties by addition of ZrO_2 [6–8]. Addition of ZrO_2 into HfO_2 has been shown to stabilize the tetragonal phase (the phase with higher dielectric constant) as well as improving device performance and reliability.

To overcome P+ A-silicon depletion issue, metal gate electrodes are needed to replace Si gate. A number of simulation studies have shown that to achieve comparable device performance to n^+ and p^+ Si, the metal electrodes must have work function near the conduction and valence band edge of silicon, respectively. Two promising approaches to obtain high-k / metal gate stack with the desired work function is to add a dielectric capping layer (such as La_2O_3, Mg_2O_3, Al_2O_3) on the high-k (HfO_2, $Hf_xZr_{1-x}O_2$, $HfSiO_x$) or to directly alloy these metals (La, Mg, Al) into the metal gate (TaC, TiN, TaN, MoN). Excellent transistor characteristics have been reported using both approaches. IBM reported on n-type band-edge results HfO_2 with La_2O_3 dielectric capping layer and TiN metal gate [4, 5]. Similar V_t tuning for n-type band-edge could also be achieved by doping HfO_2 with other rare earth materials [4]. Alloying Al into MoN resulted in p-type band edge as reported by SEMATECH and Freescale [13, 14]. Additionally, SEMATECH has also shown V_t tuning for p-type band-edge by using SiGe channel on Si with HfSiON/TiSiN gate stack [15]

High-k is also being used in capacitor application to increase capacitance density. Metal insulator metal (MIM) capacitors using high-k dielectrics such as HfO_2, Ta_2O_5, Al_2O_3 or a combination thereof have been extensively reported in the literature [16–18 for example]. The metals used in conjunction with these high-k dielectrics are typically TiN or TaN. Capacitance density in the range of 5–8 fF/μm^2 can be achieved by changing the thickness of Al_2O_3 dielectric. Higher capacitance densities have been reported by others. Chiang et al. demonstrated 23 fF/μm^2 capacitance density with Ir/TiTaO/TaN structure [18]. ST has reported 3D MIM capacitor with Ta_2O_5 with capacitance density as high as 17 fF/μm^2 [16].

Many methods have been investigated to deposit these high-k dielectrics and metal for transistor and capacitor applications. Three commonly studied deposition techniques include physical vapor deposition (PVD), chemical vapor deposition (CVD), and atomic layer deposition (ALD). Each method has its own advantages and disadvantages. PVD would give a high throughput and pure films (as pure as the target used). Plasma induced damage and conformality are a concern. CVD technique gives a good throughput, good conformality without plasma damage. However, film will contain impurities that will come from the precursors. ALD and MOCVD have been extensively explored for high-k and metal gate deposition [6, 11, 12]. ALD was developed in the early 1970s as a chemical gas phase thin film deposition method based on alternative saturative surface reactions [19, 20]. In ALD, the precursors are alternately pulsed into the reaction chamber, separated by purging or evacuation periods. Each pulse of precursor saturates the surface with a sub-monolayer of the precursor. This results in a self-limiting film growth with excellent thickness control and conformality. The surface chemistry of ALD has been recently reviewed in detail [19, 20]. Due to its self-limiting nature, ALD is a promising technique to deposit thin, conformal high-k dielectrics and metal films. ALD has the advantage of excellent thickness control, self-limiting

growth, conformality, and low deposition temperatures. The disadvantage of ALD is its relatively lower throughput compared to PVD and CVD.

As stated earlier, the silicon / high-K interface has been extensively studied and reported for more than 20 years. High level of trapped charges at this interface, Fig. 3, is causing reliability issues that continue to cause many semiconductor companies to be cautious about implementing high-k/metal gate in high volume production. In addition, the effects of this advanced gate stack on analog passive devices also need to be scrutinized. The backend capacitor such as MIM, Plate or MOM/VNCAP capacitors may not be significantly affected by the change in the gate stack materials. The N&P MOS capacitors (N & P MOS Varactors) are more commonly used because they provide higher quality factor and extended tuning range than junction capacitors. However, these capacitors will see significant impact on 1/f noise, linearity and mismatch due to the high Dit at the high-K/silicon substrate interface. The high-K/metal gate process integration and reliability solutions for the transistor may also address those of MOSCAPs. On the other hand, the precision analog gate resistor as shown in Fig. 4 is expected to be significantly affected by the change to metal gate stack.

The typical unsilicided P+ A-silicon gate resistor is now replaced by the unsilicided P^+ A-silicon over metal resistor. Because of the metal layer, the current is expected to flow through it instead of through the A-silicon layer so the resistor characteristics maybe more of those of the metal resistor. Also the added interface – A-silicon/metal gate interface – will need to be characterized for its effects on temperature coefficient (TC) control, temperature / voltage linearity, and 1/f noise. TaN and TiN are two of the most common metal films considered for the metal gate stack. Figure 5 showed the Rs for a typical TiN thin film resistor and the resistance versus temperature. The Rs is approximately 35 Ω/sq, and the TCR is approximately 450 ppm/C. The Rs of metal gate resistor is estimated to be approximately 1/3 to 1/2 of P+ A-silicon gate resistor, in the range of 140 Ω/sq.

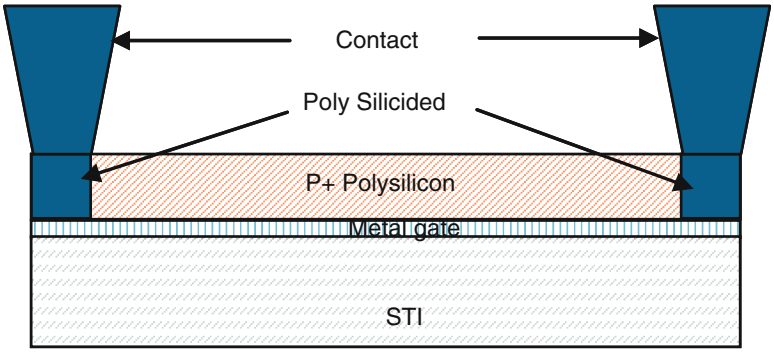


Fig. 4 P+ A-silicon/metal resistor

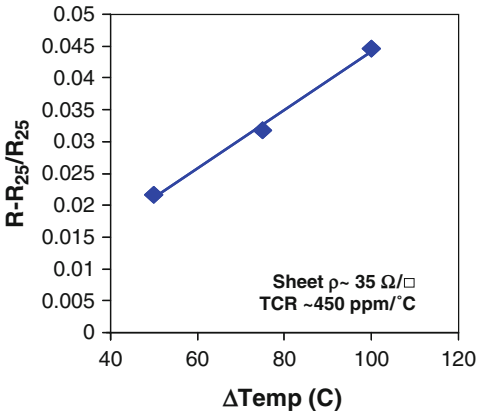


Fig. 5 TiN resistance vs temperature

3 P+ A-Silicon/Metal Resistor

A P+ A-silicon/Metal resistor has two main components: a resistor body composed with a stack of P+ A-silicon/Metal (Rbody) and the end component (Rend) composed of contacts and the A-silicon/Metal layer as shown in Fig. 6 [22]. Because of the low resistivity of the metal layer, the resistor body behavior is dominated by the characteristics of the metal resistor. On the other hand, the characteristics of the P+ A-silicon dominate the Rend behavior (P+ A-silicon thickness of 800 Å). It is well known that metal resistor has positive temperature coefficient (TC) – resistance increased with increasing temperature as shown in Fig. 5 – compared to negative TC for a typical P+ A-silicon resistor [21]. Since there are two competing components Rbody with metal resistor property and Rend with P+ A-silicon resistor property, the P+ A-silicon/Metal resistor behavior depends on the relative contribution of each component into the total resistance.

The zero bias resistance at a given temperature can be described by the following general equation

$$R = \frac{Rs(L+\Delta L)\left[1+(TCRs1\,\Delta T)+\left(TCRs2\,\Delta T^2\right)\right]+2\mathrm{Re}nd\left[1+(TCRn1\,\Delta T)+\left(TCRn2\,\Delta T^2\right)\right]}{W+\Delta W}$$

where R is resistance in Ω, Rs is sheet resistance in Ω/sq, Rend is end resistance in Ω-μm, L is design length in μm, ΔL is length bias in μm, W is design width, ΔW is width bias in μm, TCRs1 is linear body temperature coefficient, TCRs2 is quadratic body temperature coefficient, TCRn1 is linear end temperature coefficient, TCRn2 is quadratic end temperature coefficient, and ΔT is temperature difference to the nominal temperature (Tnom). The Rs, Rend, ΔL, and ΔW are calculated from the measured zero bias resistance at Tnom.

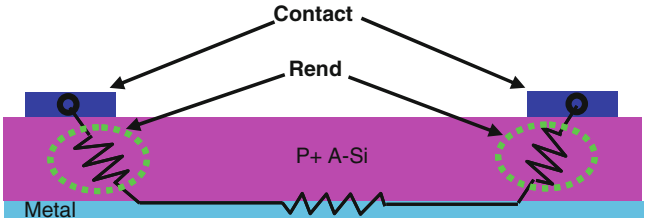


Fig. 6 X-section view of a metal gate resistor electrical behavior

In order to characterize Rs, ΔL, ΔW, and Rend accurately at Tnom, many structures with variation of L and W are required in addition to a wide/long resistor, a narrow/long resistor, a wide/short resistor, and a narrow/short resistor. The parameters can be obtained from optimization, which starts from initial extraction of Rs from the wide/long resistor, ΔL and ΔW from the wide/short resistor and the narrow/long resistor.

All resistors are measured at various temperatures with the four terminal Kelvin set-up to minimize the parasitic resistance from the cable and metal connections to the pads. Current sweeping is preferred to voltage sweeping because the forcing current for each resistor with varying size can be easily calculated from the current density sweep condition for the bias dependence characterization. Rs, ΔL, ΔW, and Rend are optimized from the zero bias resistance values at Tnom, and the temperature coefficients TCRs1, TCRs2, TCRn1, and TCRn2 were extracted from the measured zero bias resistances at various temperatures. In our case the extracted Rs and Rend of the P+ P+ A-silicon/metal gate resistor at 25° C are about 140 Ω/sq and 160 Ω-μm respectively. The die-to-die variation of the Rs and Rend is very tight based on the measurement taken over 54 die locations. The temperature coefficient of the Rend (TCRn1) is estimated to about −1,300 ppm/°C while the temperature coefficient of Rbody (TCR1) is approximately +280 ppm/°C. The negative temperature coefficient for Rend is due to the dominant behavior of the P+ A-silicon layer and the positive temperature coefficient for Rbody is due to the dominant behavior of the metal layer as mentioned before. Figure 7 shows the normalized temperature dependence of the resistance on the resistor length L. As expected, the temperature behavior of P+ A-silicon/metal resistors shows a strong dependence of the length because the contribution of Rbody and Rend changes with the length. For short length devices, the resistance is dominated by the characteristics of Rend, showing negative temperature coefficient. As the length increases, the resistance becomes more dominated by the characteristics of Rbody, showing positive temperature coefficient. Figure 8 shows the normalized temperature dependence of the resistance on the resistor width W. As the width increases the resistor becomes more dominated by the end characteristic because of higher value of Rend and shows more negative temperature coefficient. The temperature coefficient of P + A-silicon/Metal resistor depends on the device size and the modeling of temperature behavior for an arbitrary device size is challenging. It is common practice that

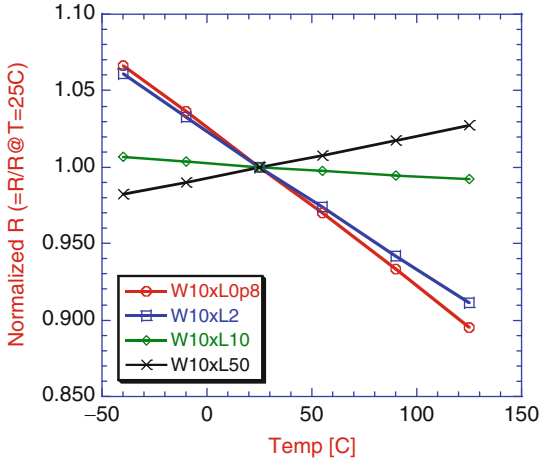


Fig. 7 Normalized R over temperatures for varying resistor L

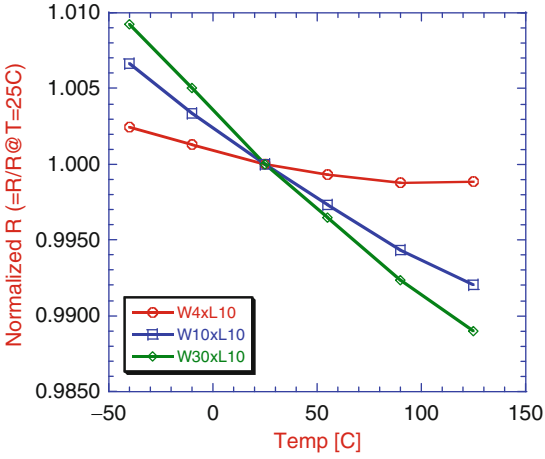


Fig. 8 [Normalized R over temperatures for varying resistor W

Ge dopants are employed in P-type transistor channel to improve mobility. Figure 9 shows that the normalized P+ A-silicon/metal gate resistance with Ge has very similar temperature and length dependence as those without Ge.

The bias dependence of a resistor in circuit simulation is described as

$$R = R0\left(1 + VCR1\,V + VCR2\,V^2\right)$$

where R0 is the zero bias resistance, VCR1 is the first order voltage coefficient of resistance, VCR2 is the second order voltage coefficient of resistance, and V is applied voltage. For a semiconductor resistor the voltage dependence of resistance comes from self-heating, velocity saturation and depletion of the semiconductor layer due to the bias. For a P+ A-silicon/metal resistor, self-heating and velocity

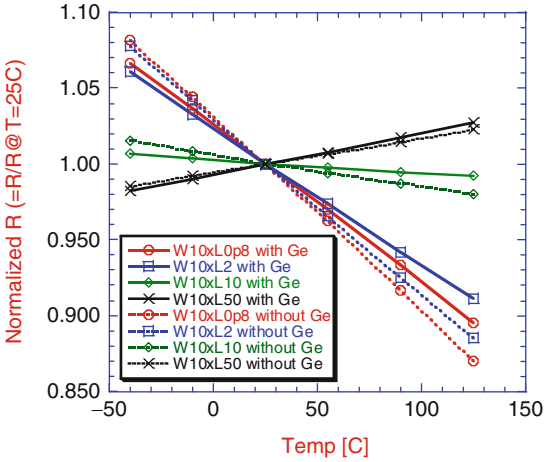


Fig. 9 Normalized R over temperatures for resistors with and without Ge

saturation are the sources for the voltage dependence of resistance. The self-heating effect increases the temperature of a resistor and the resistance changes according to the temperature coefficient of the resistor: a resistor with negative temperature coefficient decreases its resistance and a resistor with positive temperature coefficient increases its resistance. On the other hand, the velocity saturation effect increases its resistance as the voltage for the resistor increases. Two effects compete in a resistor with negative temperature coefficient and add together in a resistor with positive temperature coefficient. Figure 10 shows bias dependence of resistance for resistors of various lengths. Figure 10a shows the normalized resistance at various temperatures for a short resistor. The self-heating effect is strong for all biases and the resistance decreases as the bias increases because the resistor has negative temperature coefficient. Figure 10c shows the normalized resistance at various temperatures for a long resistor which has positive temperature coefficient. The resistance increases as the voltage across the resistor increases because of self-heating and velocity saturation at high biases.

Figure 10b shows the normalized resistance at various temperatures for a intermediate length resistor. The resistor has negative temperature coefficient, and as the temperature of the resistor increases due to self-heating, the resistance decreases according to the negative temperature coefficient. However, when the bias is high enough where velocity saturation effect overcomes the self-heating effect, the resistance starts to increase as the bias increase.

Recently, more sophisticated and physics-based models have been reported by adding the geometry dependence of voltage coefficient [23], and by including nonlinearities from velocity saturation and self-heating [24, 25]. The voltage dependence of the resistance can also be expressed in terms of field (E) instead of voltage (V) for scalability of the model [24, 25]. However, the modeling of the voltage dependence for a P+ A-silicon/metal resistor with an arbitrary size is very challenging. Since accurate modeling of voltage dependence of a resistor is

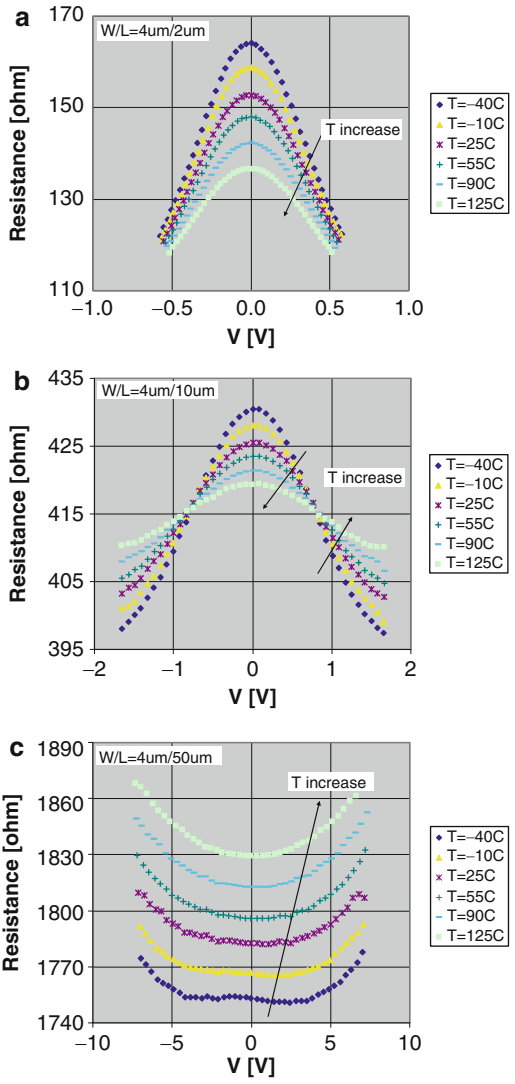


Fig. 10 Measured R versus bias over temperatures for varying resistor L: (**a**) W/L = 4/2 μm, (**b**) W/L = 4/10 μm, (**c**) W/L = 4/ 50 μm

important for analogue and RF applications, the device sizes of P+ A-silicon/metal resistors may have to be limited to a narrow range.

The resistor noise is modeled using thermal and flicker (1/f) noise components. The flicker noise depends on DC current through the resistor and scales with geometry. Its formula is given by the following equation

$$\frac{\overline{i_f^2}}{\Delta f} = \frac{kf \cdot I^{af}}{L \cdot W \cdot f^{f\,\exp}}$$

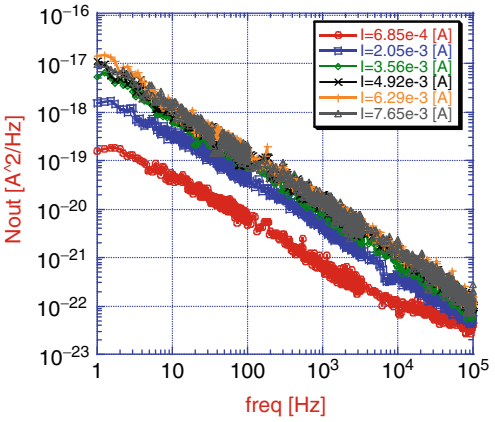


Fig. 11 Noise current spectral density at different biases for a 700 Ω metal gate resistor

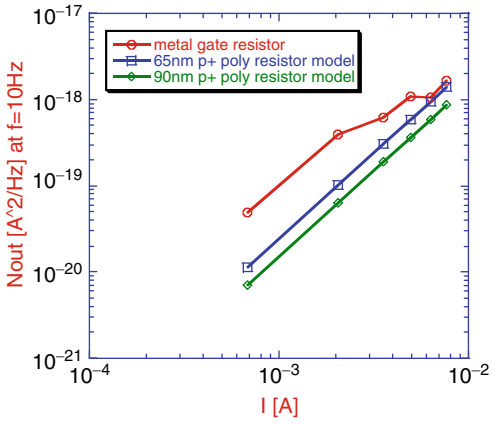


Fig. 12 Comparison of noise current spectral density at f = 10 Hz for a W/L = 10/50 μm metal gate resistors. The 65 nm P+ poly resistor model has Rsh ∼ 710 Ω/sq and the 90 nm P+ poly resistormodel Rsh ∼ 440 Ω/sq

where f is frequency in Hz, af is flicker noise exponent, fexp is flicker noise frequency exponent, kf is flicker noise coefficient, and I is the DC current in the resistor. Figure 11 shows a result of the noise current spectral density of a 700 Ω P+ A-silicon/metal gate resistor at different applied biases for the frequency range from 1 Hz to 100k Hz. The noise spectral density shows the characteristic of 1/f noise: the noise spectral density decreases inversely proportional to frequency. As expected, the noise increases with increasing bias.

Although the 1/f noise of a metal resistor is very low, for a P+ A-silicon/metal resistor the traps located at the grain boundary in P+ A-Si and the interface traps between P+ A-Si and metal can induce noise by emitting and capturing carriers [26]. In Fig. 12, the low frequency noise of a P+ A-silicon/metal gate resistor at 10 Hz is compared with the simulated noise from two resistor models: 65 and 90 nm poly resistor models. The noise simulation was done for the same size as the P+ A-silicon/metal resistor at the same DC current.

It was observed that the P+ A-silicon/metal gate resistor exhibited higher noise with slightly different bias dependence than the P+ polysilicon resistors from the other two technologies. This shows that the interface between P+ A-Si and metal plays an important role for the noise behavior.

4 Summary

At the 32 nm technology node, the change from a P+ polysilicon / SiON gate stack to a P+ A-silicon/metal gate/high-K dielectric will have a significant impact on RF passives, especially the MOSCAPs and gate resistors.

The low sheet resistance value of the metal gate resistor will cause its size to increase. As shown in Figs. 7, 8, and 9, the metal gate resistor TC ranges from positive values for resistors with longer length and narrower width to negative values for resistors with shorter length and wider width. This strong temperature performance dependence on L&W may limit the L&W ranges of analog metal gate resistor designs. The resistor's 1/f noise may depend on the quality of the A-silicon to metal gate interface. An in-situ atomic layer deposition (ALD) process may be used for metal and A-Si deposition to ensure that the lowest amount of charge is trapped at the metal/A-Si interface to improve 1/f noise and linearity characteristics. The metal gate resistor's TC trend was found to be similar for silicon substrates with or without Ge.

Acknowledgement This work has been supported by the Bulk CMOS technology development project at the IBM Microelectronics, Div. Semiconductor Research & Development Center, Hopewell Junction, NY 12533.

References

1. R.B. Staszewski et al., IEEE 2006 Custom Integrated Circuits Conference, pp. 517–522
2. W.M. Elgharbawy et al., IEEE Circ. Syst. Mag. 4th Quarter 2005, 8 (2005)
3. V. George, V. Jahagirdar, S. Chao Tong, K. Smits, S. Damaraju, S. Siers, V. Naydenov, T. Khondker, S. Sarkar, P. Singh, 2007 IEEE Asian Solid-State Circuits Conference, pp. 14–17
4. P. Sivasubramani et al., VLSI Technology, Digest of Technical Papers, pp. 68–69
5. V. Narayanan et al., VLSI Technology, 2006. Digest of Technical Papers, pp. 178–179
6. D.H. Triyoso et al., Appl. Phys. Lett. **88**, 222901–222903 (2006)
7. D.H. Triyoso et al., J. Vac. Sci. Technol. B. **25**, 845–852 (2007)
8. R.I. Hegde et al., IEEE International Electron Devices Meeting 2005, pp. 39–41
9. K. Iwamoto et al., Jap. J. Appl. Phys. **46**, 7666–7670 (2007)
10. W. Taylor et al., IEEE International Electron Devices Meeting 2006, pp. 11–13
11. C.Y. Kang et al., 2008 IEEE International Reliability Physics Symposium (IRPS), pp. 663–664
12. A. Delabie et al., J. Vacuum Sci. Technol. A **25**, 1302–1308 (2007)

13. J.K. Schaeffer et al., ECS Transactions (submitted)
14. H.-C. Wen et al., 2007 VLSI Technology, Digest of Technical Paper, pp. 160–161
15. R. Harris et al., 2007 VLSI Technology, Digest of Technical Paper, pp. 154–155
16. M. Thomas et al., 2007 VLSI Technology, Digest of Technical Paper, pp. 58–59
17. Y.-K. Jeong et al., 2004 VLSI Technology, Digest of Technical Papers pp. 222–223
18. C. Chiang et al., VLSI Technology 2005 Digest of Technical Papers, pp. 62–63
19. M. Ritala, M. Leskela, in *Handbook of Thin Films Materials*, ed. by H.S. Nalwa (Academic Press, San Diego, CA, 2001), p. 103
20. R. Puurunen, J. Appl. Phys. **98**, 121301 (2005)
21. *The Electrical Engineering Handbook*, 2nd edn. Passive components. (CRS Press & IEEE Press, Boca Raton, FL/New York, 1997), p. 5
22. I. Aureli et al., 2007 International Conference on Microelectronic Test Structures, pp. 268
23. Seok Yong Ko, Jin Soo Kim, Gwang Hyeon Lim, Sung Ki Kim, A new P+ A-siliconsilicon resistor model considering geometry dependent voltage characteristics for the deep sub-micron CMOS process. 2006 International Conference on Microelectronic Test Structures, pp. 27–30
24. r2_cmc: Two-Terminal Nonlinear Resistor Model, available at http://www.eigroup.org/ cmc/downloads/r2_cmc/r2_cmc_v1.0_r0.0_2005nov12.pdf
25. r3_cmc: Three-Terminal Nonlinear (Diffused and P+ A-silicon-Silicon) Resistor Model and JFET Model, available at http://www.geia.org/GEIA/files/ccLibraryFiles/Filename/000000003012/r3_cmc_release1.0.0_2007Jun12.pdf
26. R. Brederlow, W. Weber, C. Dahl, D. Schmitt-Landsiedel, R. Thewes, A physically based model for low-frequency noise of P+ A-silicon-silicon resistors. Tech. Digest IEDM, 1998, pp. 89–92

Part IV
Reliability and SEU

Threshold Voltage Shift Instability Induced by Plasma Charging Damage in MOSFETS with High-K Dielectric

Koji Eriguchi, Masayuki Kamei, Kenji Okada, Hiroaki Ohta, and Kouichi Ono

1 Introduction

With the decrease in dimensions of ULSI circuits in accordance with the scaling law [1], the electrical thickness of gate dielectric materials should be decreased. Dielectric materials with higher dielectric constant (high-k) should replace conventional SiO_2. Hafnium-based gate stacks have been proposed to be one of the most promising candidates, although reliability issues are being discussed [2–4]. On the other hand, as the critical dimension of feature size in devices has shrunk and new materials have been introduced, plasma-induced damage (PID) have been pointed out [5–8]. PID is a significant reliability issue for high-k gate dielectrics as long as plasma is used for device fabrication and has been studied by many groups [8–10]. However, there have been few comprehensive studies of PID to high-k devices caused by different plasma sources so far, in particular, of the charging polarity during plasma processing, extensively studied during the last two decades for SiO_2. Consideration of the charging polarity has recently been recognized as one of indispensable guidelines in designing an antenna rule [11] for different devices (n- or p-channel) to prevent a yield loss [12]. In order to understand the PID mechanisms, the charging polarity to high-k devices should be investigated in detail. The purpose of this paper is to address the polarities of charging damage to high-k dielectric as well as those of charging stress driven by plasmas, by focusing on the effects of the plasma source type (Ar- and Cl-based gas mixture) and device type (n- or p-channel) on performance degradation. Threshold voltage shift instability induced by plasma charging damage will be reported and the implication will be discussed.

K. Eriguchi (✉), M. Kamei, H. Ohta,and K. Ono
Graduate School of Engineering, Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan
email: eriguchi@kuaero.kyoto-u.ac.jp

K. Okada
MIRAI-AIST 16-1, Onogawa Tsukuba Ibaraki, 305-8569, Japan

A. Amara et al. (eds.), *Emerging Technologies and Circuits*,
Lecture Notes in Electrical Engineering 66, DOI 10.1007/978-90-481-9379-0_7,

2 Experimental

N- and p-channel MOS devices were fabricated on a p-type Si (100) substrate using a conventional CMOS process technology. A high-k gate stack ($HfAlO_x$/SiO_2 = 6.4/1.0 nm), denoted as "High-k", and thermally grown SiO_2, denoted as "SiO_2", with approximately similar physical thicknesses (~7 nm) were used as gate dielectric materials. The $HfAlO_x$ film was deposited by atomic layer deposition on thermally grown SiO_2 [4]. The electrical thicknesses determined by capacitance-voltage (C-V) measurements were ~2.7 and ~7.4 nm for High-k and SiO_2, respectively. The gate width (W) and length (L) of MOSFETs were W/L = 1/1, 10/10 and 100/100 μm, giving gate areas of 1, 100 and 10,000 μm^2, respectively. Al interconnects were employed. As shown in Fig. 1, Al-probing pads with an area of 42,000 μm^2 served as antennas of box-type structures. Current-voltage ($I_g - V_g$, $I_d - V_g$) and C–V measurements were conducted for at least 12 difference devices to evaluate deviation. All the measurements were carried out at room temperature.

Processed samples were mounted on the wafer stage and exposed to electron cyclotron resonance (ECR) plasma sources. The source power was 600 W and the working pressure was 1.0×10^{-2} Torr. Two different Ar/O_2 (= 28/12 sccm) and Cl_2/O_2 (= 20/20 sccm) gas mixtures, denoted as "Ar" and "Cl", respectively, were utilized. A 13.56 MHz-RF bias with a power of 200 W was supplied to the wafer stage. Process durations were 150 and 30 s for Ar and Cl, respectively. The Langmuir probe measurements determined the electron temperature and density for Ar plasma as 1.5 eV and $5.1 \times 10^{10} cm^{-3}$. (In the case of Cl plasma, it was found to be difficult to determine the parameters with high accuracy.)

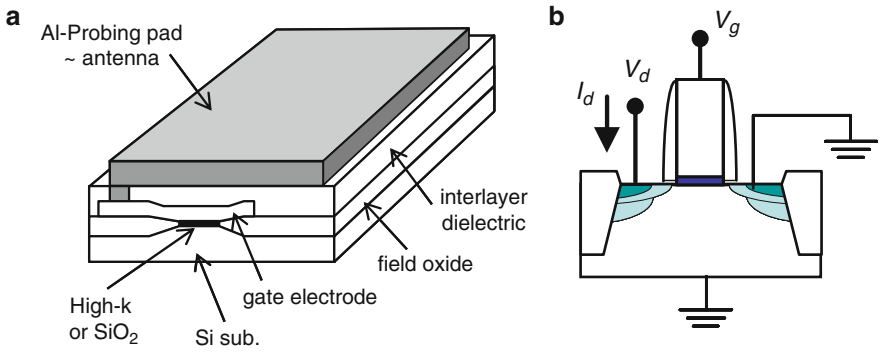


Fig. 1 Test sample structure used in this study: (**a**) cross-sectional view, (**b**) illustration of an electrical measurement setup

3 Results and Discussion

3.1 Gate Leakage Characteristics

Figure 2 shows cumulative probability plots of gate leakage currents for MOSFETs with a gate area of 1 μm^2. Figure 2a shows the results for High-k devices and Fig. 2b, for SiO_2 devices. It is confirmed that charging damage is more severe for Ar-plasma than for Cl-plasma in this study. Not that the physical thicknesses of these devices are approximately similar to each other. These observations are consistent with the flat band voltage shift determined by C–V measurement (not shown here). Thus, it can be concluded that Ar-plasma induced more severe damage to both devices.

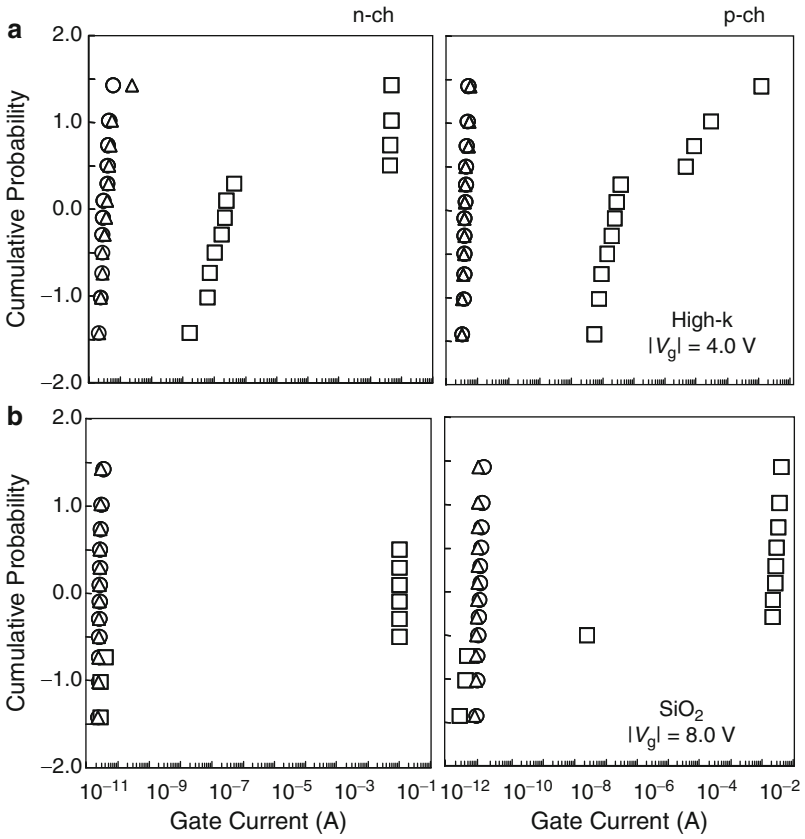


Fig. 2 Cumulative probabilities of gate leakage current for (**a**) high-k gate stack ($HfAlO_x/SiO_2$) and (**b**) SiO_2. The left is for n-channel MOS devices, and the right, for p-channel MOS devices. *Circles*, control; *squares*, Ar-plasma exposure; *triangles*, Cl-plasma exposure

3.2 MOSFETs Characteristics

Figure 3a and b show typical examples of drain current-gate voltage (I_d-V_g) characteristics for MOSFETs with High-k and SiO_2, respectively. One can find that, for Ar-plasma exposure case, the threshold voltage (V_t) shifts in MOSFETs with High-k are toward the opposite directions between n- and p-channel MOSFETs (a more positive over-drive voltages are required to obtain the same drain current), while those in SiO_2 devices, toward the same directions between n- and p-channel MOSFETs (a more positive over-drive voltage is required for n-channel MOSFETs and a more negative over-drive voltage, for p-channel MOSFETs).

Regarding the V_t shifts in MOSFETs with SiO_2, the present observations are consistent with a previous report [13] It is also observable that the direction of the V_t shift is opposite in the case of MOSFETs with High-k exposed to Ar- and

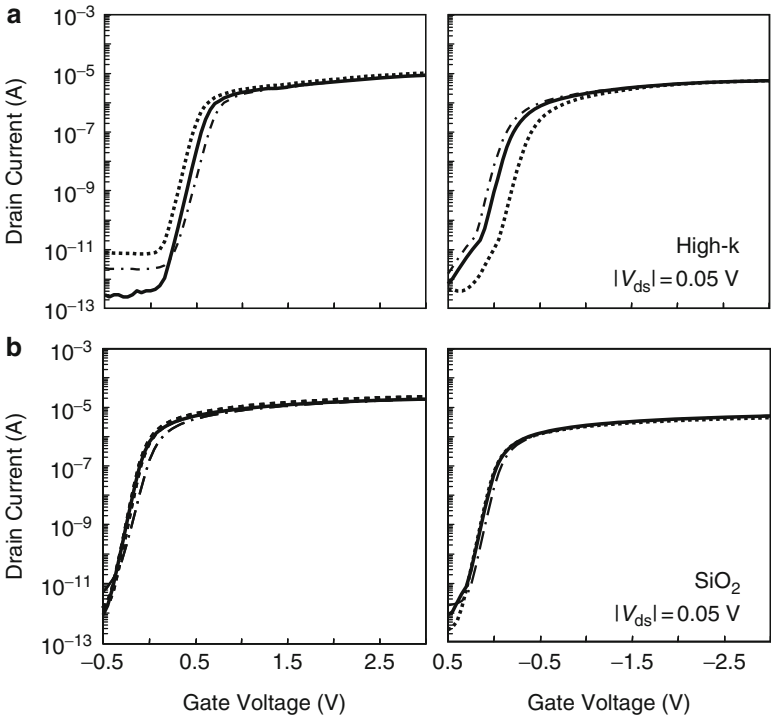


Fig. 3 Drain current versus gate voltage characteristics for n- and p-channel MOSFETs with (**a**) high-k gate stack and (**b**) SiO_2. The gate length and width are 100 and 100 μm, respectively. The left is for n-channel MOSFET, and the right, for p-channel MOSFET. *Solid lines* indicate the samples without plasma exposure, *dotted lines*, those with Cl-plasma exposures, and *dash-dotted line*, those with Ar-plasma exposures

Cl-plasmas, respectively, in contrast to the previous report [9]. The results are summarized in Fig. 4.

In Fig. 4, the results for different test structures ($W/L = 10/10$ and 100/100 μm, i.e., the antenna ratio) are also compared. From this figure, several key features are found;

1. In the case of High-k, the direction of V_t shift in the case of Ar-plasma exposure is opposite to that of Cl-plasma exposure for both test structures, while in the case of SiO_2, the direction is the same.
2. From the V_t shifts for SiO_2, Ar-plasma exposure is confirmed to induce a larger damage than Cl-plasma exposure since the larger V_t shift is observed for both n- and p-channel MOSFETs in the case of Ar-plasma exposure. Note that this is also confirmed from Fig. 2.
3. On the basis of the above discussions, it can be concluded that MOSFETs with High-k exhibit the transition from the negative V_t shift to the positive as the charging damage becomes more severe. That is, MOSFETs with High-k suffer from the positive charge trapping by the Cl-plasma (lower charging damage) and from the negative charge trapping, by the Ar-plasma. It is worthy to note that Young et al. [9] observed the V_t shift similar to the present result for Ar-plasma exposure. This may be due to the long plasma exposure in their case, resulting in severe charging damage, which is consistent with some of the present observations.

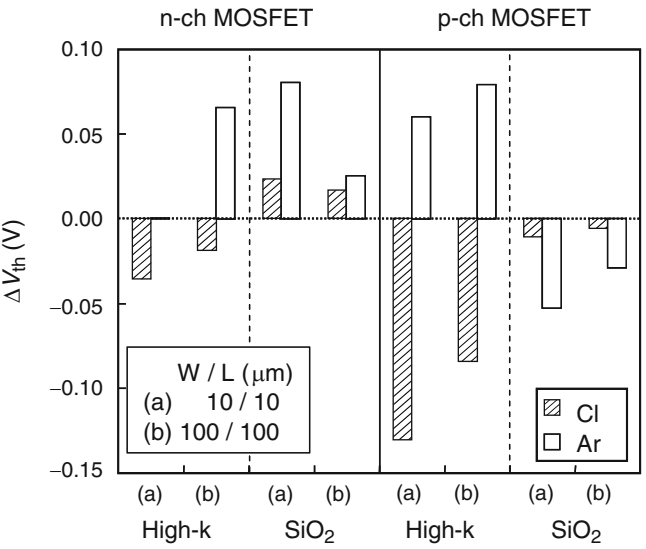


Fig. 4 Summary of threshold voltage shifts for Ar- and Cl-plasma exposures. Two different device structures are compared

3.3 Mechanisms of Charging Polarity and Vt Shift Instability

In order to clarify the mechanisms underlying our findings mentioned above, we conducted a post-process electrical stress test. Since plasma is considered to be a constant-current source in terms of charging damage [10, 14–17], the constant-current stress (CCS) test was performed to simulate the charging damage observed in this study. The range of injected stress currents in the test is determined from the gate areas of test sample structures and Al-probing pads ($\sim$ the antenna ratio), and the ion fluxes expressed by the Bohm sheath criterion [18], i.e., $0.61\, n_{\mathrm{ion}} \sqrt{T_{\mathrm{e}}/M}$, ($n_{\mathrm{ion}}$ is the ion density, T_{e} is the electron temperature, and M is the mass of ion) on the assumption that the ion density is the same as the electron density. In this study, n_{i} and T_{e} is determined from the Langmuir probe measurement as mentioned in the above. To understand the difference in the direction of V_{t} shift (V_{t} shift instability) observed in Figs. 3 and 4, the time evolutions of applied gate voltage under various current stresses for n- and p-channel MOSFETs were investigated in detail. In addition, in order to understand the polarity of plasma-driven stress, the decrease in a linear region peak transconductance (Δg_{mmax}) versus the V_{t} shift under CCS is investigated in detail for n- and p-channel MOSFETs.

Figure 5 shows typical examples of the time evolution of applied gate voltage under various current stresses for n- and p-channel MOSFETs. Clear time dependence of applied gate voltage is observed in the course of stress time, and the dependence is larger in the case of MOSFETs with High-k. This is due to the larger carrier trapping cross section of High-k. The negative gate voltage shift primarily corresponds to the hole trapping process and the positive shift, to the electron trapping process [19, 20]. At the early stage of stressing, hole trapping is obvious, and with stress time electron trapping becomes predominant, which is consistent

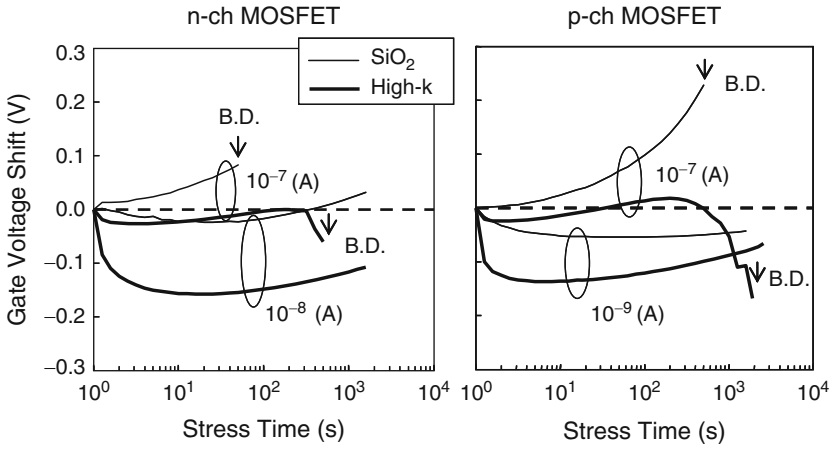


Fig. 5 Time evolution of applied gate voltage under constant-current stress for High-k and SiO_2 devices. The left is for n-channel MOS devices, and the right, for p-channel MOS devices

with the previous observations on high-k gate dielectrics [4]. This trend is more obvious for High-k and for lower stress current density, as observed in these figures. Thus, we speculate that from the viewpoint of device characteristics change, the plasma-source-dependent charging polarity observed in the Cl- and Ar-plasma exposures for High-k (Figs. 3 and 4) is attributed to (1) the stress current density-dependent charge trapping phenomenon and/or (2) the time-dependent charge trapping phenomenon [17], as conducted in Fig. 5. With regard to the stress current density dependence, it is expected that the lower stress current density caused by Cl-plasma results in hole trapping predominance for degradation in High-k, and the higher stress current caused by Ar-plasma, in the electron trapping. With regard to the time dependence, one might expect the positive V_t shift when the High-k devices are exposed to Cl-plasma for a longer time, which may be in consistent with the previous report by Young et al. [9].

Figure 6 shows the decrease in a linear region peak transconductance (Δg_{mmax}) versus the V_t shift (ΔV_t) under CCS for (a) SiO_2 and (b) High-k devices. In the

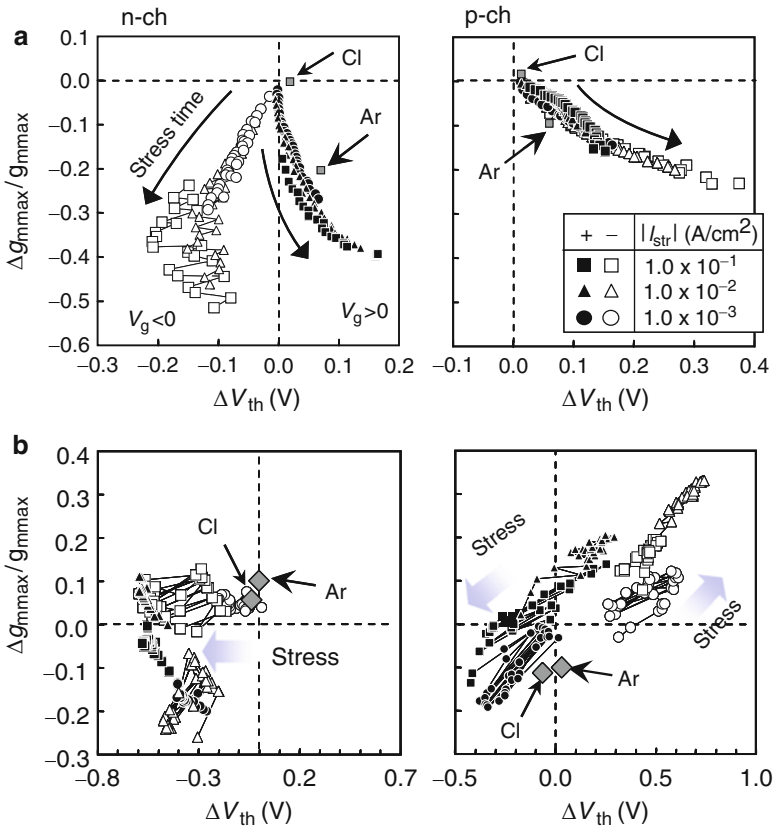


Fig. 6 Time evolution of transconductance change versus threshold voltage shift for (**a**) SiO_2 and (**b**) high-k devices under various stress current densities. The left is for n-channel MOS devices, and the right, for p-channel MOS devices

course of stress time, Δg_{mmax} decreases for both SiO_2 and High-k devices in the positive and negative gate bias cases. For the n-channel SiO_2 device case, it is clear that the $\Delta g_{mmax} / g_{mmax}$ and the direction of ΔV_t depend on the applied gate bias polarity (V_g): In the case of positive gate bias, the V_t shifts toward the positive direction, while in the case of negative, toward the negative. This indicates that ΔV_t and the direction observed for plasma-damaged devices may be a fingerprint of plasma-driven stress polarity. In the case of p-channel device, one can see no clear dependence of the direction on the stress polarity.

For the p-channel High-k device case, one can see the stress polarity-dependent V_t shifts: In the case of positive gate bias, the V_t shifts toward the negative direction, while in the case of negative, toward the positive. The data for plasma damaged samples fall on the trend in the positive gate bias case for the n-channel SiO_2 and p-channel High-k devices, while, that in the negative, for n-channel High-k. Thus, it can be speculated that the polarity of plasma-driven stress is dependent on device structures (High-k or SiO_2) as well as plasma sources [21, 22] as schematically summarized in Fig. 7. Although we have no positive answer to this mechanism at present, these structure- and plasma-source dependent charging polarities are induced by a layout of source, drain, and well surrounding gate stack structures [23].

3.4 *Impacts on Damage Characterization of High-k Devices*

As mentioned in Figs. 3 and 4, MOSFETs with high-k suffer from the negative or positive V_t shift in accordance with plasma sources, i.e., an amount of charging damage. Thus, one should be aware that, in characterizing a plasma charging

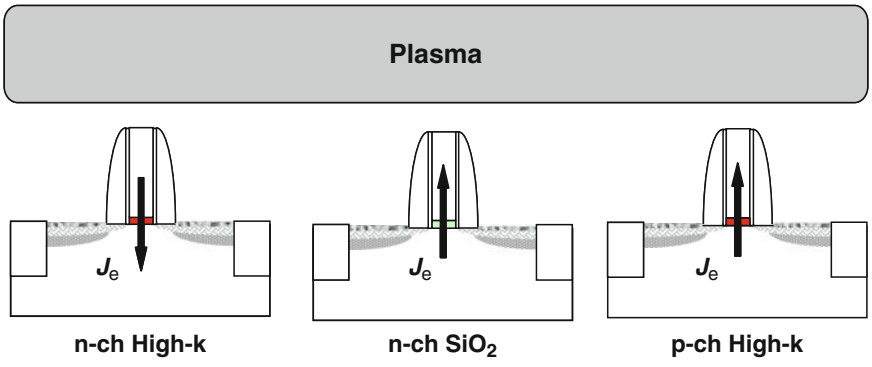


Fig. 7 Difference in polarity of plasma-driven stress to devices deduced from the post-process stress tests conduced in the present study

damage of MOSFETs with high-k, a use of V_t shift might result in erroneous conclusion since a result of "no V_t shit" between Ar- and Cl-plasma exposures may be expected for a certain plasma treatment even if the device was degraded. In another scenario, one may see a process-time dependent V_t shift instabilities in the same plasma equipment, i.e., a longer process duration results in a smaller V_t shift as deduced from Fig. 5. Moreover, a conflicting result between gate leakage increase and V_t shift may be expected as previously discussed for two time-dependent dielectric breakdown test results for thick SiO_2 [14, 15, 22]. The plasma charging damage characterizations based on various methods are necessary for MOSFETs with high-k dielectrics.

4 Conclusion

We found that high-k devices are more susceptible to plasma charging, than SiO_2 devices and that the charging polarity strongly depends on plasma sources such as Ar- and Cl-based gas mixtures. Also we identified a unique charging polarity,i.e.,the V_t shift instability. High-k devices are found to exhibit the opposite charging polarity in accordance with the plasma sources, in contrast to SiO_2, primarily owing to the characteristic hole/electron trapping phenomena: The direction of V_t shift itself depends on a plasma source and/or a amount of plasma charging damage. The present experimental observations provide important implications; the plasma source-dependent charging polarity for high-k devices should be considered. Damage characterizations by various techniques should be conducted in future plasma process design and device design rules for accurate plasma charging damage evaluation.

Acknowledgement We thank Dr. H. Ota of MIRAI-AIST, Dr. T. Nabatame of MIRAI-ASET, and Professor A. Toriumi of the University of Tokyo for fruitful discussions. This work was supported in part by the NEDO/MIRAI project and a Grant-in-Aid for the Twenty-First Century COE from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

1. SIA, The International Technology Roadmap for Semiconductors 2007 update, 2007
2. C. Rino, S.C. Song, C.D. Young, B. Gennadi, L. Byoung Hun, Charge trapping and detrapping characteristics in hafnium silicate gate dielectric using an inversion pulse measurement technique. Appl. Phys. Lett. **87**, 122901 (2005)
3. S. Zafar, Statistical mechanics based model for negative bias temperature instability induced degradation. J. Appl. Phys. **97**, 103709 (2005)
4. K. Okada, H. Ota, T. Nabatame, A. Toriumi, Dielectric breakdown in high-k gate dielectrics – mechanism and lifetime assessment, *Proceedings of the International Reliability Physics Symposium*, 2007, pp. 36–43
5. Y. Yoshida, T. Watanabe, Gate breakdown phenomena during reactive ion etching process, *Proceedings of the Symposium of Dry Process*, 1983, pp. 4–7

6. A.T. Krishnan, V. Reddy, S. Krishnan, Impact of charging damage on negative bias temperature instability, IEDM Technical Digest, 2001, pp. 865–868
7. K. Eriguchi, K. Takahashi, K. Ono, plasma-induced damage and its impacts on the reliability of advanced semiconductor devices. *Proceedings of the 6th International Conference Reactive Plasmas and 23rd Symposium. Plasma Processing*, 2006, pp. 5–6
8. K. Eriguchi, M. Kamei, D. Hamada, K. Okada, K. Ono, A comparative study of plasma source-dependent charging polarity in MOSFETs with high-k and SiO2 gate dielectrics, Ext. Abs. Solid State Dev. Mat., 2007, pp. 722–723
9. C.D. Young, G. Bersuker, F. Zhua, K. Matthewsb, R. Choi, S.C. Song, H.K. Parkc, J.C. Lee, B.H. Leed, Comparison of plasma-induced damage in SiO2/TiN and HfO2/TiN gate stacks. *Proceedings of the International Reliability Physics Symposium*, 2007, pp. 67–70
10. W.T. Weng, A.S. Oates, T.-Y. Huang, A comprehensive model for plasma damage enhanced transistor reliability degradation. *Proceedings of the International Reliability Physics Symposium*, 2007, pp. 364–369
11. V. Shukla, V. Gupta, C. Guruprasad, G. Kadamati, Automated antenna detection and correction methodology in VLSI designs. *Proceedings of the International Symposium on Plasma Process-Induced Damage*, 2003, pp. 158–161
12. T.B. Hook, D. Harmon, W. Lai, Gate oxide damage and charging characterization in a 0.13 mm, triple oxide (1.7/2.2/5.2 nm) bulk technology. *Proceedings of the International Symposium on Plasma Process-Induced Damagei*, 2002, pp. 10–13
13. K.P. Cheung, C.P. Chang, Plasma-charging damage: A physical model. J. Appl. Phys. **75**, 4415–4426 (1994)
14. K. Eriguchi, Y. Uraoka, New method for lifetime evaluation of gate oxide damaged by plasma processing. IEEE Electron Device Lett. **16**, 187–189 (May 1995)
15. K. Eriguchi, Y. Kosaka, Correlation between two time-dependent dielectric breakdown measurements for the gate oxides damaged by plasma processing. IEEE Trans. Electron Devices **45**, 160–164 (Jan 1998)
16. K. Eriguchi, Y. Uraoka, H. Nakagawa, T. Tamaki, M. Kubota, N. Nomura, Quantitative evaluation of gate oxide damage during plasma processing using antenna-structure capacitors. Jpn. J. Appl. Phys. **33**, 83–87 (1994)
17. K. Eriguchi, T. Yamada, Y. Kosaka, M. Niwa, Impacts of plasma process-induced damage on ultra-thin gate oxide reliability. Proceedings of the International Reliability Physics Symposium, 1997, pp. 178–183
18. M.A. Lieberman, A.J. Lichtenberg, *Principles of Plasma Discharges and Materials Processing*, 2nd edn. (Wiley-Interscience, Hoboken, NJ, 2005)
19. I.-C. Chen, S.E. Holland, C. Hu, Electrical breakdown in thin gate and tunneling oxides. IEEE Trans. Electron Devices **ED-32**, 413–422 (1985)
20. K. Eriguchi, M. Niwa, Temperature and stress polarity-dependent dielectric breakdown in ultrathin gate oxides. Appl. Phys. Lett. **73**, 1985–1987 (1998)
21. T. Yamada, K. Eriguchi, Y. Kosaka, K. Hatada, Impacts of antenna layout enhanced charging damage on MOSFET reliability and performance. IEDM Technology Digest, 1996, pp. 727–730
22. Y. Kosaka, K. Eriguchi, T. Yamada, Stress mode of gate oxide charging during the MERIE and the ICP processing and its effect on the gate oxide reliability. *Proceedings of the International Symposium on Plasma Process-Induced Damage*, 1998, pp. 209–212
23. M. Kamei, K. Eriguchi, K. Okada, K. Ono, Investigation of junction characteristics of MOSFETs with high-k gate stack by plasma processing. *Proceedings of the International Conference on Integrated Circuit Design and Technology*, 2007, pp. 117–120

Analysis of SI Substrate Damage Induced by Inductively Coupled Plasma Reactor with Various Superposed Bias Frequencies

Y. Nakakubo, A. Matsuda, M. Kamei, H. Ohta, K. Eriguchi, and K. Ono

1 Introduction

Plasma-induced Si substrate damage has become one of the critical issues in advanced MOSFETs with shallower junction in source/drain (S/D) extension regions, since the damaged layer thickness will be in conflict with the device design margin such as junction depth [1]. This Si substrate damage is realized as "Si recess" [2] as shown in Fig. 1. Although Si recess is considered to induce dopant loss and performance degradation in MOSFETs, few attentions has been paid to suppress Si recess from the viewpoint of plasma design. Moreover, one can easily observe a constant thickness of Si recess in ultra-scaled MOSFETs recently presented at many conferences [3–5]. It is worthy to note that the thickness is considered to be governed by plasma parameters such as ion energy distribution function (IEDF), electron temperature and plasma density [6]. In order to understand the mechanism and to control the damage, the plasma-induced defects in Si surface layer should be quantitatively estimated, and then, plasma designs should be optimized. Defect generation probability was proposed from an optical analysis as a measure of the damage [7], on one hand. With regard to plasma design, on the other, a plasma source driven by the superimposed dual bias frequency was reported in order to control an IEDF [8] and believed to be a promising candidate for future plasma processes. Recently the effects of superposed bias-frequency on the defect generation probability have been preliminary reported [9]. In this paper, the effect of bias power with the superposed configurations on defect generation process will be investigated in detail from various aspects. The structure of interfacial damaged layer (IL) will be analyzed by spectroscopic ellipsometry and photoreflectance spectroscopy. Also for the purpose of verifying the above methods and clarifying the interfacial growth and carrier trap site generation in the vicinity of

Y. Nakakubo (✉), A. Matsuda, M. Kamei, H. Ohta, K. Eriguchi, and K. Ono
Graduate School of Engineering, Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan
Email:y.nakakubo@ks7.ecs.kyoto-u.ac.jp

A. Amara et al. (eds.), *Emerging Technologies and Circuits*,
Lecture Notes in Electrical Engineering 66, DOI 10.1007/978-90-481-9379-0_8,

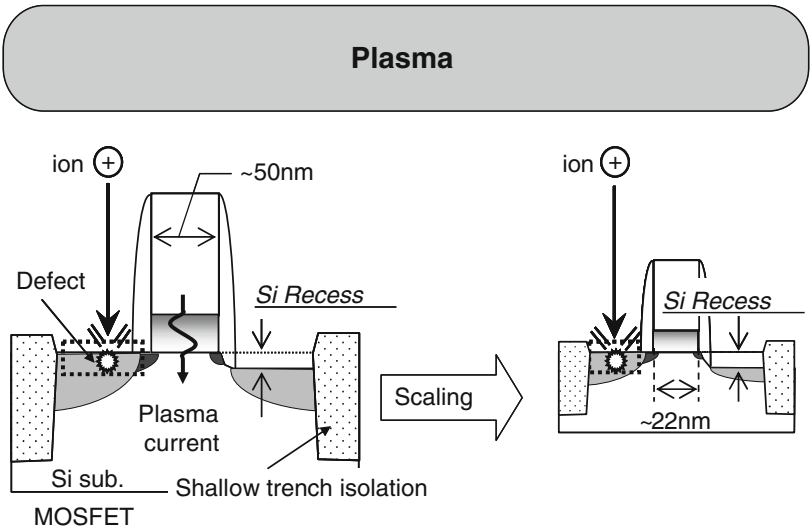


Fig. 1 Schematic illustration of Si substrate damage by energetic ion bombardment usually associated with "Si recess". Although the feature size of MOSFETs has been shrunk dramatically by the scaling law, the thickness of Si recess is believed to be unchanged (see text). Thus, in near future, the thickness of Si recess will be in conflict with device size such as source/drain extension depth

plasma-exposed surface, the result by a capacitance–voltage (C–V) measurement (electrical test) will be discussed.

2 Experimental

2.1 Plasma Exposure with Superposed Bias Frequencies

N-type (100) Si substrates (0.02 Ωcm) with native oxide layer (∼1 nm) were mounted on the stage and exposed to an inductively coupled plasma reactor (ICP). No processed sample was to be the control. Ar gas was utilized for eliminating chemical reactions with Si. The operating pressure was 20 mTorr and a background pressure was less than 3.0×10^{-5} Torr. The substrate temperature was ∼500 K. The Langmuir probe and bias voltage measurements were carried out for plasma diagnostics. As illustrated in Fig. 2, an rf bias with a power of 100 or 200 W in total were applied with the frequency of 13.56 MHz and that of 400 kHz, and their superposed bias through a filter box. Applied source power in the ICP was 300 W with a frequency of 13.56 MHz. Self bias voltage (V_{dc}) and peak-to-peak voltage (V_{pp}) were monitored by an oscilloscope. From the result by plasma monitoring, it was found that V_{dc} which corresponds to the mean impacting ion energy varies with

the superposed bias configuration even if the total bias power was the same, as shown in Fig. 3. Figure 4a–c are the example wave forms of single- and superposed frequencies conducted with ICP reactor shown in Fig. 2. We analyze the plasma

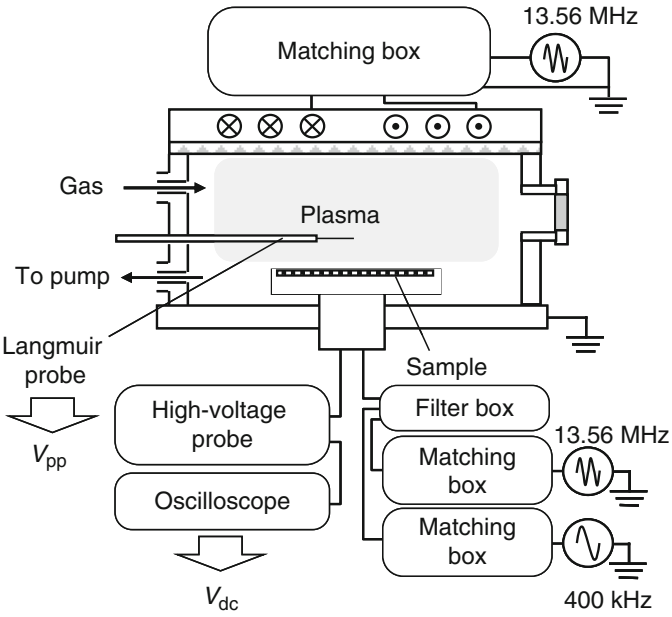


Fig. 2 Inductively coupled plasma reactor used in this study. Two different bias frequencies of 13.56 MHz and 400 kHz can be supplied in this equipment

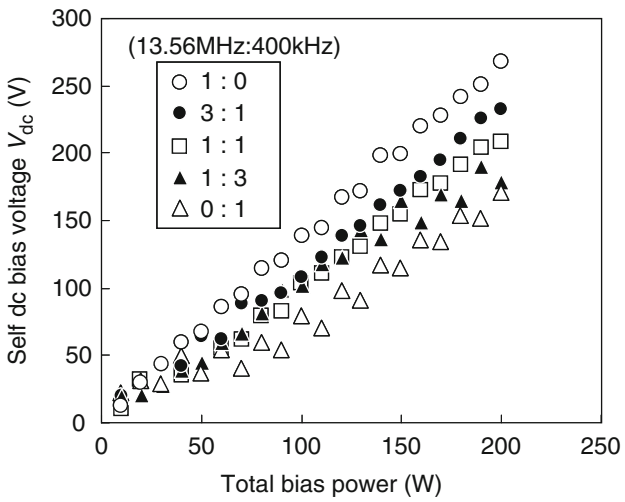


Fig. 3 Measured self dc bias voltage (V_{dc}) as a function of total input bias power with various bias configurations

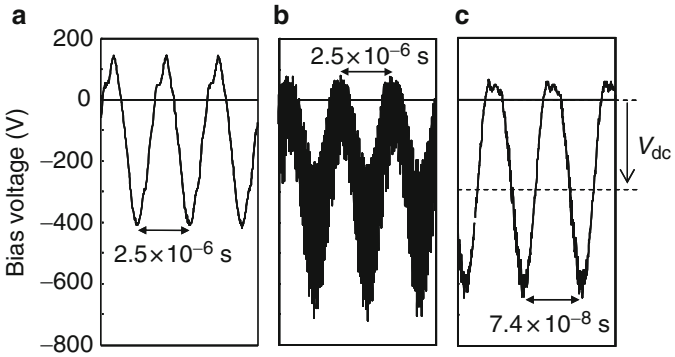


Fig. 4 Wave forms for several bias-frequency configurations; (**a**) Single frequency (400 kHz) of 200 W, (**b**) Superposed frequencies (400 kHz/13.56 MHz) of 200 W in total. The power of 400 kHz and 13.56 MHz was 100 W and 100 W, respectively, **(c)** Single frequency (13.56 MHz) of 200 W

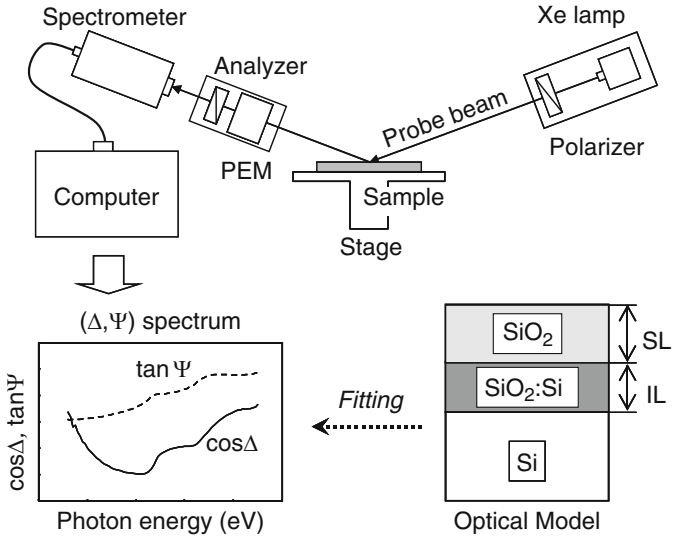


Fig. 5 Illustration of SE analysis setup with an optical model of Si surface structure

damage on silicon substrate surfaces exposed to plasmas with these bias configurations. This will be again discussed later in detail.

2.2 Spectroscopic Ellipsometry

Figure 5 shows an experimental setup of spectroscopic ellipsometry (SE) with an optimized optical model of Si surface structure used in this study. The thickness

and the refractive index of damaged layer were measured by SE before and after plasma exposure. The measurement was performed over the 1.6–5.5 eV range at 0.05 eV intervals. All the SE spectra were taken at an angle of incidence of 75°. For the quantitative analysis of the substrate damage, an optimized optical model (SiO_2/SiO_2:Si/Si Sub.) was introduced [8, 10] as shown. Surface layer (SL) consists of SiO_2, and interfacial layer (IL) consists of SiO_2 and Si, respectively. The interface layer between SiO_2 and Si substrate was introduced as a composite of SiO_2 and *c*-Si with the Bruggeman effective medium approximation (EMA) [11] with the thickness and volume fraction of SiO_2 (f_{SiO_2}) being used as the fitting parameters. Plasma-induced defect sites are considered to be distributed mainly in IL. On the basis of this model, SE determines the thicknesses of SL and IL.

2.3 *Photoreflectance Spectroscopy*

The photoreflectance (PR) spectroscopy was employed to identify a carrier trap site (defect site) density as conducted in [7, 9]. The PR spectroscopy is based on the electroreflectance spectroscopy which measures the change in reflective spectrum modulated by external applied field. In this method, the analysis of the reflective spectrum is based on the mechanism that the change in reflectance due to the change in the permittivity of silicon substrate surface by the photomodulation. Figure 6 shows schematic diagram of the PR analysis setup. The Si substrate surface was modulated by the light from Ar+ ion laser ($\lambda = 514.5$ nm)

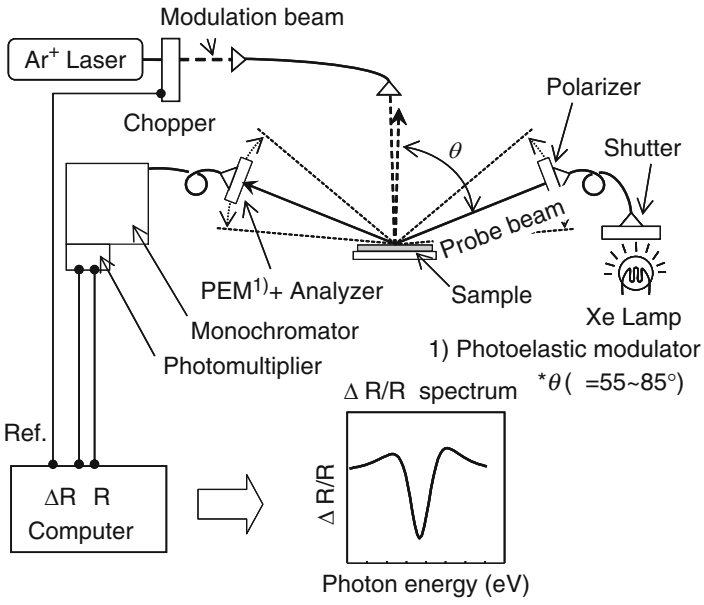


Fig. 6 Illustration of PR analysis setup

chopped with frequency of 500 Hz. A Xe discharge lamp was used as a probe light source. The p polarized probe beam through the polarizer was irradiated. Since the spectral line-shape was found to be clearly observable for an angle of incidence near the pseudo Brewster angle, the incident angle of this probe beam is set at 75°. The measurement was performed over the 2.9–3.9 eV range at 0.01 eV intervals.

For analyzing the structure of IL and the aerial defect site density (N_{dam} (cm^{-2})), the PR technique is used. As described previously [7], the surface potential (V_{s}) was extracted from the spectral line shape ($\Delta R/R$) by the following equations [12],

$$\Delta R/R = \Re\left[Ce^{i\theta}\left(E - E_g + i\Gamma\right)^{-n}\right] \tag{1}$$

$$C = A_1 \ln\left[A_2 I_{\mathrm{p}} \exp(eV_{\mathrm{s}}/kT) + 1\right] \tag{2}$$

where C, θ, E and n are amplitude, phase factor, photon energy and dimension of the critical point associated with the optical transition ($n = 5/2$ in this study),respectively. E_{g} and Γ are the transition energy and the broadening parameter, respectively. From E_{g} in Eq. 1, the strain change in IL by the plasma exposure is estimated with the assumption of a hydrostatic strain model [13]. In Eq. 2, I_{p}, e, k and T are laser power, elementary charge, Boltzmann constant and temperature, respectively. A_1 and A_2 are material and structural dependent parameters described by the following equations.

$$A_1 = \frac{c_1}{\mu_{||}} \frac{\eta kT}{e} \tag{3}$$

$$A_2 = c_2 \frac{e\gamma(1-R)}{A^* T^2 h\nu} \tag{4}$$

where $\mu_{||}$ and η are the interband reduced mass evaluated in the field direction and the ideal factor [14] or ideality factor [15] , respectively. γ, R, A^*, h and ν are the quantum efficiency, reflectance of the probe beam, modified Richardson constant [16], Planck's constant and the frequency of the light, respectively. c_1 and c_2 are constants. In this study, parameters A_1 and A_2 are assumed to be constant.

In order to quantify the plasma-induced defect site density, we focus on the characteristic mechanism that $\Delta R/R$ (C) is related to Si surface potential (V_{s}) [14] as expressed by Eq. 2. We modify and apply this model to evaluate the plasma-induced defect site per unit area (N_{dam}) as follows.

The basic concept is shown in Fig. 7. The effect of the charges trapped into the plasma-induced defect site (N_{dam}) is taken into consideration. By assuming that the trapped charge induces the change in V_{s} ($\Delta V_{\mathrm{s}} = V_{\mathrm{s}}^{\mathrm{ref}} - V_{\mathrm{s}}^{\mathrm{dam}}$) as seen in Fig. 7, N_{dam} can be calculated by solving Poisson equation.

$$N_{dam} = \frac{2\varepsilon_0 \varepsilon_{Si} \Delta V_s}{e d_{IL}} \tag{5}$$

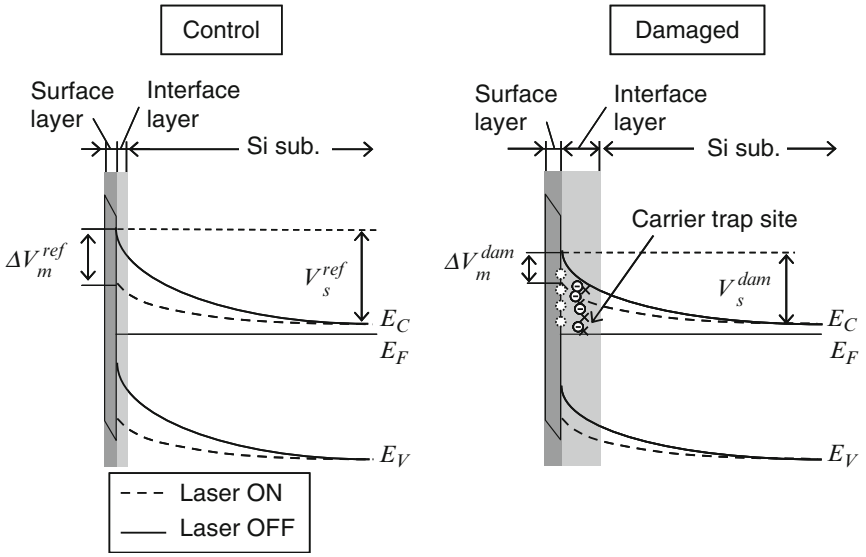


Fig. 7 Schematic energy diagram of the model for defect induced surface potential change during the photoreflectance measurement. Owing to charges trapping into defect sites near surface region, the surface potential decreases

where $\varepsilon_0$and ε_{Si} are the permittivity of vacuum and that of silicon, respectively. ΔV_s is the difference of Si surface potential between the control and the damaged samples. d_{SL} is the IL thickness of damaged sample.

Thus, N_{dam} is used as a measure of damage [9]. Details for this model are reported elsewhere [7].

2.4 Current–Voltage Measurement

Figure 8 shows a measurement system for electrical characteristics of sample surfaces. The damaged structures were analyzed by capacitance–voltage (C–V) measurement [16]. HP-4284A LCR meter connected to a mercury probe system was used to evaluate the capacitance of the surface and interfacial layer for the control and the plasma-exposed samples. The samples were assumed to be pseudo-MOS (Metal (Hg)/Oxide (damaged layer)/Semiconductor (Si substrate)) structures: The damaged layer was assigned to include SiO_2 layer, thus, we denote this structure as pseudo-MOS in this article. The frequency of modulation bias was 1 MHz. The amplitude was 100 mV. The modulation bias was superimposed on the DC voltage sweeping form 0 to 4.0 V by 0.1-V increment. The area of metal contact was 2.1×10^{-2} cm^2. All the measurements were performed at the room temperature.

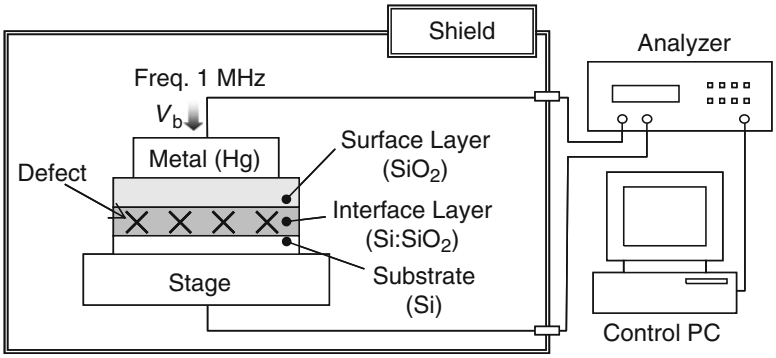


Fig. 8 Illustration of C–V measurement setup

3 Results and Discussion

3.1 SE and PR analyses

Table 1 summarizes the obtained results by optical analysis for various bias-frequency configurations. In Table 1, f_{SiO2} is the fraction of SiO_2 in the IL. Figure 9 shows the thickness in SL and IL (d_{SL} and d_{IL}) determined by SE for various bias configurations as a function of measured self bias voltage (V_{dc}). The V_{dc} corresponds to the mean impacting ion energy [6]. As described in Fig. 3, V_{dc} varies with the superposed bias configuration even if the total power is the same (100 and 200 W in total). The thickness in IL (d_{IL}) also varies with the bias configuration. As seen in Fig. 9, the IL thickness increases with V_{dc}, indicating more severe damage. The SL thickness increases with V_{dc} and tends to decrease in the higher V_{dc} region. This observed dependence may be attributed to the sputtering of the surface layer by Ar ion with higher energy [17]. The observed thickness range is found to be consistent with a simulation result as seen in Fig. 10. Figure 10 illustrates the depth profile of Ar ion permeating into Si substrate calculated by the simulation developed for an etch profile projection [18]. The simulation model includes the binary elastic collision between an Ar ion and a Si atom. Ions are found to penetrate silicon with the peak of the profile located from 2 to 5 nm from surface in the case of the energies ranging from 50 to 200 eV. By taking into account the thickness and composition of the surface layer listed in Table 1, the thickness range of IL obtained by SE analysis is consistent with the theoretical calculation in Fig. 10.

Figure 11 shows an example of PR spectra. By fitting the obtained spectra with Eq. 1, we can obtain E_g corresponding to the mechanical strain change developing in IL, as listed in Table 1, and amplitude C. One of the mechanisms for the observed E_g shift is considered to originate from the change in a mechanical strain developing in the surface region of Si substrate. On the basis of the hydrostat ic strain model, the change in the strain developing in interfacial layer can be estimated,

Table 1 Extracted parameters determined by SE and PR analyses

Process	Bias (W) (400k/13.56 M)	d_{SL} (nm)	d_{IL} (nm)/f_{SiO2}	E_g (eV)	ΔV_S (mV)	V_{dc} (V)
ICP-A	100/0	4.1	0.6/0.75	3.40	−1.8	−52
ICP-B	75/25	3.5	0.7/0.73	3.39	−11.0	−62
ICP-C	50/50	3.5	0.7/0.74	3.38	−11.7	−74
ICP-D	25/75	3.4	0.8/0.68	3.38	−16.6	−87
ICP-E	0/100	3.1	1.0/0.78	3.38	−30.6	−115
ICP-F	200/0	4.5	0.7/0.78	3.40	−16.5	−82
ICP-G	150/50	3.4	1.1/0.72	3.39	−33.3	−93
ICP-H	100/100	3.2	1.4/0.77	3.39	−80.1	−103
ICP-I	50/150	2.2	1.6/0.72	3.37	−114.0	−175
ICP-J	0/200	1.7	1.6/0.58	3.38	−18.5	−211
Control	–	0.9	0.2/0.59	3.40	–	–

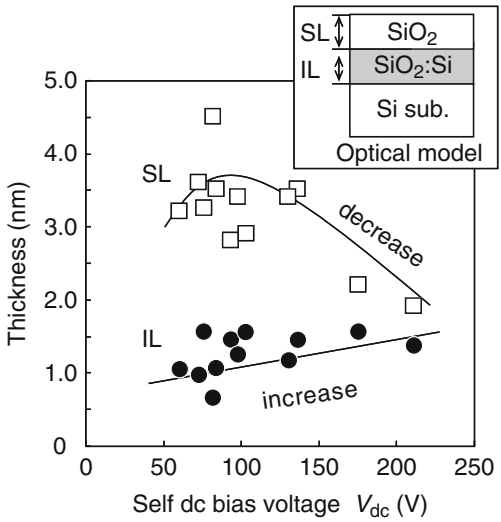


Fig. 9 Surface and interfacial layer thicknesses as a function of self bias voltage for various superposed bias-frequency configurations

as approximately 0.004 ($\Delta E_g = 0.02$ eV) between the control and damaged samples. Also by comparing the difference in C between the damaged and control samples in Eq. 2, the surface potential change (ΔV_s) by plasma-induced carrier trap site is calculated. From the obtained ΔV_s, N_{dam} is calculated.

Figure 12 shows that N_{dam} increases clearly in the range of $|V_{dc}|$ below ~150 V, while, in the range of $|V_{dc}|$ above ~50 V, N_{dam} tends to saturate for various superposed bias-frequency configurations. From Fig. 12, the areal defect site density of $\sim 10^{13}$ cm^{-2} are estimated in the exposed surface region (near the IL and Si substrate interface). Wada et al. also determined plasma-induced defect site density as $\sim 10^{18}$ cm^{-3} [19]. Both experimental observations discuss the

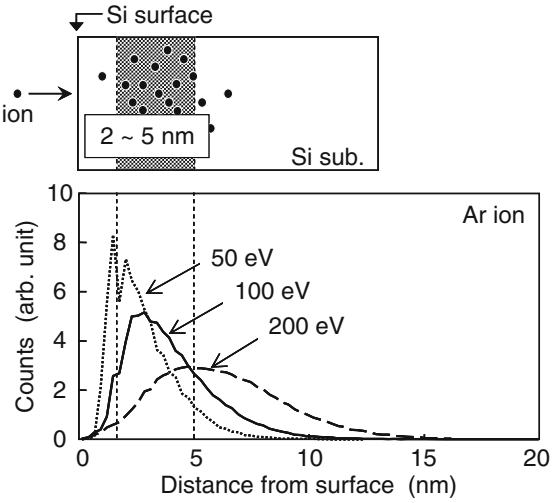


Fig. 10 Simulation results of the depth profiles of Ar ions impinging into plasma-exposed Si surface layer with the energies correspond to the present study

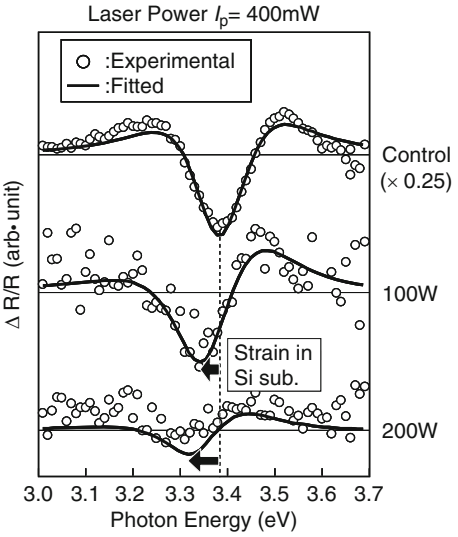


Fig. 11 Examples of PR spectra for the control and plasma exposed samples. Bias-frequency configuration is 400 kHz/13.56 MHz = 1/1

distribution of defect site located within 10 nm from the surface since the probe beam penetration depth is approximately 10 nm. Thus, by comparing the above values ($\sim 10^{13}$ cm^{-2} and $\sim 10^{18}$ cm^{-3}), it can be speculated that plasma-induced carrier trap sites are locally distributed near the IL and Si substrate interface. Further depth profile analyses based on this technique should be conducted in the future.

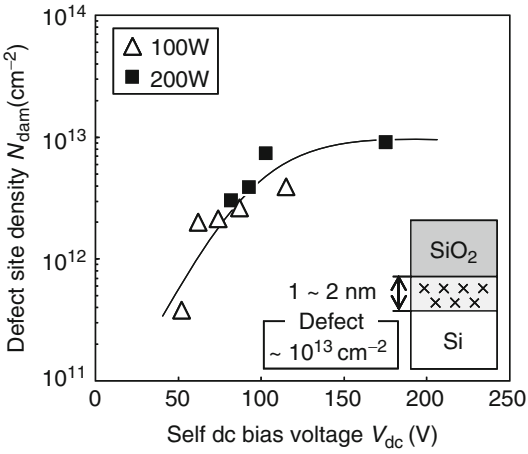


Fig. 12 Calculated defect site density (N_{dam}) as a function of measured V_{dc} for various superposed bias configurations

3.2 C–V Analysis

Figure 13 shows C–V characteristics for the control and plasma-exposed samples obtained by a mercury probe system. Distortion of C–V curve is clearly observed for the plasma-exposed samples. Two mechanisms are speculated for the distortion; (1) the shift of curve toward the positive bias implies the generation of negative charge trapping near the interfacial layer, (2) the decrease in capacitance in the region of positive bias voltage corresponds to the increase in surface layer thickness. This is consistent with the results (d_{SL}) by SE analysis.

Figures 14 show the bias voltage shift (ΔV_b) as a function of the decrease in maximum capacitance (ΔC_{max}) (left) and of V_{dc} (right), respectively. Since ΔV_s is in proportion to trapped charge density, ΔV_s is dependent on the plasma-induced defect density N_{dam}. On the other hand, since ΔC_{max} is in proportion to the inverse of surface layer thickness, thus ΔC_{max} depends on d_{SL}. Therefore, the left figure implies the relationship between N_{dam} and d_{SL}. The right figure shows that ΔV_b ($\sim N_{dam}$) depends on V_{dc} for various superposed configurations, which is consistent with the result in Fig. 12. Therefore C–V measurement can identify plasma-induced defect generation. The damage identified by C–V measurement is found to depend on the plasma parameter, V_{dc} for various superposed bias-frequency configurations.

It is concluded from PR and C–V measurement that an ion energy and/or IEDF are confirmed to be a predominant parameter for Si substrate damage (Si recess thickness) in a plasma reactor with superposed bias-frequency.

All of present findings are summarized in Fig. 15. From SE analysis, the thicknesses of SL and IL are determined as approximately 5 and 2 nm, respectively. The obtained thickness is consistent with widely observed Si recess thicknesses. From PR analysis, the carrier trap site density of the order of 10^{12} cm^{-2} is identified

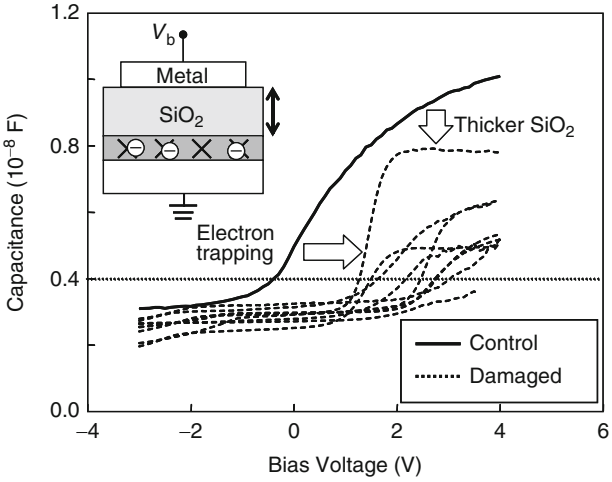


Fig. 13 Capacitance–voltage curves obtained by a mercury probe system

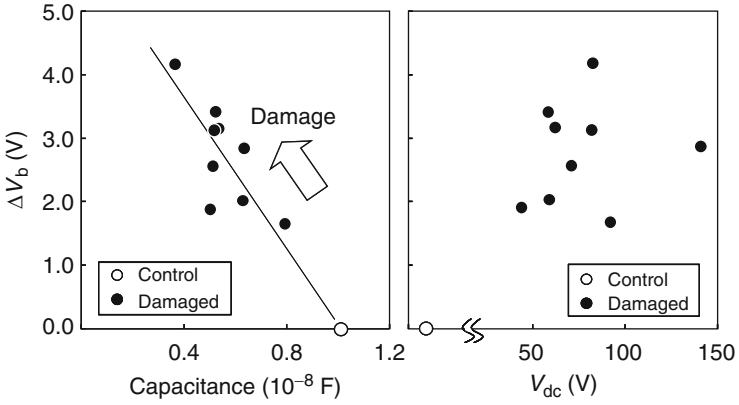


Fig. 14 Relationship between bias voltage shift at 4.0×10^{-9} F in Fig. 13 and maximum capacitance (*left*) and self bias voltage (V_{dc}) monitored by an oscilloscope in Fig. 2 (*right*)

in an IL. Moreover, it is confirmed from PR and C–V measurement that the damaged layer thickness (d_{SL} and d_{IL}) and the defect site density in IL (N_{dam}) depend on V_{dc}. The trap site plays a role as a tunneling site for leakage even after removal of surface layer in conventional manufacturing processes. Finally the present findings imply that the SL and IL structure, in particular, the carrier trap site involved should be evaluated from various aspects; for future plasma designs such as an optimization of superposed bias frequency configurations, and also for device designs by controlling Si recess by plasma damage.

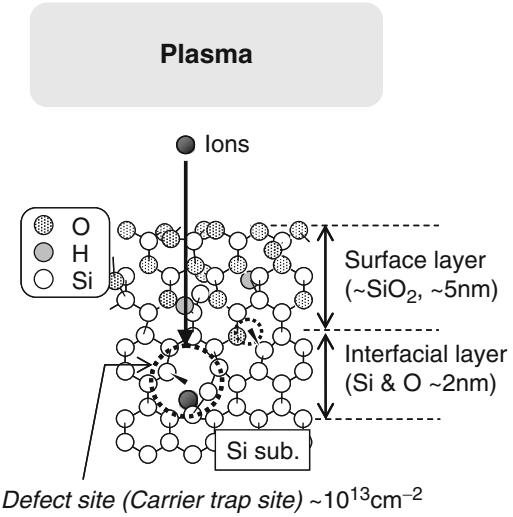


Fig. 15 Schematic illustration of plasma-exposed Si surface layers. These layers result in the increase in Si recess thickness and that in junction leakage due to the created carrier trap site

4 Conclusion

The plasma-induced defect site in Si layer was studied by the optical and electrical methods for various bias-frequency configurations. The growth of surface and interfacial layers (corresponding to the thickness increase in Si recess) was assigned by both SE and C–V measurement. The carrier trap site was identified by PR and C–V measurement. The above mentioned plasma damage is found to strongly depend on the superposed bias-frequency configuration. The present results provide a key guideline for future plasma and device designs.

Acknowledgement We thank Drs. M. Yoshimaru, M. Nakamura, H. Nakagawa, S. Noda and K. Ishikawa at Semiconductor Technology Academic Research Center (STARC) for their helpful discussion. This work was financially supported in part by STARC.

References

1. K. Eriguchi, K. Takahashi, K. Ono, Plasma-induced damage and its impacts on the reliability of advanced semiconductor devices, *Proceedings of the 6th International Conference Reactive Plasmas and 23rd Symposium Plasma Processing*, 2006, pp. 5–6
2. T. Ohchi et al., Reducing damage to Si substrate during gate etching processes, *Proceedings of the Symposium on Dry Process*, 2007, pp. 285–286
3. W.-H. Lee et al., High performance 65nmSOI Technology with enhanced transistor strain and advanced-low-K BEOL, *IEDM Technical Digest*, 2005, pp. 61–65
4. A. Steegen et al., 65nm CMOS technology for low power applications, *IEDM Technical Digest*, 2005, pp. 69–72

5. S.A. Vitale, B.A. Smith, Reduction of silicon recess caused by plasma oxidation during high-density plasma polysilicon gate etching. J. Vaccum Sci. Technol. **B21**, 2205–2211 (2003)
6. M.A. Lieberman, A.J. Lichtenberg, *Principles of Plasma Discharges and Materials Processing*, 2nd edn. (Wiley-Interscience, New York, 2005)
7. K. Eriguchi, K. Ono, Quantitative and comparative characterizations of plasma process-induced damage in advanced metal -oxide-semiconductor devices. J. Phys. D Appl. Phys. **41**, 024002 (2008)
8. A. Kojima et al., Dual frequency superimposed (DFS) rf capacitive coupled plasma etch process*, Proceedings of the Symposium on Dry Process*, 2003, pp. 13–17
9. Y. Nakakubo et al., Scaling of plasma-induced defect generation probability in Si: Effects of bias voltage at single- and superposed-frequencies, *Proceedings of the Symposium on Dry Process*, 2007, pp. 287–288
10. A. Matsuda, Y. Nakakubo, Y. Ueda, H. Ohta, K. Eriguchi, K. Ono, Significance of interface layer between surface layer and Si substrate in plasma-exposed structures and its impacts on plasma-induced damage analysis, *Ext. Abs. Solid State Dev. Mat.*, 2008, pp. 358–359
11. D.E. Aspnes, Optical properties of thin films. Thin Solid Films **89**, 249–262 (1982)
12. D.E. Aspnes, Third-derivative modulation spectroscopy with low-field electroreflectance. Surf. Sci. **37**, 418–442 (1973)
13. J.T. Fitch, C.H. Bjorkman, G. Lucovsky, F.H. Pollak, X. Yin, Intrinsic stress and stress gradients at the SiO_2/Si interface in structures prepared by thermal oxidation of Si and subjected to rapid thermal annealing. J. Vac. Sci. Technol. **A7**, 775–781 (1989)
14. H. Shen, M. Dutta, Franz–Keldysh oscillations in modulation spectroscopy. J. Appl. Phys. **78**, 2151–2176 (1995)
15. F.H. Pollak, H. Shen, Modulation spectroscopy of semiconductors: bulk/thin film, micro-structures, surfaces/interfaces and devices. Mater. Sci. Eng. R **10**, 275–374 (1993)
16. S.M. Sze, *Physics of Semiconductor Devices*, 2nd edn. (Wiley-Interscience, New York, 1981)
17. A. Kubota, D.J. Economou, A molecular dynamics simulation of ultrathin oxide films silicon: Growth by thermal O atoms and sputtering by 100 eV Ar+ ions. IEEE Trans. Plasma Sci. **27**, 1416–1425 (1999)
18. Y. Osano, M. Mori, N. Itabashi, K. Takahashi, K. Eriguchi, K. Ono, A model analysis of feature profile evolution and microscopic uniformity during polysilicon gate etching in Cl_2/O_2 plasmas. Jpn. J. Appl. Phys. **45**, 8157–8162 (2006)
19. H. Wada, M. Agata, K. Eriguchi, A. Fujimoto, T. Kanashima, M. Okuyama, Photoreflectance characterization of the plasma-induced damage in Si substrate. J. Appl. Phys. **88**, 2336–2337 (2000)

Part V
Power, Timing and Variability

CMOS SOI Technology for WPAN: Application to 60 GHZ LNA

A. Siligaris, C. Mounet, B. Reig, P. Vincent, and A. Michel

1 Introduction

This chapter discusses the design of a 60 GHz Low Noise Amplifier (LNA) using a standard low power SOI CMOS process from ST Microelectronics. First, we outline the technology as well as the mm-wave design challenges. Using recent work on Coplanar Waveguide (CPW) modeling, we describe how it's possible to use parametric, 3D electromagnetic simulation to complete or replace analytical models of on-chip passive devices. A short description of the transistor model is also provided. Finally, we discuss the details of the LNA design and show how the simulation results compare to the measurements.

Recently, much effort has been performed in order to develop millimeter wave circuits for 60 GHz wireless applications. SOI CMOS technology is a good candidate for such applications because it offers high quality passive elements and good substrate isolation due to high resistivity substrate [1]. Accurate modeling up to high frequency of passive as well as active devices of a technology is needed for robust circuit design. Moreover, models have to be available for both steady state (harmonic balance) and transient simulations. We first show the modeling flow for passive and active elements applied to a standard industrial SOI CMOS 65nm technology from STM. The generated models are used to design circuits for 60 GHz applications. Finally, we discuss the design of an LNA at 60 GHz and show how the simulation compares to the measurements.

A. Siligaris (✉), C. Mounet, B. Reig and P. Vincent,
CEA-LETI,
e-mail: Alexandre.Siligaris@cea.fr

A. Michel
ANSOFT, France

A. Amara et al. (eds.), *Emerging Technologies and Circuits*,
Lecture Notes in Electrical Engineering 66, DOI 10.1007/978-90-481-9379-0_9,

2 Modeling SOI CMOS 65 NM Technology

The LNA was implemented in low power SOI 65 nm SOI process. This technology features a back-end with six copper metal layers on which an aluminum cap-layer is added. One key parameter for reliable design at millimeter wave frequencies is accurate modeling of the transistors as well as the transmission lines (TL) and parasitic. The 65 nm SOI design kit provides a BSIM4 model for the standard CMOS design flow. Nevertheless, the model is not validated for high frequency applications. For that reason, we have used an empirical model with the benefit of fast extraction and accuracy which has been verified from DC to 110 GHz. The empirical transistor's model uses analytical equations for intrinsic I(V) and Q(V) description and an electrical small signal equivalent circuit for the extrinsic elements of the device. Details about this model are available in references [2] and [3]. Figure 1 shows a comparison of the measured and simulated S parameters of a 50 μm width (40 gate fingers) and 60 nm gate length transistor. The transistor is biased at moderate inversion (Vg = 0.8 V) and saturation (Vd = 1.2 V). A good matching is obtained from low frequency to 110 GHz. The model was implemented in Nexxim, state of the art circuit simulator (Ansoft), with care taken to make the model to work seamlessly and accurately in both time and frequency domain. Indeed, for many applications one needs to use the models in mixed mode. As an example, Fig. 2 shows the time domain output voltage of the transistor simulated with both a transient simulation and HB in steady state.

The passive elements that are used in the circuits (mainly CPW TL) are extracted through 3D full wave simulations with HFSS of Ansoft. The modeling is undertaken in two phases: first a 3D EM analysis is undertaken from which the CPW passive structures are characterized (S parameters) as a function of the frequency and the geometrical characteristics. The generated data helps to build accurate analytical models that are suited for electrical simulations and circuit design at

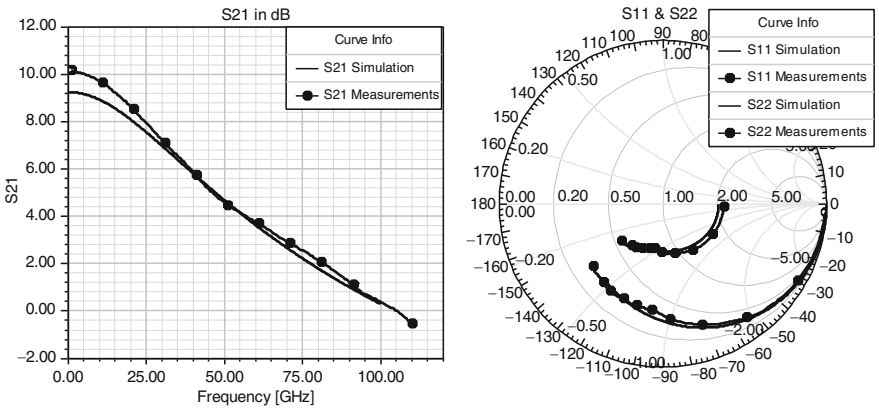


Fig. 1 Measured and simulated S parameters of a 50 μm SOI CMOS65 nm transistor up to 110 GHz

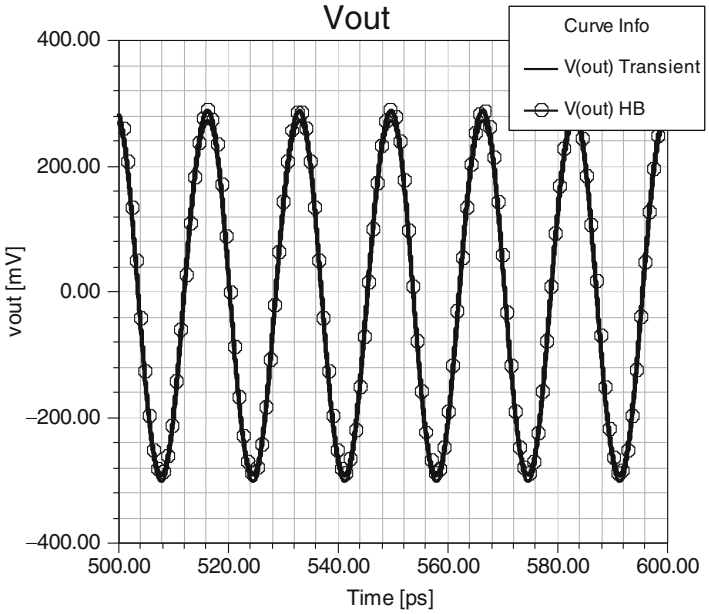


Fig. 2 Output voltage time domain voltage calculated with transient simulation and steady state HB

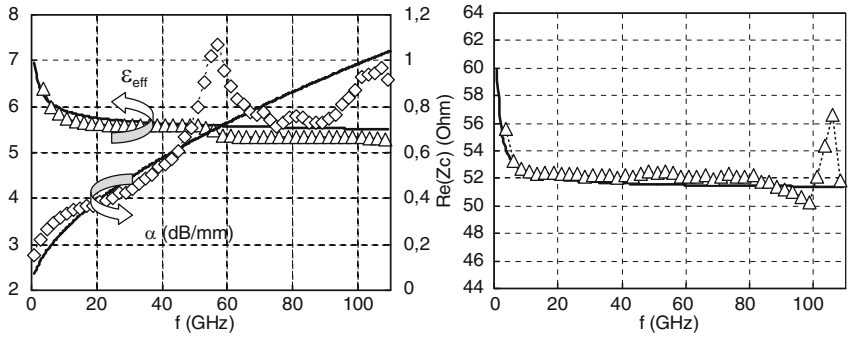


Fig. 3 Measured (*symbols*) and simulated (*solid lines*) insertion loss and effective permittivity. Measured (*symbols*) and simulated (*solid line*) real part of the characteristic impedance up to 110 GHz

high frequency. Analytical models for CPW TL on CMOS SOI technology can be found in reference [4]. Figure 3 a compares simulations and measurements of the attenuation, and the effective permittivity of a CPW line. The characteristic impedance is shown in Fig. 3b. Very good agreement is observed between measured and simulated data. Some of the blocks of the transceiver, like the VCO or the frequency Divider (frequency synthesis), will need to be simulated both in frequency domain

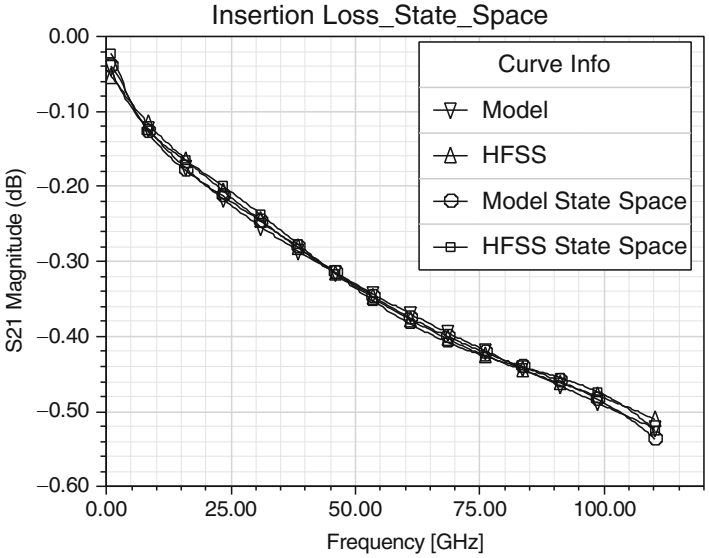


Fig. 4 State Space compared to originated S parameters

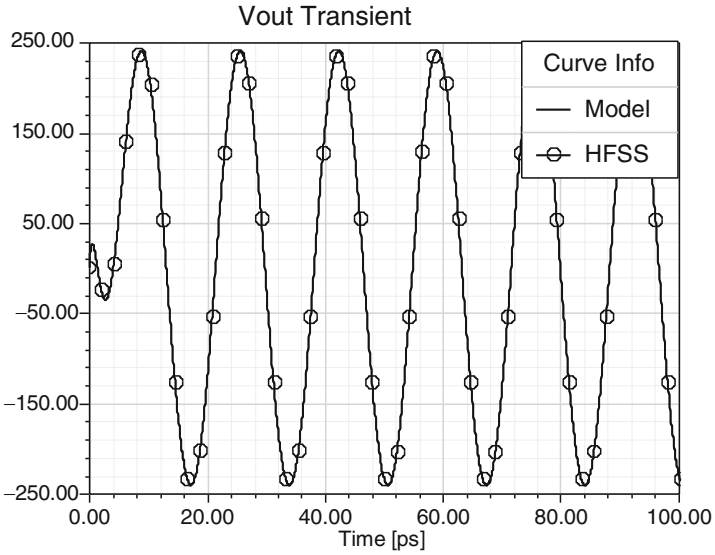


Fig. 5 Transient simulation results

(harmonic balance) and time domain (transient). So, it is very important at the modeling stage to validate the models in both frequency and time domain. The analytical and HFSS models are frequency dependent and are seen as an S parameter file by the simulators. Nexxim's standard transient solution for S-parameters uses

a state-space formulation to represent the model in the time domain, which is also guaranteed to be causal. This is highlighted in Fig. 4 which shows perfect match between S parameter and state space formulation of the insertion loss of the CPW transmission line. We also run a transient simulation with analytical model and HFSS model as shown in Fig. 5. We observe that transient simulation works with the transmission line even though the line model is defined in the frequency domain.

3 Low Noise Amplifiers in 65-NM SOI CMOS

A two-stage low noise amplifier was implemented in the CMOS65 nm SOI technology from STM. The amplifier uses a cascode topology for each stage (Fig. 6), because it offers higher isolation and gain than the common source (CS) topology. Moreover, the cascode topology ensures an unconditional stability for the amplifier in the 0–110 GHz frequency range. The input, the output, and the inter-stage matching networks are constructed from series CPW transmission lines and short ended stubs. The drain biasing of each stage is achieved through quarter-wave ($\lambda/4$) at 60 GHz short ended stubs. The RF shorts are constructed with the help of high density MOM capacitor. The circuit has been designed with the help of the transistors and TLs models that are described in the previous paragraph. Thanks to the geometry dependent models of the TLs, the layout design is easy and fast. The microphotograph of the chip is shown in Fig. 7. On wafer measurements of the S parameters were performed with an HP-8510XF VNA up to 110 GHz. Figure 8 shows the simulated and the measured S-parameters of the two-stage-65nm LNA. The gain of the amplifier is 12 dB at 64 GHz, while the input and output matching are better than -10 dB. From Fig. 8 we observe that very accurate simulation is obtained for the full chip, thanks to accurate active and passive models. The LNA consumption is 36 mW under 2.2 V voltage supply.

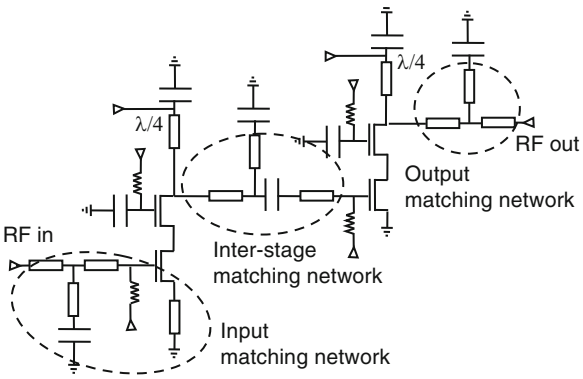


Fig. 6 Schematic of electrical circuit of a 60 GHz two-stage LNA implemented in SOI CMOS technology

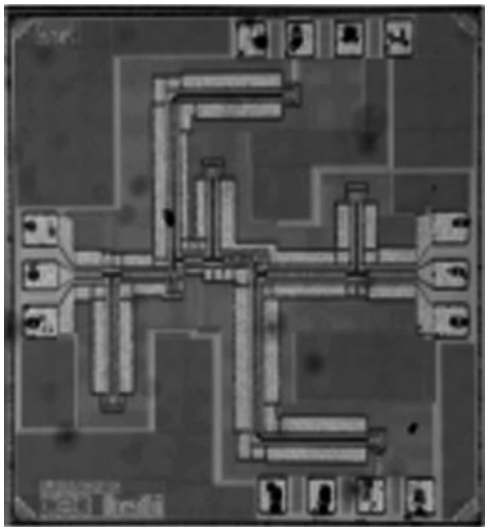


Fig. 7 Chip microphotograph of the two-stage-65nm 60 GHz LNA. Dimensions : $0.96 \times 1.05\ mm^2$

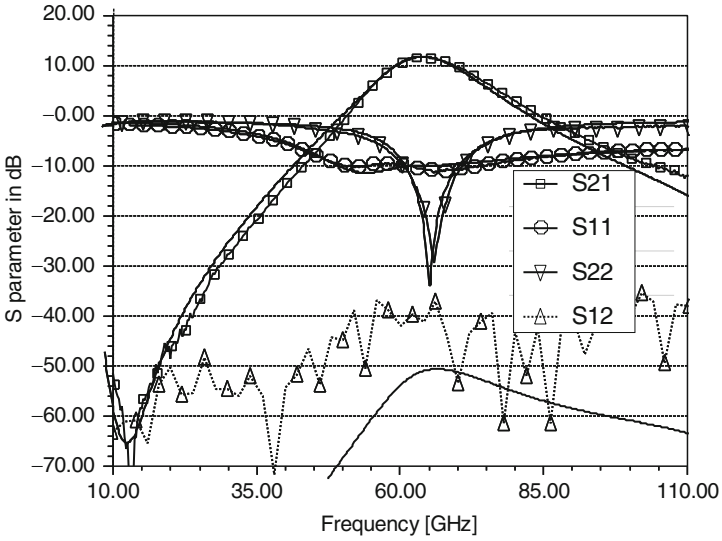


Fig. 8 Simulated and measured S parameters of the two-stage-65nm LNA. *Symbols*: measured. *Solid lines*: simulated

The noise figure of the LNA was measured at only one frequency point. Indeed, with the available noise measurements facilities it was only possible to measure at 60 GHz. The measured noise figure of the LNA is shown in Fig. 9 versus the current density of the first stage of the LNA. We observe that the noise figure is 8 dB and

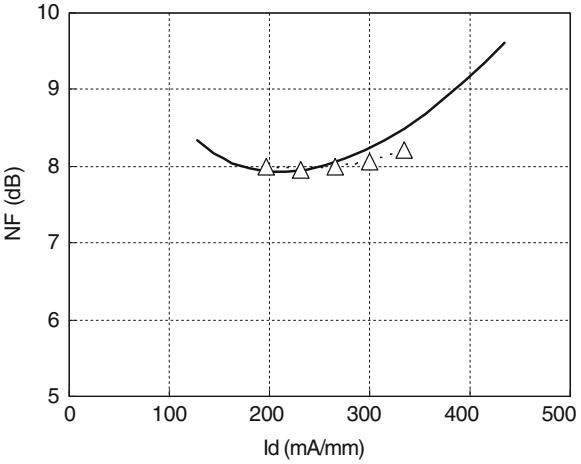


Fig. 9 Measured (*symbols*) and simulated (*solid line*) noise figure at 60 GHz versus the current density of the first cascode stage

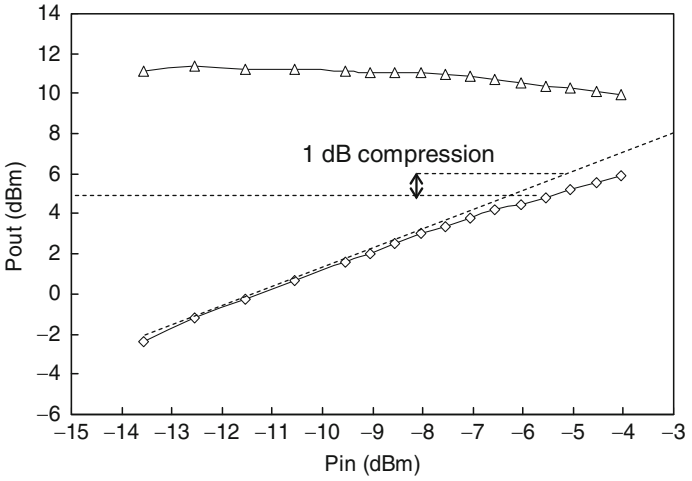


Fig. 10 Measured gain and output power at 64 GHz. The extracted output power at 1 dB compression point is 5 dBm

Table 1 General RF performances of the LNA

S21	S11	S22	NF	OCP_{1dB}	Pdc
12 dB (64 GHz)	<-10 dB (64 GHz)	<-20 dB (64 GHz)	8 dB (60 GHz)	5 dBm (64 GHz)	36 mW

shows a very small variation when the first stage transistor is biased at weak or moderate inversion. Power measurements were performed at 64 GHz in order to extract the compression point of the LNA. Figure 10 shows the measured gain and output power at 64 GHz. The extracted 1 dB compression point of the gain gives an output power of 5 dBm. The high linearity is obtained thanks to large transistors that constitute the LNA. Table 1 summarizes the general RF performances of the LNA.

4 Conclusion

In this chapter we have described the design flow of a 60 GHz LNA in a standard CMOS SOI technology. First we presented the modeling approach for transmission lines as well as for transistors. Passive elements are modeled with the help of HFSS 3D electromagnetic simulator, whilst the transistors use an empirical high frequency oriented CMOS model. Finally, we describe the design and the performances of a key element in high frequency transceivers, the LNA.

Acknowledgement The authors wish to acknowledge ST microelectronics for circuit fabrication and MC2 technologies for measurement support.

References

1. F. Gianesello et al., State of the art integrated millimeter wave passive components and circuits in advanced thin SOI CMOS technology on high resistivity substrate. *IEEE SOI Conference Proceedings* (Oct. 2005) pp 52–53
2. A. Siligaris, G. Dambrine, D. Schreurs, F. Danneville, A new empirical nonlinear model for sub-250 nm channel MOSFET. IEEE Microwave and Wireless Components Letters **13**(10), 449–451 (Oct. 2003)
3. A. Siligaris, G. Dambrine, D. Schreurs, F. Danneville, 130-nm Partially depleted SOI MOSFET nonlinear model including the kink effect for linearity properties investigation. IEEE Transaction on Electron Devices **52**(12), 2809–2812 (Dec. 2005)
4. Siligaris et al. CPW and discontinuities modeling for circuit design up to 110 GHz in SOI CMOS technology. *IEEE Radio Frequency Integrated Circuits (RFIC)Symposium Proceedings* (June 2007), pp 295–298

SRAM Memory Cell Leakage Reduction Design Techniques in 65 nm Low Power PD-SOI CMOS

Olivier Thomas, Marc Belleville, and Richard Ferrant

1 Introduction

As the technologies scale down into the nanometer range, the transistor leakage currents become a major concern. To overcome this problem, advanced control methods are mandatory, especially for circuits such as memories.

Different techniques have been proposed to reduce leakage in conventional 6T SRAM cell. Two main approaches exist: a process one, using dual VT [1, 2] and a design one, controlling the cell node voltages [3, 4]. A comparison of the most promising design techniques in sub 100 nm BULK technology has been presented in [5]. The aim of this chapter is to investigate these techniques in 65 nm PD-SOI technology, in order to determine the best one.

On the other hand, these leakage reduction techniques must not have a noticeable impact on SRAM cell performances. In BULK technology, the threshold voltage (VT) variation is only spatial. It depends on the process, voltage and temperature (PVT) variations. The floating body of PD-SOI technology adds temporal VT variations. As the body potential of the cell transistors depends on their node polarization, the leakage reduction techniques may induce a mismatch in the memory cell that temporary affects the performances of the memory cell. This mismatch can significantly degrade the electrical characteristics of the cell such as the static noise margin (SNM) considered as a key figure-of-merit. Hence, the temporal variability is a parameter that must be taken into account to evaluate adequately the best leakage reduction technique in PD-SOI. In addition to the evaluation of the saved leakage current, the impact of leakage reduction techniques on the SNM is

O. Thomas (✉) and M. Belleville,
CEA-LETI, Minatec 17 rue des Martyrs, 38054 Grenoble Cedex 9, France
e-mail: oliver.thomas@cea.fr

R. Ferrant
STMicroelectronics, 850rue Jean-Monet, 38920 Crolles, France

A. Amara et al. (eds.), *Emerging Technologies and Circuits*,
Lecture Notes in Electrical Engineering 66, DOI 10.1007/978-90-481-9379-0_10,

also investigated. To do so an SOI specific methodology has been developed for accurate SNM evaluation.

The chapter is organized as follows. Section 2 introduces analyzes and compares the design leakage reduction techniques considered in this work. Section 3 discusses the impact of leakage reduction techniques on cell stability. Section 4 comments the results. Finally, concluding remarks are drawn in the last section.

2 Design Leakage Reduction Techniques

For sub 100 nm technologies, gate (I_G), sub-threshold (I_{STH}) and GIDL (I_{GIDL}) currents are the dominant leakage mechanisms. Figure 1 illustrates their distribution in a BULK 6T SRAM cell. It results in three leaking paths: a first one from VDD to ground across the cell inverters, a second one from the bit-lines ($BL_{T/F}$) to ground and a third one from $BL_{T/F}$ to the word line (WL).

The leakage currents are directly related to the electric fields inside the device. Reducing the node voltages drastically decreases the leakage current. In [5], an analysis has been done on leakage current saving by controlling the different 6T SRAM cell node voltages (WL, BL, VDD, VSS, V_{NWELL}, V_{PWELL}). For a low standby power BULK technology, it has been shown that reducing the supply voltage of the cell (LVDD) and setting the bit-lines floating and equalized (FBL) leads to the best leakage reduction technique. LVDD reduces the leakage components flowing through the cell inverters, while FBL diminishes the leakage current of the access transistors. Raising the ground cell voltage (UVSS) is also efficient in the reduction of the leakage currents in the cell inverters but adds extra GIDL and gate currents flowing through access transistors. On the other hand, excessive GIDL and gate currents limit the interest of Reverse Body Biasing (RBB) and negative word-line (NWL) approaches, reducing only I_{STH}.

In SOI technology, because of the floating body isolation, all the BULK leakage reduction techniques are applicable, with the exception of the RBB technique. In

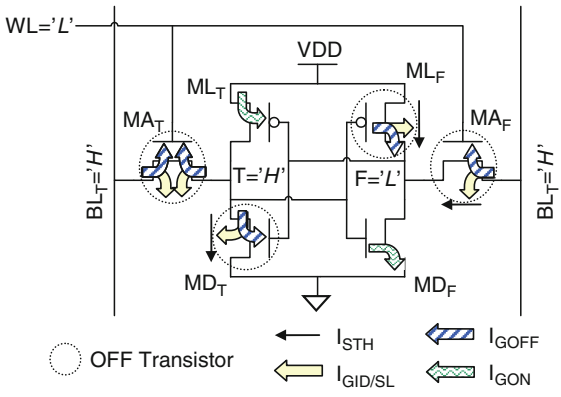


Fig. 1 Leakage current contribution in a BULK 6T SRAM cell

fact, the body nodes cannot be connected without dramatically increasing the size of the cell. It results that the cell leakage is affected by the polarities of the floating bodies of the cell transistors [7, 8], which are related to the voltage on the drain, the source and the gate of each transistor according to the memorized logical level

The DC body voltage (V_B) value of a PD-SOI transistor depends on the body leakage mechanisms such as the impact ionization, GIDL, gate tunneling and generation and recombination effects, as illustrated in Fig. 2 [9]. As example, for an NMOS transistor, the increase of V_B increases I_{STH} while reducing I_{GIDL}.

Figure 3 shows the leakage distribution in a PD-SOI 6T-SRAM cell and depicts the V_B amplitude of the transistors in OFF state ('H' = 1.2 V, 'L' = 0 V). Compared with BULK technology (cf. Fig. 1), the body voltage weakens I_{GIDL} and reinforces I_{STH}. It can also be noticed that I_{GIDL} is eliminated in the access transistor MA_T connected to the high logic level storage node T ($V_{BD/S} \sim 0$ V). It results that I_{STH} is the main leakage component in PD-SOI SRAM cell.

Comparing the leakage reduction techniques, in PD-SOI technology, UVSS and LVDD leakage reduction techniques have the same efficiency regarding the reduction of the cell inverters' leakage current. This comes from the source

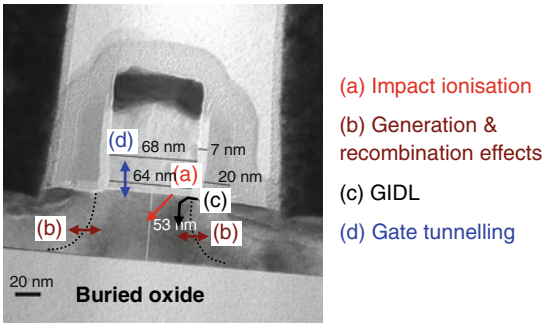


Fig. 2 Cross-sectional TEM image of a 65 nm NMOS transistor, showing the body leakage mechanisms

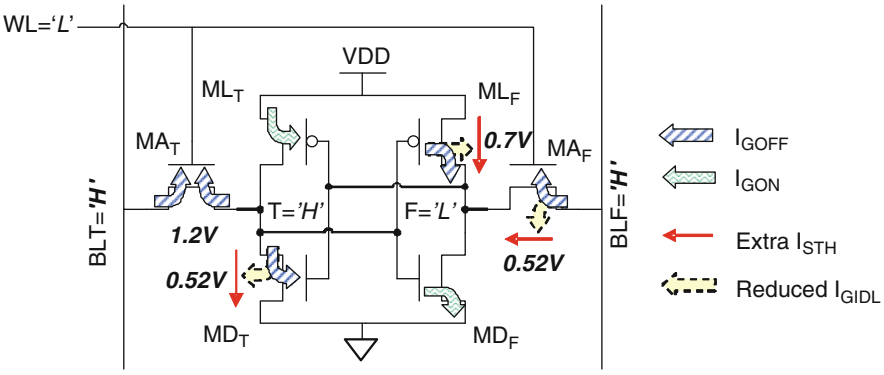


Fig. 3 Leakage current distribution in a PDSOI 6T SRAM cell. V_B of the transistor in OFF state are depicted (VDD = 1.2 V)

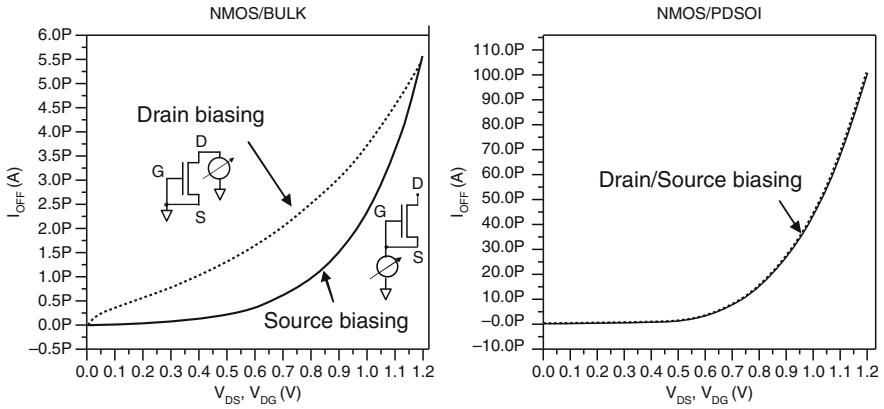


Fig. 4 Drain and source biasing effect on leakage in BULK and PDSOI technologies

follower effect in PD-SOI technology which removes the reverse source biasing effect on the driver and load transistors in off state, presents in BULK technology (see Fig. 4). Regarding the access transistors, in BULK technology, the drawback of UVSS versus LVDD is the extra I_G and I_{GIDL} flowing through the access transistor (MA_F) connected to the low logic level storage node F. On the other hand, UVSS reduces I_{STH} of MA_F ($V_{GS} < 0$) drastically. In PD-SOI technology, UVSS do not add extra I_{GIDL} flowing through MA_F, thanks to the increase of its body potential with F node voltage ($V_{FB} < 0$). Therefore, UVSS approach appears to be the most efficient leakage reduction technique.

Now, to fully successful in reducing leakage current of access transistors, the bit-line pair can be left floating and just equalized between accesses. Compared with BULK technology, the FBL technique is much more efficient in PD-SOI technology thanks to the body voltage drop of the access transistors with the bit line voltage drop, reducing the I_{STH} current drastically.

Thus, mixing FBL with UVSS allows reducing all the cell leakage current and leads to the best leakage reduction technique.

3 Cell Stability Versus Leakage Strategy

The fundamental cell functionality considerations must be addressed in conjunction with any leakage reduction strategies. In particular, the static noise margin (SNM) in read mode considered as a key figures of-merit. The leakage reduction techniques should not have any noticeable impact on it.

With technology scaling, the fluctuations of the process parameters (V_T, L, W, TOX ...) introduce an unbalance in the inverter behaviours between each side of the cell decreasing its stability. The primary parameter that causes variations is the threshold voltage, because it simultaneously depends on the fluctuations of the

transistor dimensions and on the doping concentration in the channel [6]. In SOI technologies, the floating body isolation introduces an additional source of fluctuation that further worsens the V_T variation of the transistors. This effect depends on the process tolerances and on the operating conditions of the cell.

Two mechanisms create the body voltage variations: A fast one (∼some ps) that is characterized by the body charge sharing during transitions and a slow one (∼100 µs to ∼1 s) linked to the body charge variation on steady state conditions. Figure 5 depicts these two effects by showing the body voltage (V_B) and charge (Q_B) variations of the NMOS transistor of an inverter.

During the switching period (1 to 2) there is a fast evolution of V_B and Q_B remains unchanged. Then (2 to 3), there is a slow variation of body voltage V_B driven by the Q_B evolution (ΔQ_B). The same simulation has been done for three different drain junction capacitances. It illustrates that the charge sharing mechanism depends on the drain, source and gate coupling capacitance network, while the charge evolution comes from the body leakage mechanisms (as described in Section 2). For each polarization of the transistor (logic state), there is a unique DC body voltage V_B that result from the equilibrium between the coupling capacitance network and Q_B. V_B behavior can be described by the following Eq. 1:

$$V_B = \frac{Q_B + C_D V_D + C_S V_C + C_G V_G + C_E V_E}{C_D + C_S + C_G + C_E} \tag{1}$$

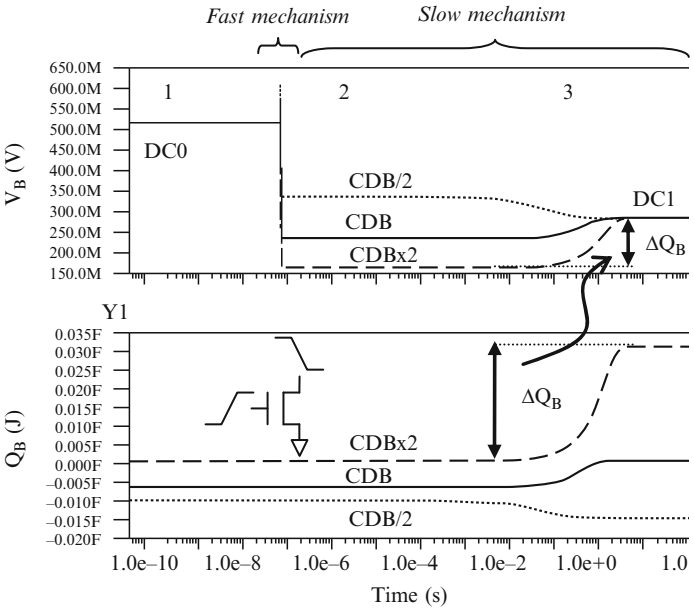


Fig. 5 Body potential fluctuation

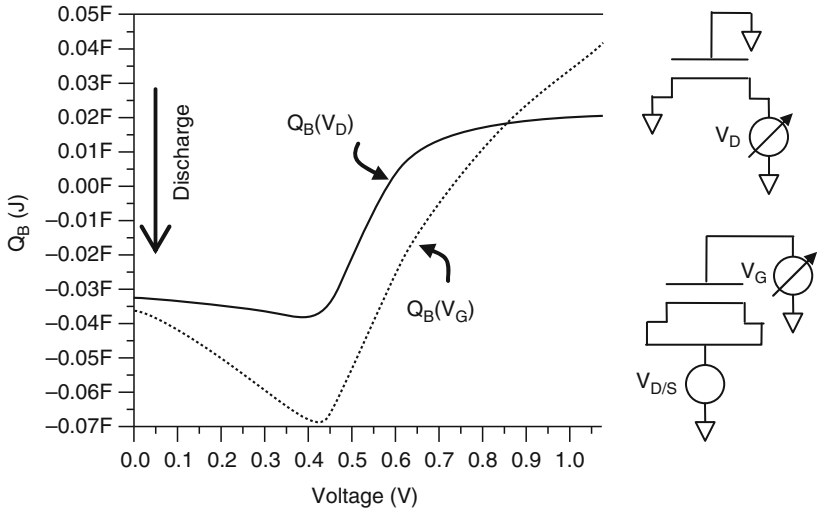


Fig. 6 Body charge variation of a NMOS transistor versus V_D and V_G

In read mode the body voltage mismatch will unbalance the cell, reducing the SNM. Indeed, the SNM depends simultaneously on the ability to maintain both the low (F) and high (T) internal nodes voltage. The stability of F is function of the r gain ratio (around 2) between the driver (MD_F) and access (MA_F) transistors. While H stability is guaranteed by the load transistor (ML_T) drive [10].

Body voltage mismatch resulting from the Q_B DC equilibrium in standby or idle mode is different according to the leakage reduction techniques applied. UVSS and LVDD affect the body charge of the cell transistors by reducing the gate-source/drain voltages ($V_{G\text{-}D/S}$), while FBL only affects the body charge of access transistors by reducing the drain-source voltage (V_{DS}). As shown in Fig. 6, the reduction of V_{DS} and $V_{G\text{-}D/S}$ leads to a discharge of the floating body, resulting in a reduction of the body voltage and thereby an increase of the threshold voltage.

It results in UVSS and LVDD degrading the SNM by weakening the driver transistors versus the access transistors leading to a reduction of the r ratio. On the other hand, the bit-lines voltage reduction when using FBL weakens the access transistors which this time increases the effective r ratio and improves the SNM.

4 Results

Figure 7 presents, for a 6T-SRAM cell, in 65 nm PD-SOI technology, the normalized leakage current obtained for LVDD, UVSS, LVDD+FBL and UVSS+FBL leakage reduction techniques versus the delta voltage applied to supply or ground of the cell (VDD-delta, VSS+delta). The simulations have been done in worst case (WC) i.e. a fast process (FF), an increase of VDD by 10% (1.32 V) and a temperature of 125°C.

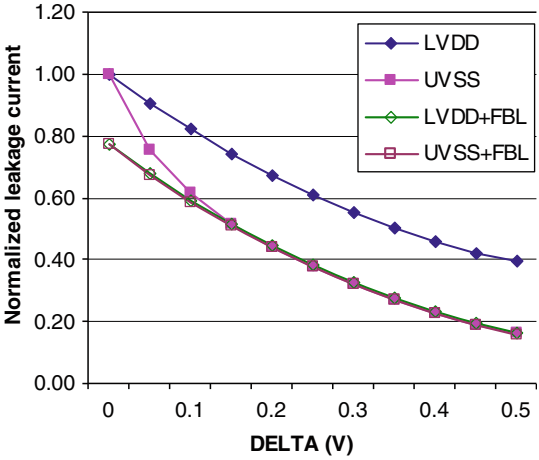


Fig. 7 Normalized 65 nm PD-SOI 6T SRAM cell leakage versus leakage reduction technique at worst case (FF, 1.32 V, 125°C)

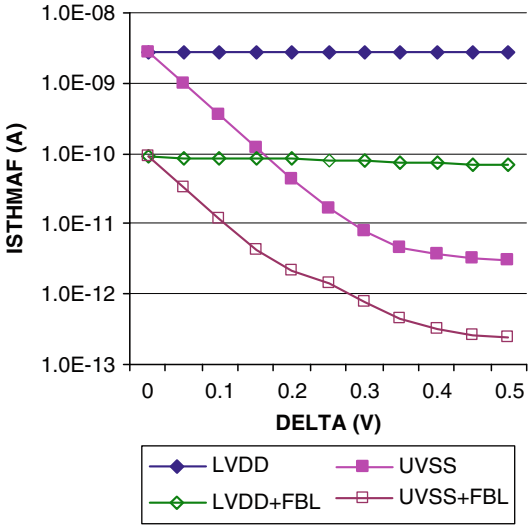


Fig. 8 ISTH of MAF versus leakage reduction techniques at worst case (FF, 1.32 V, 125°C)

Compared with LVDD, UVSS is much more efficient thanks to the reduction of I_{STH} of MA_F transistor, as depicted in Fig. 8. Though, if the bit lines are left floating (FBL) the access transistor leakage are reduced by more than 1 decade becoming negligible and resulting in LVDD+FBL being as much efficient as UVSS+FBL. Those techniques give a total leakage current reduction around 80% for a delta of 0.4 V.

Figure 9 shows the SNM variations obtained for LVDD, UVSS, LVDD+FBL and UVSS+FBL leakage reduction techniques versus the delta voltage. The SNM measurements correspond to a read cycle following a long period of time in idle mode. The results have been normalized to the default SNM measured without leakage reduction techniques. Both UVSS and LVDD approaches degrade the SNM

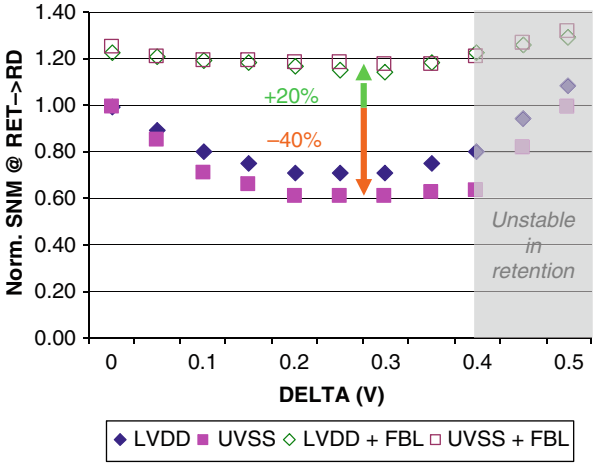


Fig. 9 Normalized SNM (FS, 1.08 V, −40°C) versus the delta voltage applied to the ground or the supply voltage of the cell

by up to 40%, UVSS representing the worst case. It originates from a lower r ratio. In order to have a SNM equal to or greater than the default one, FBL technique has to be added. The results show an improvement of 20%.

In retention mode, the cell stability (RNM) depends on the cell inverters, only. Due to the source follower effect in the cell inverters transistors, the RNM is the same for each leakage reduction technique. The results obtained by Monte-Carlo analysis in worst case (FS, 1.08 V, −40°C), including mismatch, give a minimum cell voltage of 0.7 V. This limit has been defined following the ratio between the RNM average value and the standard deviation, for an acceptance criteria higher than 7 (100 ppm fails for 10 Mbit/chips).

5 Conclusion

Whereas below 65 nm, I_{GIDL} and I_G become as important as I_{STH} in BULK technology, in PD-SOI technology I_{STH} is still the dominant current due to the floating body. Thereby, to overcome the leakage problem in PD-SOI, the main parameters to address are the body voltages of the cell transistors. It has been shown that the best results are obtained when letting the bit lines floating (FBL) and increasing the cell ground voltage (UVSS) or reducing the cell supply voltage (LVDD). Leakage reduction up to 80% is achieved.

Regarding the impact of leakage reduction techniques on the cell stability during the first read cycle, the main parameters to address are the body charge of the access transistors. Letting the bit lines float during the idle mode and prior to a read access, improves the SNM of the cell thanks to the V_T increase of the access transistors and thus protects the cell content against the bit-line aggressions.

References

1. F. Hamzaoglu et al., Analysis of dual-Vt SRAM cells with full-swing single-ended bit line sensing for on-chip cache, *IEEE TVLSI Systems* **10**, Apr 2002
2. N. Azizi et al., Low-leakage asymmetric-cell SRAM, *IEEE VSLI Systems* **11**(4), Aug 2003
3. A. Agarwal et al., DRG-cache: A data retention gated-ground cache for low power, *DAC*, June 2002, pp. 473–478
4. K. Nii et al., A 90-nm low-power 32-kB embedded SRAM with gate leakage suppression circuit for mobile applications, *IEEE JSSC* **39**(4), Apr 2004
5. O. Thomas et al., Impact of CMOS technology scaling on SRAM standby leakage reduction techniques, *ICICDT*, May 2006
6. S. Mukhopadhyay et al., Modeling and estimation of total leakage current in nano-scaled -CMOS devices considering the effect of parameter variation, *ISPLED*, Aug 2003
7. T.Y. Chan et al., The impact of gate-induced drain leakage current on MOSFET scaling, *IEDM*, Dec 1987
8. BSIM 4 user manual
9. K. Bernstein, N.J. Rohrer, *SOI Circuit Design Concepts*. Kluwer 2000
10. O. Thomas et al. Sub-1V, robust and compact 6T SRAM cell in double gate MOS technology, *ISCASS*, May 2007

Resilient Circuits for Dynamic Variation Tolerance

Keith A. Bowman and James W. Tschanz

1 Introduction

Integrated circuits are susceptible to dynamic variations in supply voltage (V_{CC}) and temperature. Abrupt changes in die-level switching activity induce large current transients in the power delivery system, resulting in V_{CC} droop and overshoot fluctuations. The magnitude and duration of V_{CC} droops and overshoots depend on the interaction of capacitive and inductive parasitics at the board, package, and die levels with changes in current demand [1]. Temperature variations depend on workload, environmental conditions, and the heat-removal capability of the package. These dynamic variations in V_{CC} and temperature degrade the clock frequency (F_{CLK}) of microprocessors. Conventional designs build a guardband into the operating F_{CLK} to ensure correct functionality within the presence of worst-case dynamic variations. Consequently, these inflexible designs cannot exploit opportunities for higher performance by increasing F_{CLK} or lower power by lowering V_{CC} during favorable operating conditions. Since most systems usually operate at nominal conditions where worst-case scenarios rarely occur, these infrequent dynamic variations severely limit the performance and energy efficiency of conventional microprocessor designs.

On-die V_{CC} and temperature variation sensors coupled with adaptive circuit techniques have been demonstrated to adjust F_{CLK}, V_{CC}, or body bias in response to changes in V_{CC} and temperature [2–4]. These schemes reduce the F_{CLK} guardband from slow-changing V_{CC} and temperature variations, resulting in higher average F_{CLK}. Alternatively, the average F_{CLK} benefits may be converted to lower average power by decreasing V_{CC}. The disadvantages of on-die sensors and adaptive approaches include the inability to respond to fast-changing variations such as high-frequency V_{CC} droops [1]. Furthermore, sensors and adaptive circuits

K.A. Bowman (✉) and J.W. Tschanz
Intel Corporation, Hillsboro, OR, USA

A. Amara et al. (eds.), *Emerging Technologies and Circuits*,
Lecture Notes in Electrical Engineering 66, DOI 10.1007/978-90-481-9379-0_11,

require substantial calibration time per die, leading to increased testing costs. Although sensors may be tuned during test to reduce the delay mismatch between sensors and critical paths from within-die (WID) process variations, an F_{CLK} guardband is still necessary to ensure coverage across a wide range of V_{CC} and temperature conditions as well as for transistor aging.

In digital logic design, error-detection sequential (EDS) circuits have been proposed to monitor timing faults for on-line testing of digital circuits within the presence of environmental influences and reliability concerns [5–7]. The combination of timing-error detection and error-recovery [7–12] enables the microprocessor to operate at an F_{CLK} determined by nominal V_{CC} and temperature. When dynamic variations induce a timing error, the error is detected and corrected to maintain proper logic functionality. Since additional clock cycles are required for error recovery, instructions per cycle (IPC) reduce as errors occur. Assuming infrequent timing errors, the IPC reduction is relatively small as compared to the F_{CLK} gains from removing the guardband for dynamic V_{CC} and temperature variations, resulting in higher overall system throughput. In addition, further F_{CLK} benefits are possible by exploiting path-activation probabilities [7–12]. If the slowest paths on the die are infrequently activated, the F_{CLK} may increase higher than the critical path operating frequency. When these critical paths are activated, the timing error is detected and corrected. As an alternative to performance benefits, the F_{CLK} gains may be traded-off for lower power by reducing V_{CC}.

In Section 2, the application of EDS circuits for dynamic variation tolerance is reviewed and five separate EDS circuits are described. Next in Section 3, two error-recovery designs with different trade-offs in recovery cycles and design overhead are presented. In Section 4, test-chip measurements from a resilient circuit prototype [10, 11] are presented to demonstrate the effectiveness of resilient circuits in eliminating the F_{CLK} guardband from dynamic V_{CC} and temperature variations as well as exploiting path-activation probabilities to maximize performance efficiency. Section 5 concludes by summarizing the chapter and providing recommendations to further enhance the performance and energy efficiency of resilient circuits.

2 Error-Detection Sequential Circuits

2.1 Overview

The basic concept of timing-error detection circuits for dynamic variation tolerance is described in Fig. 1. A conventional path with master-slave flip-flops (MSFF) is provided in Fig. 1a along with conceptual timing diagrams in Fig. 1b, illustrating the arrival times of the input data (D) to the receiving flip-flop during worst-case dynamic variations and nominal conditions. Within the presence of worst-case dynamic variations, the input data to the receiving flip-flop must arrive

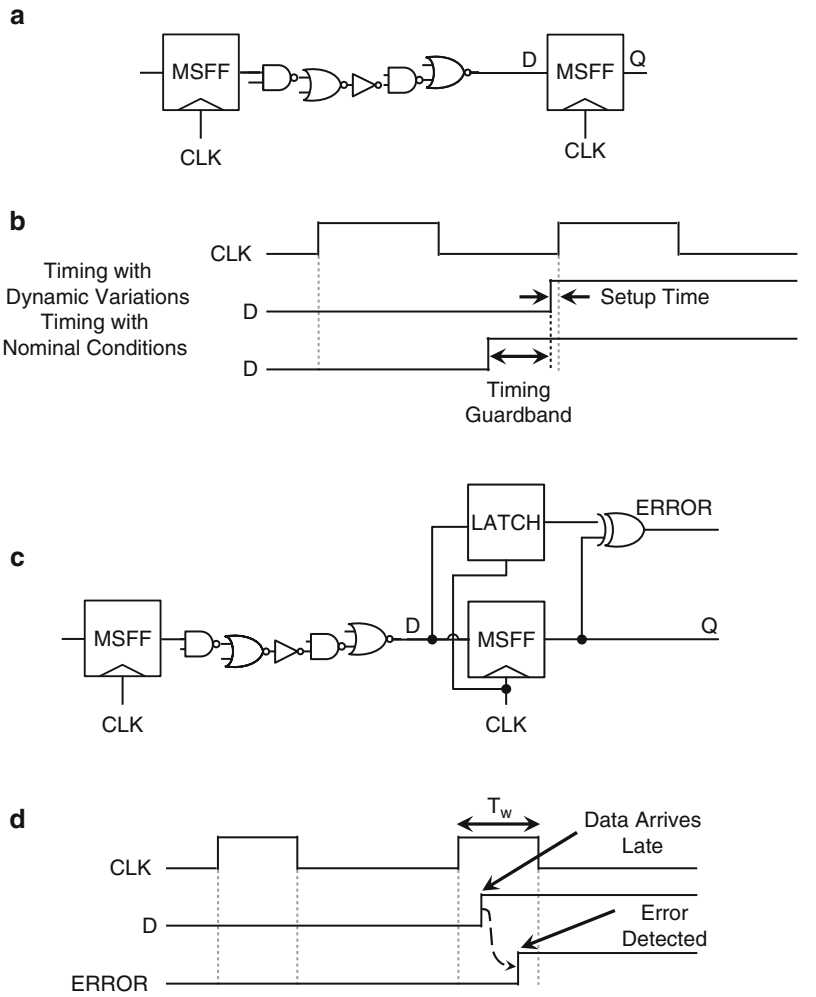


Fig. 1 (**a**) Conventional path design and (**b**) conceptual timing diagrams for worst-case dynamic variations and nominal conditions. (**c**) Resilient path design with a double-sampling (DS) error-detection sequential (EDS) circuit [5–7] and (**d**) conceptual timing diagram for late arriving input data [11] (© 2009 IEEE)

a setup time prior to the rising clock edge to ensure correct functionality. In comparison, the input data for the same path arrives much earlier during nominal conditions. The difference between the input data arrival times for these two cases represents the effective timing guardband required for dynamic variations. A resilient path is created by replacing the receiving MSFF of the conventional path with an EDS circuit as described in Fig. 1c. The conceptual timing diagram in Fig. 1d illustrates late arriving input data. The EDS circuit in Fig.1c is a double-sampling

(DS) error-detection design, consisting of a datapath MSFF, a shadow latch, and an XOR logic gate [5–7]. This EDS circuit double samples input data with the datapath MSFF on the rising clock edge and the shadow latch on the falling clock edge. The MSFF and latch outputs are compared with the XOR gate to produce an error signal (ERROR). If input data transitions late as described in Fig. 1d, MSFF and latch outputs differ, resulting in a logic-high error signal. The error signal is handled at the microarchitecture level to enable error recovery. Since the resilient circuit can detect and correct late arriving data, the timing guardband for dynamic variations in the conventional design can be removed, allowing the resilient circuit to operate at a higher F_{CLK}.

In comparison to a conventional MSFF, the error-detection capability of the DS EDS circuit in Fig.1c is attained at a cost in clock energy and area since an additional latch is needed. Although the increase in area at the flip-flop level appears large, the overall area penalty at the microprocessor level is expected to be relatively small (<1%). For scan flip-flops, the additional latch can be shared with scan circuitry to further reduce the area overhead. In contrast, the clock energy overhead at the flip-flop level significantly impacts the total dynamic energy of a microprocessor since sequential clock energy represents a large portion of the overall total dynamic energy. Moreover, a well-tuned microprocessor has a substantial number of critical paths, requiring a large fraction of the total sequentials to be protected with the timing-error detection capability.

A critical issue for the DS EDS circuit in Fig. 1c is the susceptibility to datapath metastability. If the input data to the datapath MSFF arrives slightly after the setup time prior to a rising clock edge, the output of the MSFF can become metastable. In this scenario, the MSFF requires an indefinite amount of time to resolve the output to a valid logic value, corresponding to a clock-to-output (CLK-to-Q) delay push-out. During metastability, the CLK-to-Q delay push-out exponentially depends on the relationship between the setup time and the input data arrival time [13–15]. Since the MSFF output feeds an error path, which is described further in Section 3, and multiple fan-out datapaths, the CLK-to-Q delay push-out from a metastable output can affect the error path differently from one of the fan-out datapaths such that an undetected error occurs. Since the mean time between failures (MTBF) must satisfy aggressive microprocessor targets, datapath metastability is a potential show-stopper for implementing the DS EDS circuit into a microprocessor product.

For each EDS circuit described in this chapter, the error-detection window (T_w) is based on the high clock phase delay as illustrated in Fig. 1d. The maximum path delay (max-delay) constraint within the presence of worst-case dynamic conditions for max-delay is defined as:

$$T_{max} \leq T_{cycle} + T_w - T_{setup,clk-f} \tag{1}$$

T_{max} is the maximum path delay, including clock skew and jitter delay, T_{cycle} is the cycle time ($=1/F_{CLK}$), and $T_{setup,clk\text{-}f}$ is the setup time based on the falling clock

edge. The minimum path delay (min-delay) constraint during worst-case dynamic conditions for min-delay is calculated as:

$$T_{min} \geq T_w + T_{hold,clk-f} \tag{2}$$

T_{min} is the minimum path delay, accounting for clock skew and jitter delay, and $T_{hold,clk\text{-}f}$ is the hold time based on the falling clock edge. The max-delay and min-delay constraints in (1)–(2) only apply to paths with an EDS circuit as the receiving sequential. For a target T_w, min-delay requirements are satisfied in pre-silicon design by buffer insertion. As T_w increases, the number of buffers increases, leading to larger power. From (1) and (2), the fundamental trade-off in timing-error detection circuits is max-delay versus min-delay. As T_w increases, T_{cycle} may decrease to enable a higher F_{CLK} while satisfying the max-delay constraint in (1) at a cost of increased min-delay penalty in (2). For microprocessors with deep pipelines (i.e., small number of logic stages between sequentials), this trade-off may not be advantageous due to the stringent min-delay requirements. In recent technology generations, however, the microarchitecture for microprocessors has moved towards shallow pipelines (i.e., large number of logic stages between sequentials) to improve energy efficiency [16, 17]. Microprocessors with shallow pipelines greatly relax the min-delay requirements as compared to a deep pipeline design, enabling a more effective trade-off of max-delay improvement for min-delay penalty. As additional protection from min-delay violations, the high clock phase, and corresponding T_w, may be tuned with a duty-cycle control circuit. In Section 3.3, a scan-tunable duty-cycle control circuit is described to adjust T_w.

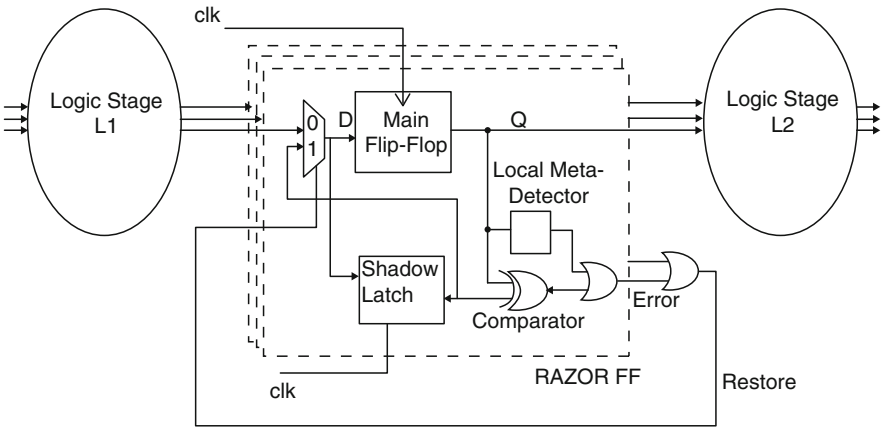


Fig. 2 Razor I EDS circuit [8, 9]. CLK is duty-cycle controlled to satisfy min-delay requirements (© 2006 IEEE)

2.2 Razor I

As described in Fig. 2, the Razor I EDS circuit [8, 9] contains a datapath MSFF and a shadow latch, which is similar to the DS EDS circuit in Fig. 1c. Razor I double samples input data and compares datapath MSFF and shadow latch outputs with an XOR logic gate. A local metastability detector is connected to the MSFF output. The XOR and metastability detector outputs feed an OR logic gate to generate an error signal. An error is generated by the EDS circuit if the MSFF and latch outputs differ or if metastability is detected at the MSFF output. In Fig. 2, a 2-to-1 MUX logic gate precedes the Razor I EDS circuit to enable local error recovery [8, 9], which is discussed further in Section 3.1. For a global error-recovery design as described in Section 3.2, the 2-to-1 MUX is removed.

The metastability detector determines if the MSFF output remains metastable for a specified duration. The metastability detector consists of N-skewed and P-skewed inverters and a dynamic XOR logic gate [9]. The MSFF output feeds the inputs of the two skewed inverters. The skewed inverter outputs drive the dynamic XOR logic gate. If a sufficient metastable condition occurs at the MSFF output, the transition through one of the skewed inverters is faster than the other, thus discharging the dynamic XOR output to generate an error. The key challenge is designing the N- and P-skewed inverters for the appropriate metastability duration at the MSFF output. If the metastability duration is designed too short, then normal MSFF output transitions with early arriving input data could induce an error from the metastability detector, resulting in a functional failure. Consequently, the design of the metastability duration would require a delay guardband to ensure correct sequential operation within the presence of WID process variations. If the metastability duration is designed too long, then the CLK-to-Q delay push-out from a metastable MSFF output could result in an undetected error as described in Section 2.1. Although a metastable MSFF output can induce metastability on the error path, error-path metastability is much simpler to manage as compared to datapath metastability [11]. The inclusion of the metastability detector increases the clock energy and area overhead relative to the DS EDS circuit in Fig. 1c.

2.3 Transition Detector with Time Borrowing (TDTB)

The EDS circuit in Fig. 3a is a transition detector with a time-borrowing latch (TDTB). The TDTB EDS circuit operation is demonstrated through a simulated timing diagram in Fig. 3b. The transition detector monitors input data (D) transitions during the high clock phase. As input data transitions, a pulse is always generated at the XOR output. During the low clock phase, the output of the dynamic gate pre-charges and the pulse does not affect the error signal (ERROR) as described in Fig. 3b. If input data arrives late, CLK is logically-high and the pulse discharges the output node voltage of the dynamic gate, thus transitioning

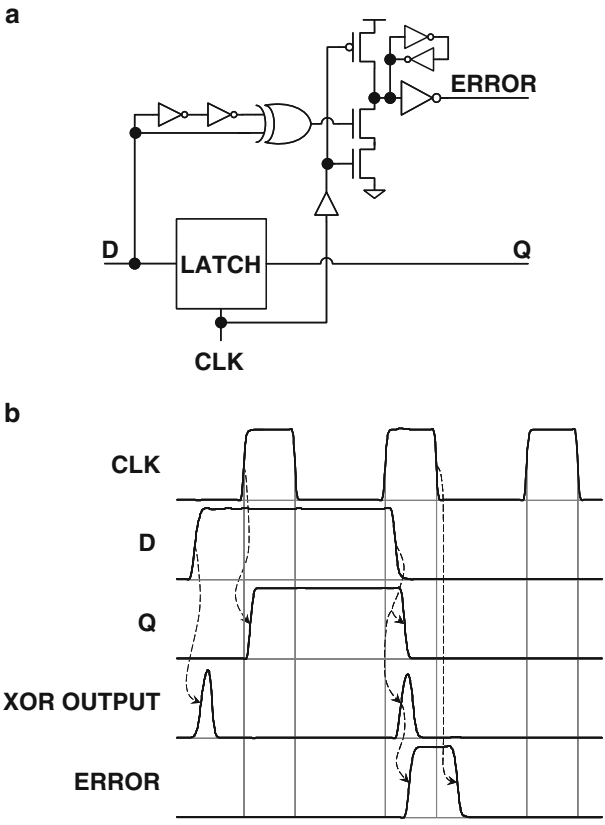


Fig. 3 (**a**) Transition detector with time borrowing (TDTB) EDS circuit [10] and (**b**) simulated timing diagram to demonstrate error generation from late arriving input data [11]. CLK is duty-cycle controlled to satisfy min-delay requirements (© 2009 IEEE)

ERROR to a logic-high as illustrated in Fig. 3b. As CLK transitions to a logic-low, the dynamic gate output pre-charges, and consequently, ERROR transitions to a logic-low. As discussed further in Section 3.2, ERROR is propagated to a set-dominant latch (SDL), where the SDL output remains logically-high while the dynamic transition detector pre-charges during the low clock phase. The SDL is transparent during the high clock phase and only allows high transitions during the low clock phase. Since min-delay paths are designed with sufficient margin as described in (2), the master latch of a datapath flip-flop is unnecessary. The datapath latch is identical to a pulse-latch, resulting in lower clock energy and eliminating datapath metastability during a rising clock edge. Datapath metastability does not occur on the falling clock edge since the max-delay constraint in (1) is satisfied.

Although TDTB employs a datapath latch, path timing constraints are still based on a flip-flop design with an error-detection window as illustrated in Fig. 1 and

modeled in (1). The purpose of the transparency window in the datapath latch is to eliminate datapath metastability while detecting timing errors. When input data arrives late, an error signal is generated even though the input data traverses to the latch output. The error signal ensures that late arriving data from the path in the current pipeline stage does not affect the max-delay constraint in (1) for adjoining fan-out paths in subsequent pipeline stages. If ample max-delay margin is available for the adjoining paths in the subsequent pipeline stage, then a pulse-latch may replace the TDTB EDS circuit at the current pipeline stage. This would enable traditional time borrowing between the path in the current pipeline stage and the adjoining paths in the subsequent pipeline stage.

Although datapath metastability is removed in TDTB, the transition-detector output can become metastable, leading to metastability in the error path. Metastability, however, is drastically simpler to manage in the error path as compared to the datapath [11]. For error-path metastability to occur, the input data must arrive close to a rising clock edge. This condition defines the boundary of a timing failure. Latch transparency allows input data to continue to the next pipeline stage. The error path for TDTB, which is described in more detail in Section 3.2, consists of an OR tree of error signals from each TDTB EDS circuit in the pipeline stage. The OR-tree output feeds an SDL, and the SDL output is the final error signal (FINAL ERROR). The SDL maintains the logic-high value for FINAL ERROR when the dynamic transition detector pre-charges during the low clock phase. FINAL ERROR is an input to an MSFF, and the MSFF output is the pipeline-error signal (PIPELINE ERROR). As long as the PIPELINE ERROR resolves to either a logic-high, resulting in error recovery, or a logic-low, resulting in no error recovery, correct functionality is maintained. With this unique characteristic, the error path behaves similar to a traditional synchronizer circuit with the exception of having combinational logic in the middle of the sequentials [13–15]. Based on an extreme worst-case metastability calculation for a microprocessor with TDTB EDS circuits, the MTBF from error-path metastability is over 10^{10} larger than microprocessor MTBF targets for soft-error rate (SER) in a 65 nm technology [11].

Since the dynamic transition detector is highly sensitive to WID process variations, the TDTB design is fairly complex. In designing the pulse width for the TDTB transition detector, the key trade-off is the transition-detector setup time ($T_{setup,clk\text{-}r}$) versus a sufficient pulse width to detect late arriving data within the presence of WID variations. For EDS circuits with a datapath latch, $T_{setup,clk\text{-}r}$ is defined as the minimum data-to-clock (D-to-CLK) delay prior to a rising clock edge such that an error signal is not generated (i.e., an error signal indicates that the input data transition did not satisfy $T_{setup,clk\text{-}r}$). As the pulse width increases, the tolerance of the dynamic transition detector to WID variations improves at a cost of a larger $T_{setup,clk\text{-}r}$.

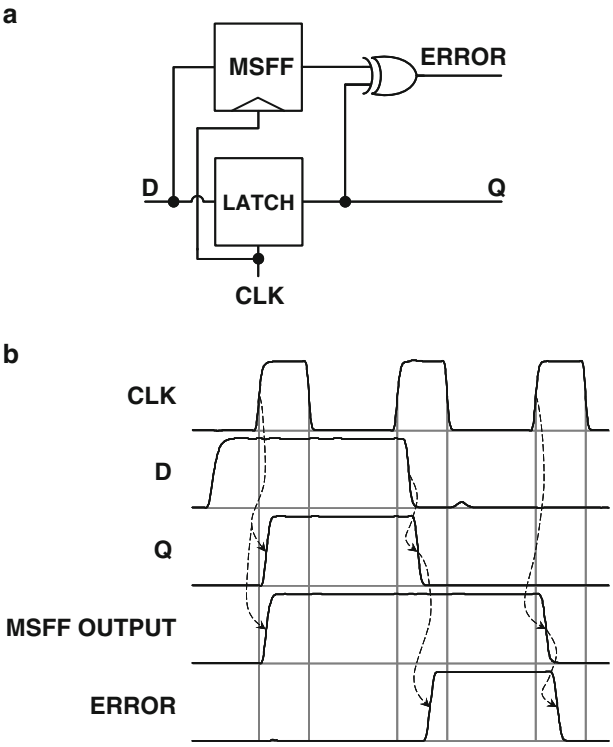


Fig. 4 (**a**) Double sampling with time borrowing (DSTB) EDS circuit [10] and (**b**) simulated timing diagram to demonstrate error generation from late arriving input data [11]. CLK is duty-cycle controlled to satisfy min-delay requirements (© 2009 IEEE)

2.4 Double Sampling with Time Borrowing (DSTB)

The EDS circuit in Fig. 4a is double sampling with a time-borrowing latch (DSTB), which is similar to TDTB except a shadow MSFF replaces the transition detector. In Fig. 4b, a simulated timing diagram demonstrates the DSTB EDS circuit operation. DSTB double samples input data similar to the DS EDS circuit in Fig. 1c and compares datapath latch and shadow MSFF outputs to generate an error signal while retaining the time-borrowing feature of TDTB to eliminate datapath metastability.

As described earlier for TDTB, the DSTB error signal can become metastable. In contrast to TDTB, the DSTB error path does not contain an SDL. The error path for DSTB, which is described in Section 3.2, is an OR tree of error signals from each DSTB EDS circuit in the pipeline stage. The OR-tree output (FINAL ERROR) feeds an MSFF, and the MSFF output is PIPELINE ERROR. The MTBF from error-path metastability in DSTB improves relative to that in TDTB since the SDL

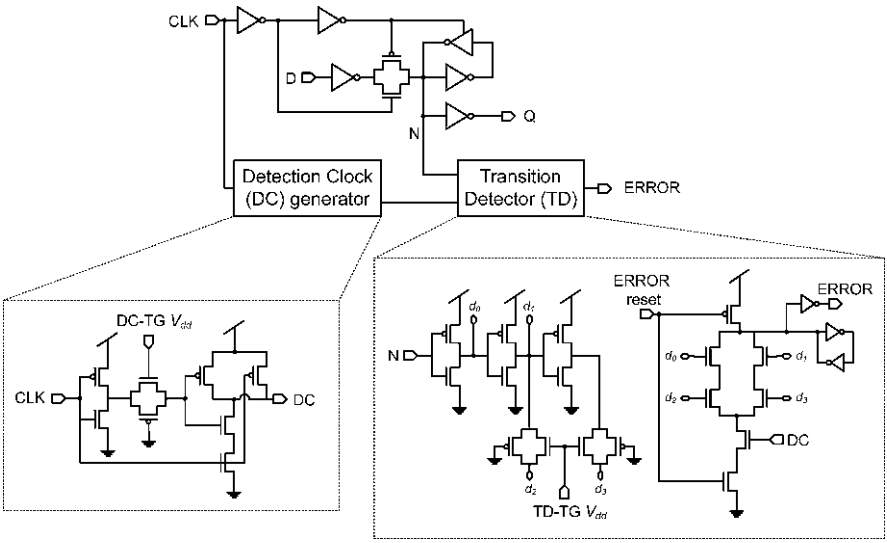


Fig. 5 Razor II EDS circuit [12]. CLK is duty-cycle controlled to satisfy min-delay requirements (© 2008 IEEE)

is removed. The MTBF from error-path metastability in DSTB for a microprocessor is over ten orders of magnitude larger than the MTBF targets for SER [11].

2.5 Razor II

The Razor II EDS circuit [12], which contains features similar to TDTB, is described in Fig. 5. Razor II and TDTB EDS circuits both utilize a datapath latch and dynamically detect late data transitions. The key difference between these two EDS circuits is that TDTB detects late transitions at the latch input where data transitions are only monitored during the high clock phase; Razor II detects late transitions at the latch pass gate output, denoted as N in Fig. 5, where data transitions are continuously monitored except for a window of time, defined as $T_{cq,max}$ [12], after the rising clock edge. In Razor II, the transition detector is suspended for $T_{cq,max}$ after the rising clock edge to guarantee that early input data transitions do not induce a timing error. The transition detector is suspended by generating a negative clock pulse to the dynamic transition detector. $T_{cq,max}$, the negative-clock-pulse width, must be larger than the CLK-to-N delay with a zero setup time plus the data-pulse width to the transition detector. WID process variations amplify the complex timing in Razor II, requiring a sufficient delay guardband on the negative-clock-pulse width. This delay guardband results in extra clock inverters, and consequently, larger clock energy overhead. Similar to TDTB and DSTB, Razor II removes datapath metastability. Although the error path can become metastable,

the MTBF from error-path metastability in Razor II for a microprocessor is expected to be similar to that of DSTB.

2.6 Advantages and Disadvantages

A salient feature of TDTB, DSTB, and Razor II EDS circuits relative to the DS and Razor I EDS circuits is moving the highly complex metastability issue from both the datapath and error path to only the error path, thus drastically simplifying metastability management. For static-EDS circuits, the DSTB clock energy overhead is slightly lower than for the DS EDS circuit since a datapath sequential is typically sized larger than a minimum-sized shadow sequential. The DSTB clock energy savings improve when compared to Razor I due to the clock energy overhead from the metastability detector [8, 9]. Since the transition detector clock energy is less than the shadow MSFF clock energy, the clock energy for TDTB is lower than for DSTB [10]. Relative to TDTB, additional clock transistors (e.g.,inverters and NAND gate) are required in Razor II to generate the negative clock pulse as illustrated in Fig. 5, resulting in larger clock energy for Razor II. The clock energy overhead comparison between Razor II and DSTB depends on the additional clock energy required to generate the negative clock pulse for Razor II versus the additional clock energy in the shadow MSFF for DSTB. With a datapath latch, the D-to-Q delays for TDTB, DSTB, and Razor II appear relatively similar, which are faster than the DS and Razor I D-to-Q delays with a datapath MSFF.

As discussed in Sections 2.2, 2.3, 2.4, and 2.5, the design complexity of DS and DSTB is lower than for Razor I with the dynamic metastability detector and for TDTB and Razor II with dynamic transition detectors, which are highly sensitive to WID process variations. DS, Razor I, DSTB, and Razor II provide SER protection for the datapath sequential during the entire clock cycle. In comparison, TDTB offers SER protection for the datapath sequential during the high clock phase and only a portion of the low clock phase. When introducing a new circuit technique into a production microprocessor, the ability to turn the functionally off is highly desirable in case unforeseen complexities arise. With DS and Razor I, the error detection capability could be turned off and operated at a low F_{CLK} with a 50% duty cycle. With TDTB, DSTB, and Razor II, a duty-cycle control would always be required for both high and low F_{CLK}. In comparison to DS, Razor I, TDTB, and DSTB, the Razor II the error-detection window shrinks by the negative-clock-pulse width, resulting in less opportunity to detect late arriving input data for a target min-delay constraint as defined by the high clock phase.

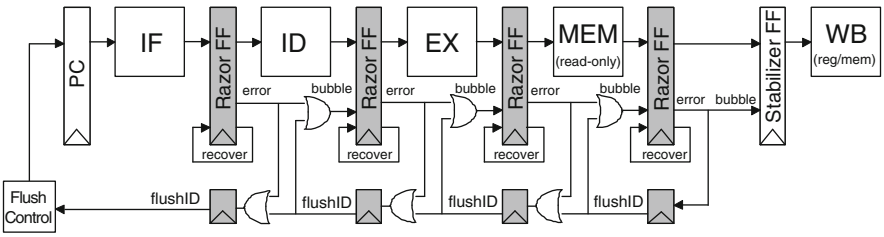


Fig. 6 Counter-flow pipeline error recovery [8, 9] (© 2003 IEEE)

3 Error-Recovery Circuits

3.1 Counter-Flow Pipeline Error Recovery

Error recovery based on counter-flow pipelining is illustrated in Fig. 6 [8, 9]. Error signals from each EDS circuit per pipeline stage are combined via an OR tree to generate a single error signal, which is also the restore signal in Fig. 2 and the recover signal in Fig. 6. As described in Fig. 2, the restore signal feeds the select of the 2-to-1 MUX prior to the EDS circuit in this error-recovery approach. The restore signal is propagated to all sequentials in the pipeline stage, both the EDS circuits and the conventional flip-flops, to select between the normal datapath value and the correct value from the previous cycle. This design imposes significant timing restrictions on the error-signal propagation delay and requires additional area for interconnect routing tracks. When an error is detected, the correct logic value is inserted back into the pipeline to enable single-cycle error recovery, ensuring forward progress for late arriving input data. The erroneous pipeline stage data is nullified via the bubble signal in Fig. 6. The bubble signal indicates an empty pipeline slot for the subsequent pipeline stages. A stabilization pipeline stage precedes the write back (WB) pipeline stage to allow the 1-cycle latency for propagating the bubble signal. This stabilization pipeline stage ensures that instructions are error-free before committing the state at the WB stage. A flush signal propagates the stage ID of the failing instruction to the beginning of the pipeline, where the flush control logic restarts the pipeline at the instruction following the failing instruction.

3.2 Instruction-Replay Error Recovery

In contrast to the single-cycle counter-flow pipeline error-recovery design [8, 9], a multi-cycle error-recovery design based on instruction replay significantly reduces the design overhead [10–12]. A test-chip implementation of the instruction-replay

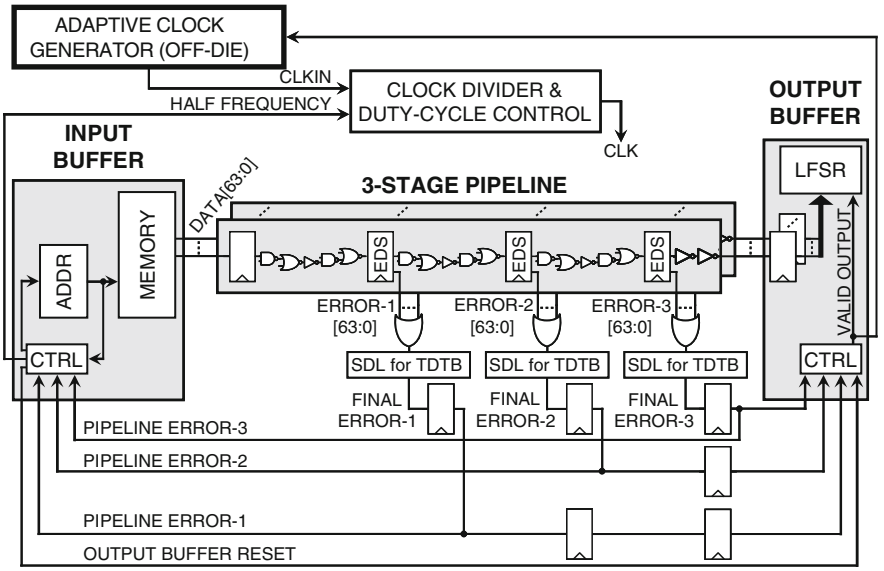


Fig. 7 Instruction-replay error-recovery design [10]. Set-dominant latch (SDL) is only used for TDTB (© 2008 IEEE)

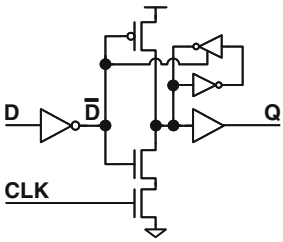


Fig. 8 Set-dominant latch (SDL) circuit schematic [11] (© 2009 IEEE)

error-recovery design is described in Fig. 7. An input buffer drives data and control signals to a three-stage pipeline circuit block to imitate a microprocessor. A stabilization pipeline stage [8, 9] follows the three-stage pipeline circuit block to accommodate the 1-cycle latency for propagating error signals in the third stage. This stabilization pipeline stage ensures that instructions are error-free before committing the state to the output buffer in the next pipeline stage. The output buffer consists of a linear-feedback-shift register (LFSR) that compresses output data into a unique signature to validate functionality. The unlabeled sequentials in Fig. 7 represent conventional MSFFs.

In the three-stage pipeline circuit block, error signals from each EDS circuit per pipeline stage are combined via an OR tree to generate a single error signal (FINAL ERROR) [8, 9]. For DS, Razor I, DSTB, and Razor II, the OR-tree output directly

feeds an MSFF. For TDTB, the OR-tree output feeds an SDL, where the schematic is provided in Fig. 8. The SDL output is FINAL ERROR, which is an input to an MSFF. The SDL in Fig. 8 is transparent during the high clock phase and only allows high transitions during the low clock phase. If FINAL ERROR transitions to a logic-high, the SDL maintains the logic-high value for FINAL ERROR when the TDTB EDS circuit pre-charges during the low clock phase. For each EDS circuit type, the output of the MSFF represents the pipeline-error signal (PIPELINE ERROR).

As illustrated in Fig. 7, the three pipeline-error signals are propagated to the input buffer in one cycle to replay the failed instruction and pipelined to the output buffer to invalidate erroneous data. Input buffer control logic determines the appropriate instruction to replay based on the three pipeline-error signals. In a microprocessor, the instruction replay circuits could leverage the existing replay design to recover from a branch miss-prediction [18]. If a pipeline-error signal transitions to a logic-high, the input buffer signals the clock divider to halve F_{CLK} while maintaining a constant high clock phase delay for min-delay protection. Reducing F_{CLK} in half ensures correct operation during replay even if dynamic variations persist. After the replayed instruction finishes, the input buffer sends a reset signal to validate output data and signals the clock divider to resume at target F_{CLK}. Since F_{CLK} is halved for all but one of the recovery cycles, the number of actual (effective) cycles for recovery is 6 (11), 7 (13), and 8 (15) corresponding to timing errors in the first, second, and third pipeline stages. Since the number of recovery cycles linearly increases with the number of pipeline stages, the average

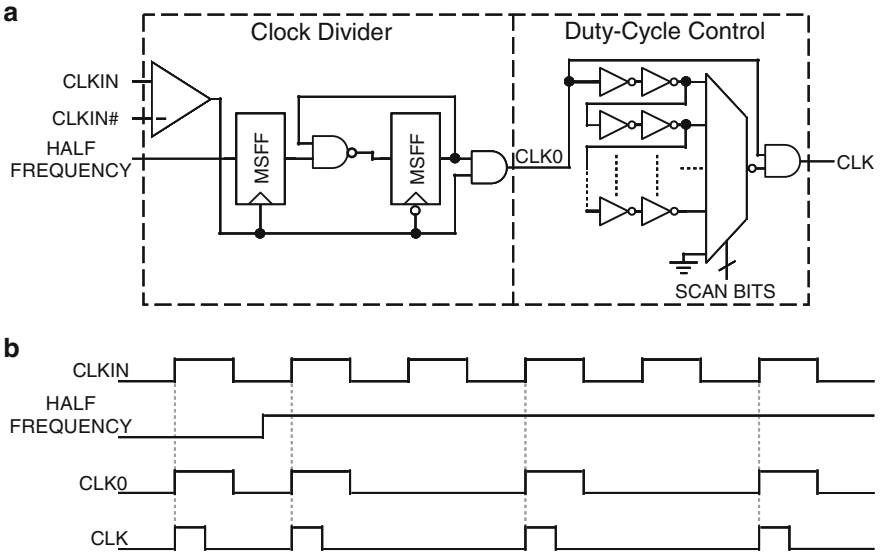


Fig. 9 (**a**) Clock divider and duty-cycle control circuit schematics and (**b**) conceptual timing diagrams [11] (© 2009 IEEE)

error-recovery penalty for a microprocessor is expected to linearly increase as compared to the test-chip implementation. If dynamic variations persist for long durations, an adaptive off-die clock generator adjusts the nominal operating F_{CLK}.

3.3 Clock Divider and Duty-Cycle Control Circuits

Clock divider and duty-cycle control circuits for the instruction-replay error-recovery design are presented in Fig. 9a along with a conceptual timing diagram in Fig. 9b. An off-die signal generator with a differential pulse-splitter creates differential inputs CLKIN and CLKIN#. The HALF FREQUENCY input is controlled by the input buffer as described in Fig. 7. The CLK output is distributed throughout the test-chip. CLKIN and CLKIN# are inputs to a differential amplifier that generates an intermediate clock signal. This intermediate clock signal and the output of the negative edge-triggered MSFF in Fig. 9a are inputs to a logic-AND gate to produce the clock divider output (CLK0). When the HALF FREQUENCY input is a logic-low, the output of the negative edge-triggered MSFF remains a logic-high, thus CLK0 and CLKIN have the same frequency. When the HALF FREQUENCY input is asserted, the output of the negative edge-triggered MSFF toggles every other cycle, enabling the clock divider circuit to skip every other high phase of CLKIN as illustrated in Fig. 9b. The duty-cycle control is performed with a logical-AND of CLK0 and a delayed CLK0# (i.e., inversion of CLK0) with CLK as the output. The

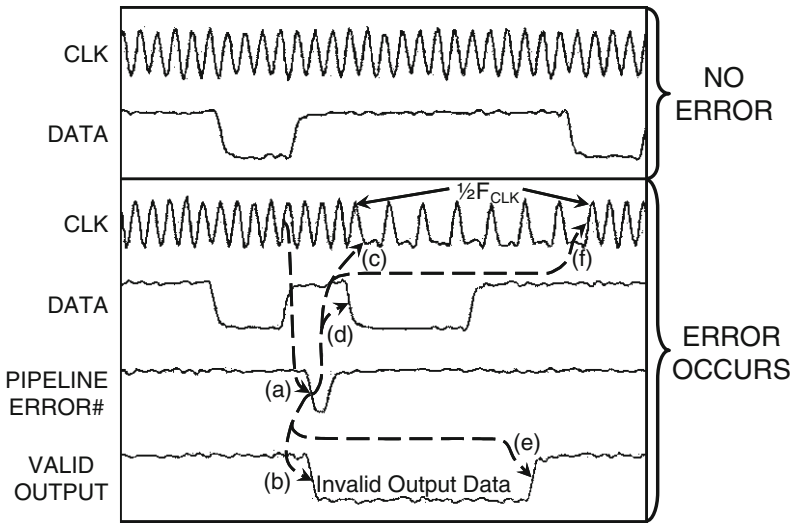


Fig. 10 Measured error detection and recovery demonstration for TDTB EDS circuits and the instruction-replay error-recovery design in Fig. 7 [10]. Data from input buffer arrives late at third pipeline stage; (**a**) Detect error; (**b**) Invalidate output data; (**c**) Halve F_{CLK}; (**d**) Replay instruction; (**e**) Validate output data; and (**f**) Resume target F_{CLK} (© 2008 IEEE)

delayed CLK0# determines the CLK high phase delay, as controlled via scan bits. With this duty-cycle control circuit, the CLK high phase delay remains constant at both high and low F_{CLK} values, which is essential for min-delay protection.

4 Resilient Circuit Measurements

4.1 Resilient Circuit Demonstration

A timing-error detection and recovery measurement is demonstrated in Fig. 10 based on an oscilloscope capture of test-chip output signals for TDTB EDS circuits and the instruction-replay error-recovery design in Fig. 7. CLK is the clock signal distributed throughout the test-chip;

DATA is 1 bit of data sent from the input buffer to the three-stage pipeline circuit block; PIPELINE ERROR# is the logic-NOR output of the three pipeline-error signals; and VALID OUTPUT is the output buffer valid signal. In the demonstration with no timing errors, DATA transitions from a logic-high to a logic-low for four cycles and then transitions to a logic-high. In the demonstration with timing errors, the high-to-low transition from DATA induces a timing error at the third pipeline stage; (a) Error is detected by the EDS circuit and propagated to the pipeline-error signal to lower PIPELINE ERROR# for one cycle; (b) VALID OUTPUT transitions low to invalidate erroneous data; (c) Pipeline-error signal triggers the input buffer to raise the HALF FREQUENCY signal in Figs. 7 and 9 to halve F_{CLK}; (d) Input buffer replays the instruction, resulting in a high-to-low DATA transition; (e) Once the instruction finishes, the input buffer resets the VALID OUTPUT signal to allow output data to reach the LFSR; (f) Finally, input buffer lowers the HALF FREQUENCY signal to resume at target F_{CLK}. In Fig. 10, note that the CLK high phase delay remains constant at both high and low F_{CLK} for min-delay protection.

4.2 Resilient Circuit Performance and Power Benefits

Imitating a microprocessor, instruction kernels are executed on a resilient circuit test-chip in 65 nm technology [19] to compare a resilient design with EDS circuits to a conventional design with MSFFs [10, 11]. The benefits from resilient circuits depend on the path-activation probabilities as determined by the instruction kernels. Although previous research has investigated node-activity probabilities for a microprocessor [20], node-activity probabilities do not directly translate into path-activation probabilities. Since path-activation probabilities have not been rigorously explored for a microprocessor, there is uncertainty about the dependency of path-activation probability on path delay. Since slow paths typically

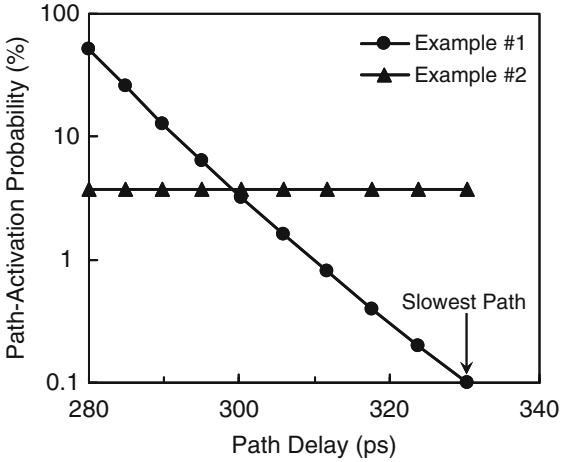


Fig. 11 Path-activation probability versus path delay for two path-activation examples [11] (© 2009 IEEE)

contain more logic depth than fast paths, the probability of activating slow paths is intuitively expected to be less than the activation probability for fast paths in general. In addressing this issue, two sets of instruction kernels are selected to induce the path-activation examples presented in Fig. 11, which attempt to represent practical approximations of favorable (example #1) and unfavorable (example #2) scenarios. In path-activation example #1, critical paths are activated less frequently than non-critical paths, where the activation probability exponential reduces from fast to slow paths. Path-activation example #2 applies an equal activation probability for all paths.

In Fig. 12, throughput (TP) and error rate are measured for a resilient design with TDTB EDS circuits and instruction-replay error recovery versus F_{CLK} for the two path-activation examples in Fig. 11. For a given F_{CLK}, error rate is governed by the path histogram, as dictated by design optimization, as well as path-activation probabilities and environmental variations, as determined by workloads. A range of V_{CC} droop magnitudes and durations are inserted via on-chip noise injectors based on data from a recent microprocessor along with assumptions on V_{CC} droop-inducing events. The worst-case V_{CC} droop magnitude is 10% and the worst-case temperature is 110°C. Nominal V_{CC} is 1.2 V and the nominal operating temperature is assumed 60°C. In Fig. 12a, throughput increases linearly as F_{CLK} increases with no errors. Once errors occur, instructions per cycle (IPC) reduce as a function of error rate and recovery time. Since V_{CC} droop events are assumed infrequent, throughput gains continue as F_{CLK} increases into the V_{CC} and temperature guardband region. When F_{CLK} reaches 3020MHz, the first path failure occurs under nominal conditions, resulting in a sharp error rate increase. Since the path-activation probability for example #1 is low for slow paths, further throughput gains are achieved at higher F_{CLK}. The maximum throughput of 3.17 billion instructions per second (BIPS) corresponds to a 3,200 MHz F_{CLK}. Increasing F_{CLK} further leads to a larger error

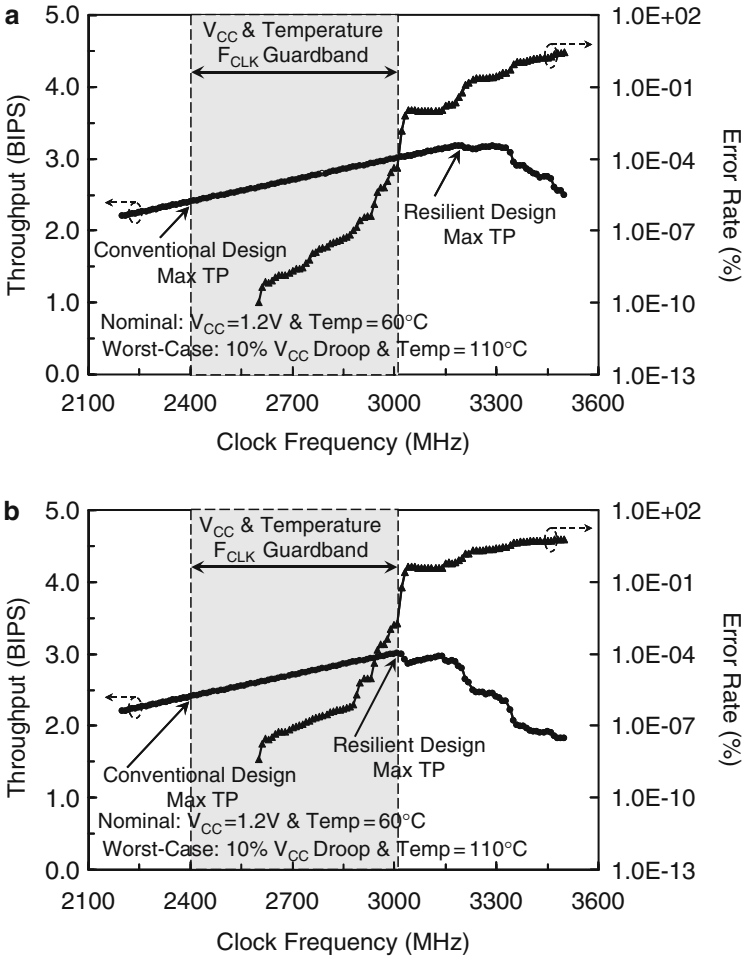


Fig. 12 Measured throughput (TP) and error rate for a resilient design with TDTB EDS circuits and instruction-replay error recovery versus clock frequency (F_{CLK}) for path-activation (**a**) example #1 and (**b**) example #2 [10] (© 2008 IEEE)

rate, where IPC reduction outweighs F_{CLK} gains. Due to the high path-activation probability for slow paths in example #2, the resilient design cannot exploit the path-activation probabilities in Fig. 12b, limiting the maximum throughput to 3.01BIPS. In Fig. 12a and b, the maximum throughput to guarantee correct functionality within the presence of worst-case dynamic V_{CC} and temperature variations for the conventional design with MSFFs is 2.4BIPS, corresponding to an F_{MAX} of 2,400 MHz. From Fig. 12, a resilient design enables 25% throughput gain over a conventional design by eliminating the F_{CLK} guardband from dynamic V_{CC} and temperature variations and

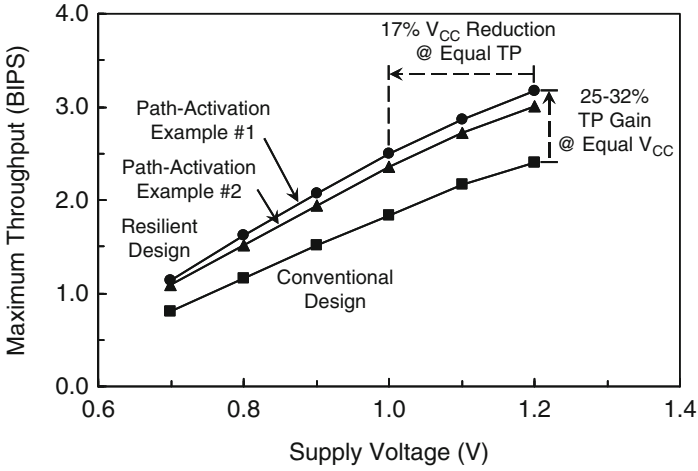


Fig. 13 Measured maximum throughput (TP) for a resilient design with TDTB EDS circuits and instruction-replay error recovery as well as a conventional design with MSFFs versus supply voltage [10] (© 2008 IEEE)

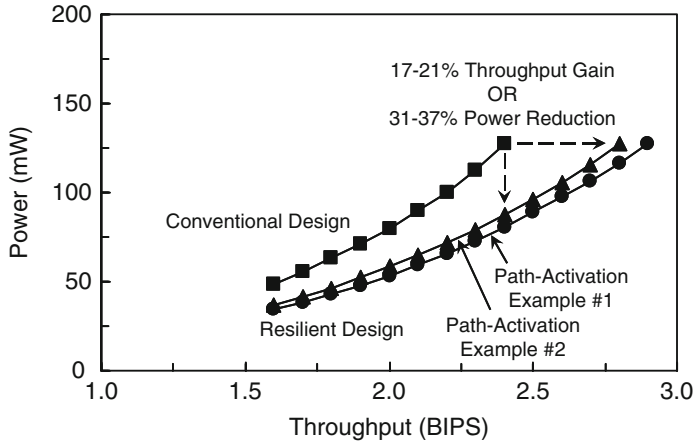


Fig. 14 Measured power for a resilient design with TDTB EDS circuits and instruction-replay error recovery as well as a conventional design with MSFFs versus throughput

an additional 7% throughput increase from exploiting the path-activation probabilities for example #1.

In Fig. 13, the maximum throughputs for resilient and conventional designs are measured versus V_{CC}. Since a worst-case V_{CC} droop of 10% is applied in these

measurements, the maximum V_{CC} droop reduces in absolute value as the nominal V_{CC} reduces. For the resilient design, the two path-activation examples in Fig. 11 are evaluated. Depending on the path-activation example in Fig. 13, the throughput for the resilient design with $V_{CC} = 1.0$ V is greater than or equal to the throughput of the conventional design with $V_{CC} = 1.2$ V. Thus, measured data indicates that resilient circuits provide either 25–32% throughput gain at equal V_{CC} or at least a 17% V_{CC} reduction at equal throughput [10] for a target V_{CC} of 1.2 V in a 65 nm technology.

In Fig. 14, total power versus throughput is measured for resilient and conventional designs. *In comparison to a conventional design, silicon measurements indicate that resilient circuits enable a 17–21% throughput gain at iso-power or a 31–37% power reduction at iso-throughput.* The potential benefits of a resilient microprocessor will differ from the test-chip depending on the fraction of total sequentials protected with EDS circuits and the amount of buffer insertion required to correct min-delay paths.

5 Conclusion

Resilient circuits with timing-error detection and error recovery mitigate the clock frequency (F_{CLK}) guardband from dynamic supply voltage (V_{CC}) and temperature variations and exploit path-activation probabilities for maximizing performance efficiency. Five error-detection sequential (EDS) circuits are reviewed: (a) Double sampling (DS), (b) Razor I, (c) Transition detector with time borrowing (TDTB), (d) Double sampling with time borrowing (DSTB), and (e) Razor II. A salient feature of TDTB, DSTB, and Razor II relative to DS and Razor I is moving the highly complex metastability issue from both the datapath and error path to only the error path, thus drastically simplifying metastability management. TDTB is the lowest clock energy EDS circuit; DSTB is the lowest clock energy static-EDS circuit with SER protection. Two error-recovery designs are presented with different trade-offs in recovery cycles and design overhead. The counter-flow pipeline design allows single-cycle error recovery. In contrast, an instruction-replay design trades-off an increase in error-recovery cycles for significantly less design overhead by replaying failing instructions at half F_{CLK} to ensure correct functionality even if dynamic variations persist. Silicon measurements indicate that resilient circuits enable 25–32% throughput gain at equal V_{CC} as compared to conventional circuits, resulting in a 17–21% throughput gain at iso-power or a 31–37% power reduction at iso-throughput. Recommendations to further enhance the performance and energy efficiency of resilient circuits include: (a) Combining resilient circuits with on-die variation sensors and adaptive-F_{CLK} schemes, (b) Mitigating delay faults induced from transistor aging, cross-coupling capacitance, and multiple-input switching, and (c) Optimizing resilient circuit designs by coupling the path-delay histogram with path-activation probabilities.

References

1. A. Muhtaroglu, G. Taylor, T. R. Arabi, On-die droop detector for analog sensing of power supply noise. *IEEE Journal of Solid-State Circuits*, Apr 2004, pp. 651–660
2. T. Fischer, J. Desai, B. Doyle, S. Naffziger, B. Patella, A 90-nm variable frequency clock system for a power-managed itanium architecture processor. *IEEE Journal of Solid-State Circuits*, Jan 2006, pp. 218–228
3. R. McGowen et al., Power and temperature control on a 90-nm itanium family processor. *IEEE Journal of Solid-State Circuits*, Jan 2006, pp. 229–237
4. J. Tschanz et al., Adaptive frequency and biasing techniques for tolerance to dynamic temperature-voltage variations and aging, in *IEEE ISSCC Digest of Technical Papers*, Feb 2007, pp. 292–293
5. P. Franco, E.J. McCluskey, Delay testing of digital circuits by output waveform analysis, in *Proceedings of the IEEE International Test Conference*, Oct 1991, pp. 798–807
6. P. Franco, E.J. McCluskey, On-line testing of digital circuits, in *Proceedings of the IEEE VLSI Test Symposium*, Apr 1994, pp. 167–173
7. M. Nicolaidis, Time redundancy based soft-error tolerance to rescue nanometer technologies, in *Proceedings of the IEEE VLSI Test Symposium*, Apr 1999, pp. 86–94
8. D. Ernst et al., Razor: A low-power pipeline based on circuit-level timing speculation, in *Proceedings of the IEEE/ACM International Symposium Microarchitecture (MICRO-36)*, Dec 2003, pp. 7–18
9. S. Das et al., A self-tuning DVS processor using delay-error detection and correction, *IEEE Journal of Solid-State Circuits*, Apr 2006, pp. 792–804
10. K.A. Bowman et al., Energy-efficient and metastability-immune timing-error detection and instruction-replay-based recovery circuits for dynamic-variation tolerance, in *IEEE ISSCC Digest of Technical Papers*, Feb 2008, pp. 402–403
11. K.A. Bowman et al., Energy-efficient and metastability-immune resilient circuits for dynamic variation tolerance. *IEEE Journal of Solid-State Circuits*, Jan 2009, pp. 49–63
12. D. Blaauw et al., Razor II: In situ error detection and correction for PVT and SER tolerance, in *IEEE ISSCC Digest of Technical Papers*, Feb. 2008, pp. 400–401
13. H.J.M. Veendrick, The behavior of flip-flops used as synchronizers and prediction of their failure rate. *IEEE Journal of Solid-State Circuits*, Apr 1980, pp. 169–176
14. C.L. Portmann, T.H.Y. Meng, Metastability in CMOS library elements in reduced supply and technology scaled applications. *IEEE Journal of Solid-State Circuits*, Jan 1995, pp. 39–46
15. C. Dike, E. Burton, Miller and noise effects in a synchronizing flip-flop. *IEEE Journal of Solid-State Circuits*, June 1999, pp. 849–855
16. V. Srinivasan et al., Optimizing pipelines for power and performance, in *Proceedings of the International Symposium of Microarchitecture (MICRO-35)*, Nov 2002, pp. 333–344
17. A. Hartstein, T.R. Puzak, The optimum pipeline depth considering both power and performance. *ACM Transactions on Architecture and Code Optimization (TACO)*, Dec 2004, pp. 369–388
18. J. Hennessy, D. Patterson, *Computer Architecture a Quantitative Approach*, 2nd edn. (Morgan Kaufmann Publishers, San Francisco, CA, 1996)
19. P. Bai et al., A 65nm logic technology featuring 35nm gate lengths, enhanced channel strain, 8 Cu interconnect layers, low-k ILD and 0.57 μm^2 SRAM cell, in *IEEE IEDM Technical Digest*, Dec 2004, pp. 657–660
20. H.L. Yeager, M.J. Patyra, R. Reyes, K.A. Bowman, Microprocessor power optimization through multi-performance device insertion, in *IEEE Symposium on VLSI Circuits Digest of Technical Papers*, June 2004, pp. 334–337

Process Variability-Induced Timing Failures – A Challenge in Nanometer CMOS Low-Power Design

Xiaonan Zhang and Xiaoliang Bai

1 Introduction

With various powerful features integrated in mobile devices, the demand for high-density, low-power, and low-leakage design keeps elevating. In 65 nm technologies and below, process intra-die variability becomes a prominent design factor.Two types of intra-die variability exist: one is spatially correlated and the other is random. In this chapter, we focus on intra-die random variability, shorted as *the variability* and its impact on circuit timing failure. The variability is independent in nature. It can have a large impact on circuit performance and robustness. This is especially true for nanometer low-power (LP) CMOS designs, where the over-drive voltage is less than two times of the threshold voltage. As the voltage scales down, circuits become increasingly sensitive to the variability. As a result, circuit design for low-power process technologies with low supply voltages becomes extremely challenging.

Numerous papers on the effects of process variability on digital circuits have been published. However, these papers focus on the maximum frequency under process variations [2, 3, 4]. In this work, we study the timing skews caused by the variability and their impact on circuit functionality. Intra-die variability causes some paths to be slower and other paths to be faster. When the timing skew is greater than what a circuit can tolerate, the circuit has failures. The impact of process variability on clock skew has also been studied [9]. However, studies have not been done on the effects of variability-induced timing skews in low-power designs. In order to design variability-resilient circuits, we must understand the effects of process variability on low-power design and quantify the impact. Then we can develop circuit design and timing verification methodologies that ensure circuits robustness.

X. Zhang (✉) and X. Bai
Qualcomm, Inc., San Diego, CA, USA
e-mail: xiaoliangbai@gmail.com

A. Amara et al. (eds.), *Emerging Technologies and Circuits*,
Lecture Notes in Electrical Engineering 66, DOI 10.1007/978-90-481-9379-0_12,

Process variability-induced scan-chain hold-time failure has been reported [7]. We have also observed hold-time failures during scan testing at various voltage levels. It is well-known that process variability determines the Vccmin of SRAM. However, for digital circuits what is the minimum voltage? What types of failures will we see at low Vdd? In this work, we attempt to answer these questions. We focus on circuit functionality and design robustness affected by the variability-induced timing skews.

2 Effects of Process Variability on Timing

In this section, we first study the effects of the variability on digital circuit timing and then investigate the most sensitive dependencies of circuit delay variation.

There are many sources of variability. Manufacturing process factors include discrete doping, line roughness, fluctuation in light exposure and etching procedure. The combined effect is a statistical performance distribution of final products. Non-process variables such as temperature, power supply noise, particle strike and hot carries effect also cause performance variations.

From timing analysis point of view, performance variation from slow, slow corner (ss) to fast, fast corner (ff) can be partitioned into systematic variations and local random variations. These catalogs are shown in Fig. 1 conceptually. Wafer to wafer and die to die variations belong to systematic or "global" part. "Global" fast (slow) means on average transistors on one chip are faster (slower) than those on the other chip. Global corner represents an averaging performance distribution. Within a given chip, there are systematic and independent variations. Within die systematic variation is layout-dependent variations caused by optical

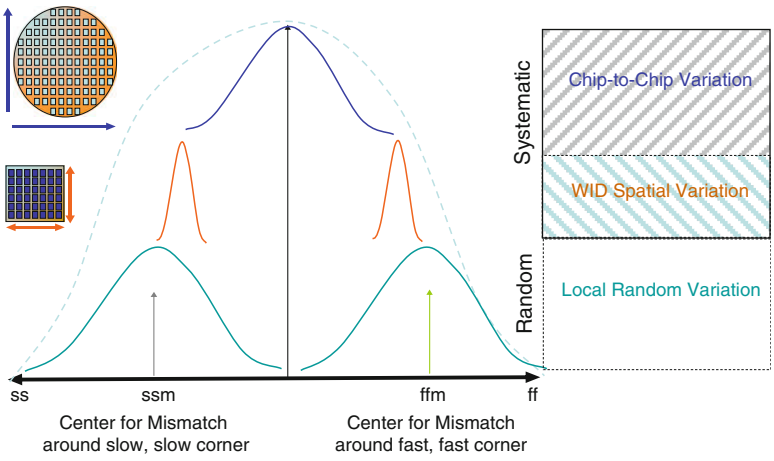


Fig. 1 Partition of variation

proximity effect (OPE), chemical-mechanical polishing (CMP), well proximity effect (WPE) and other location specific variations. This kind of variations is location and proximity layout dependent. It is modeled as spatial correlations. The third type of variations is local independent variations, caused by doping density fluctuation and line edge roughness, etc.

Local random variability is becoming the major source of pessimism in traditional STA timing flow. For 65 nm technology, we observed more than 11% performance loss due to local random variability. With feature size shrinking, the relative magnitude of local random variability is increasing. We will first study the impact of random variability on circuit delay.

2.1 Path Delay Variation

We begin with the most simple delay chain that consists of same type of gates. Figure 2 shows two example circuits with different number of logic stages.

Figure 2a shows a two-stage delay chain consists of simple inverters. The inverters are of the same type (size, p/n ratio). Monte Carlo simulations are performed around ssm corner. The 3-sigma worst case delay in Monte Carlo simulations and ss corner result (corresponding to late arrival time result in traditional STA) are also pictured. The Monte Carlo 3-sigma result resembles real life worst case path delay. We simply name it as real worst case. And the ss corner result is STA worst case. In this two-stage example, STA worst case closely resembles the real worst case.

Figure 2a shows timing results for late arrival time. For early arrival time, Monte Carlo simulations will be performed around ffm and the results will be compared with ff corner (STA best case corner). In this rest of this paper, we focus on late arrival time, but the analysis and result are equally applicable to its dual – the early arrival time.

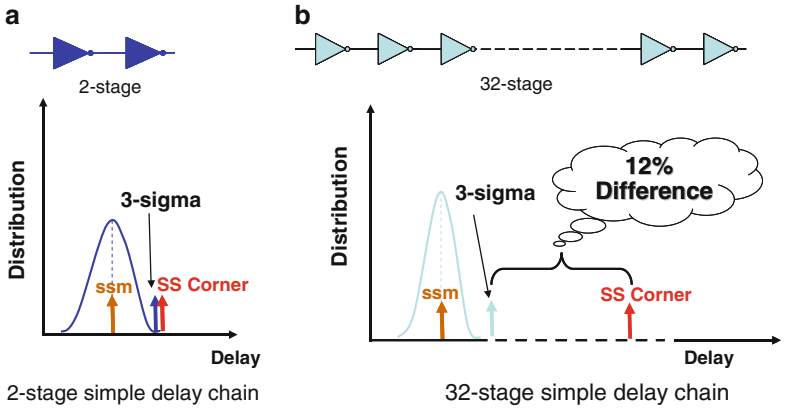


Fig. 2 Simple delay chain and delay variation

Figure 2b shows results on a 32-stage inverter chain. Here STA worst case and real worst case has about 10% discrepancy. The STA worst case is overly pessimistic. Since the physical variations along the delay chain are independent, some gates become slower, some gates become faster. The variations tend to average out each other. Suppose each stage has mean delay d_i and standard deviation σ_i. The variation versus delay of a simple delay chain can be expressed as

$$\frac{\sigma_{path}}{d_{path}} = \left(\frac{\sigma_1}{d_1}\right)\frac{n^{\frac{1}{2}}}{n} \tag{1}$$

in which n is the number of logic stages in the delay chain. As the number of stages n increases, the percentage variation decreases quickly. This is a very useful equation. In the first order, it tells us that as the number of stages increases, the variability decrease.

Under general circumstances, a delay path consists of different logic gates. Each stage *i* has a different mean delay value d_i and a different standard deviation σ_i. As a result, Eq. 1 is no longer accurate.

With the assumption that systematic variations are modeled by global models and systematic variations shift the center of local mismatch models. Local random variations are independent and path delay can be calculated by Root Sum Square (RSS). The average path delay and standard deviation are:

$$d_{path} = \sum_{i=1}^{n} d_{i\_avg} \tag{2}$$

$$\sigma_{path} = \sqrt{\sum_{i=1}^{n} \sigma_i^2} \tag{3}$$

Error exists in Eq. 1 for simple delay paths. Figure 3 shows Monte Carlo simulation results for a simple inverter chain and delay variations predicted by Eq. 1. In these Monte Carlo simulations, delay variations of each gate are caused by independent physical variables. Equation 1 models the variation stage dependency very well. However, looking into detail, the stage dependency does not follow Eq. 1 strictly, the trend is slightly off. Therefore, we propose an empirical formula to model the stage dependency with a power factor m:

$$\frac{\sigma_{path}}{d_{path}} = \left(\frac{\sigma_1}{d_1}\right)\frac{n^m}{n} \tag{4}$$

Therefore, Eq. 1 becomes a special case of Eq. 4 with $m = 0.5$. For a inverter chain, factor m falls in between (0.5, 0.6). When $m = 1$, the variations are fully correlated and the result is the same as STA worst case. When $m = 0.5$, stage delays are independent variables. Typically, m falls in between (0.5, 0.7).

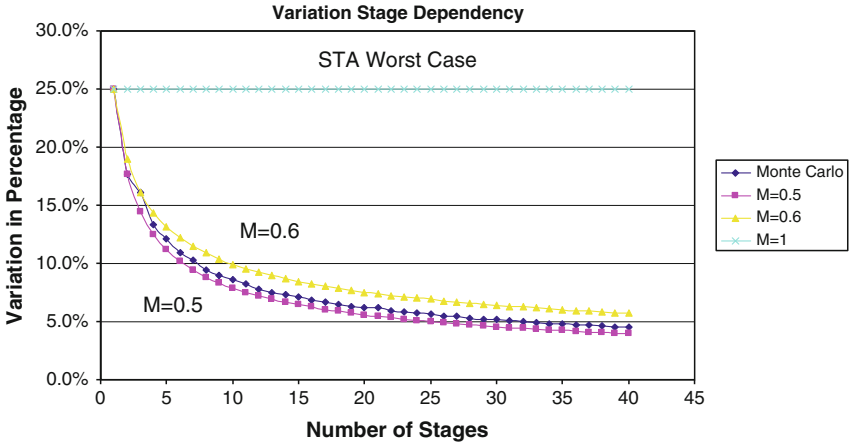


Fig. 3 Stage dependency of delay variation

2.2 *Delay Variation Voltage Dependency*

Figure 4 shows the delay trends of a typical inverter with respect to Vdd. The curve with circle marks is the inverter delay simulated using a traditional fixed process corner. The curve with triangle marks and the curve with square marks represent the 3-sigma timing variations in the Monte Carlo simulations. The variability of nanometer low power (LP) CMOS technology is modeled in the Monte Carlo SPICE model. When voltage is in the typical range around 1.1 V, the range of the timing variation is very small. As Vdd decreases, the timing variation increases dramatically. This means that at a lower voltage, two identical circuit paths on the same chip can have a large delay difference.

Figure 5 shows the voltage sensitivity of the mean delay and the delay variation. It clearly shows the trend of increased sensitivity at lower voltages. Both the delay and the delay variation are voltage dependent. Additionally, the delay variation (difference between the maximum and the minimum delay in Fig. 4) is more sensitive to voltage than the mean delay itself. The delay variation sensitivity is strongly nonlinear. This large delay variation plays an important role in the new failure mechanism for low-power designs.

2.3 *Delay Variation Transistor Size Dependency*

We use the minimum channel length to study the delay variation of an inverter. Figure 6 shows the percentage of delay variation induced by process variability versus transistor width with the minimum channel length. As the transistor width decreases toward the minimum width, percentage wise the delay variation increases sharply.

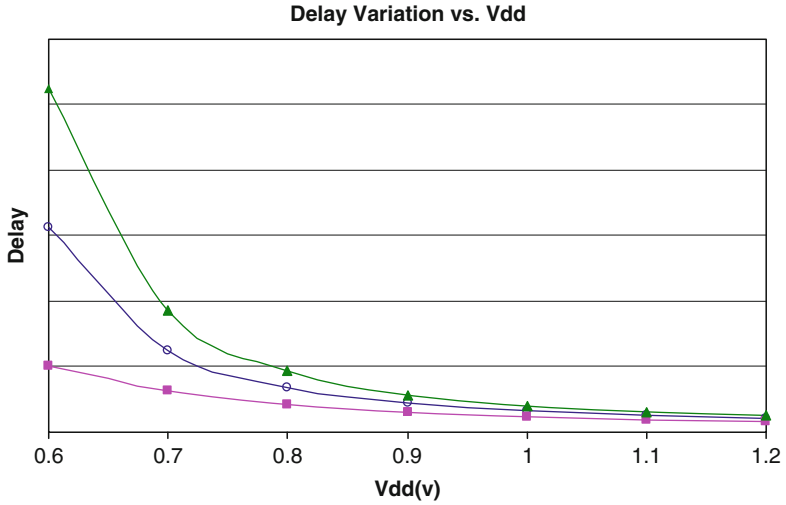


Fig. 4 Timing variation versus supply voltage

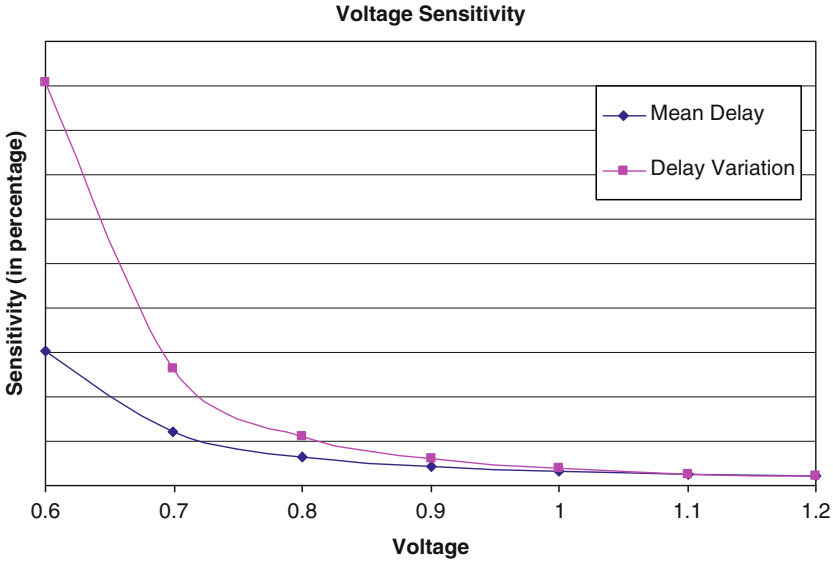


Fig. 5 Voltage sensitivity of delay and delay variation

This increase in variability becomes even more pronounced for circuits manufactured using small feature size. To reduce both the dynamic and leakage power, low Vdd and small transistors are highly preferred in mobile devices. In today's digital designs, the most frequently used transistor sizes are near minimum transistor widths. Therefore, the variability of smaller transistors presents a big challenge for low-power digital designs.

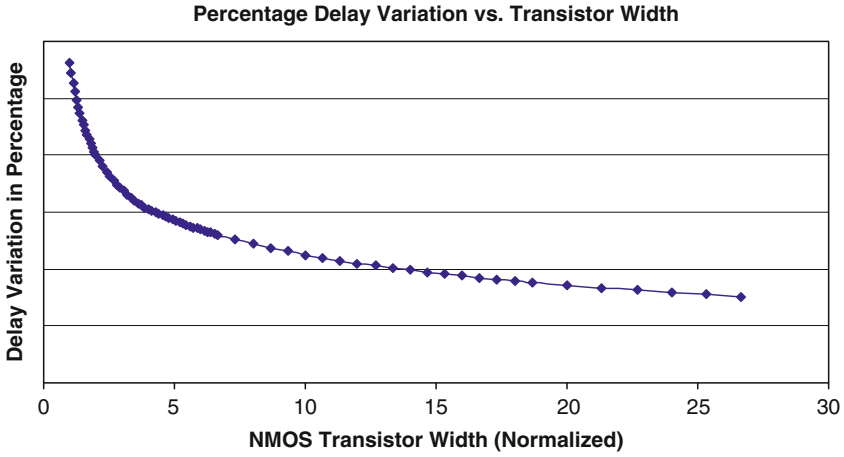


Fig. 6 Timing variation versus transistor width

2.4 Delay Variation Stage Dependency

Process variability causes delay variation in digital circuits. For two identical logic paths, path delays can be different from each other. For each stage, the delay variation exists. Along logic path, the total path delay variation increases with the number of logic stages increase. However, because the variability is independent, some stages become faster than nominal delay, some stages become slower. Percentage-wise, the delay variation tends to average out along logic path. Figure 7 shows the trend of percentage-wise delay variation for a simple inverter chain. The x axis is the number of stages and the y axis is the relative delay variation. As shown in Fig. 7, with number of logic stages increase, the relative variation reduces very quickly even though the absolute value of path delay increases.

It is interesting that delay variation transistor size dependency (Fig. 6) and stage dependency (Fig. 7) follows the same trend. The percentage-wise delay variation averages out due to correlated nature of local random process variability.

2.5 Timing Skew Due to Process Variability

We could study the absolute delay variation introduced by process variability. However, the absolute delay variation is not an effective criterion of the timing skew problem. For example, when Vdd is low, the delay variation will be large. At the same time, the circuit response time also becomes slow. Therefore, solely examining the absolute delay variation is inadequate; the relative timing skew between paths must also be checked.

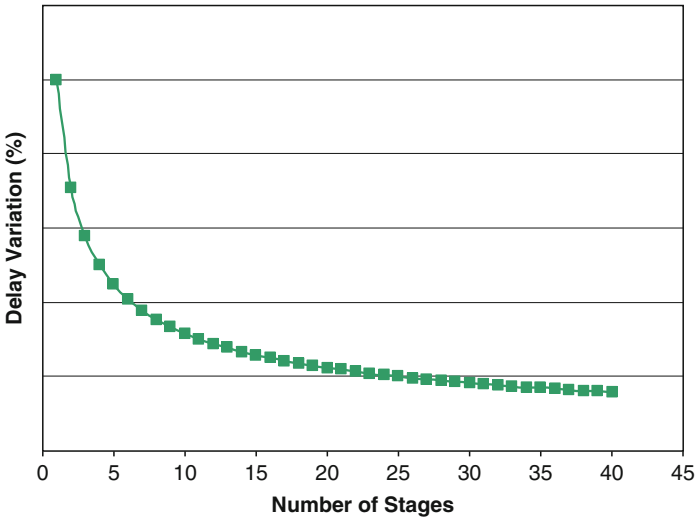


Fig. 7 Delay variation stage dependency

Traditionally, the rule of thumb for race condition verification is to check the logic depths of two timing paths. The path with more logic stages should have a later arrival time. However, with the increasing variability, is this assumption still valid? To quantify the impact of the variability on relative timing skew, we study the timing skew between the two identical timing paths. Here we use two identical inverter chains. As the number of stages increases, the absolute value of delay variation increases. The delay of stage i of chain 1 is compared to the delay of stage $I + 1$ of chain 2. We look for the number of stages where the path delay of stage i in chain 1 equal to or larger than the path delay of stage $i + 1$ in chain 2.

This simple check enables us to find out at which stage the accumulated delay variation will eclipse a single-level inverter delay. The results are captured in the delay overlapping stage (DOS) chart. Figure 8 shows the DOS versus voltage. The overlapping stage curves are not very sensitive to the process corners. Also note that the DOS chart is not sensitive to wire load. This is because when the load increases, both the stage delay and the delay variation increase. The DOS is very sensitive to voltage. According to the DOS chart, at 1.1 V for the frequently used inverter 1X, a delay overlap occurs at stage 5. Statistically, this means that for a logic depth of 5 and less, the variability-induced delay will be less than a single-level logic delay. In other words, the old rule of thumb for race conditions still applies. However, for a logic depth of 6 and larger, the delay variation will be larger than a single level-logic delay. The rule of thumb is no longer valid.

At 0.8 V, the overlap occurs at stage 2, which indicates the severity of the problem and shows how sensitive the timing variation is to the supply voltage. Note that in Fig. 6, the mismatch in wire load is not modeled. If wire load mismatches occur between the two delay chains, the situation becomes worse.

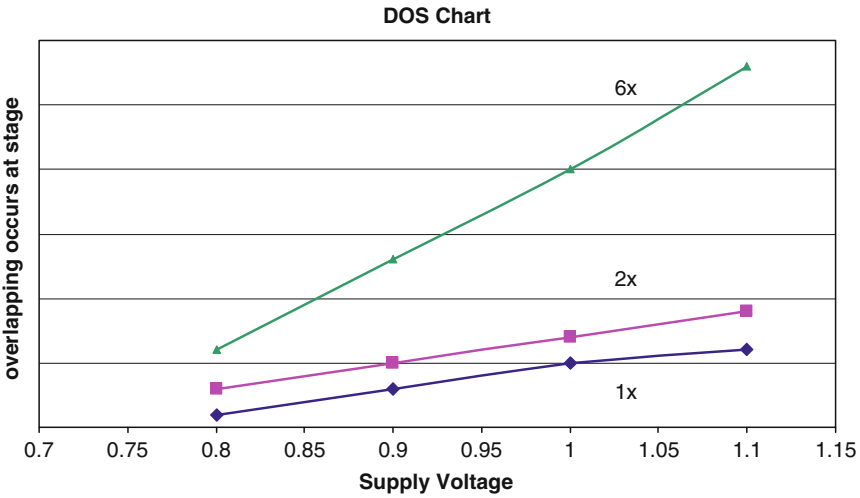


Fig. 8 Delay overlap stage (DOS) chart

The DOS chart also shows that the timing variation is very sensitive to transistor size. According to the DOS chart, at 1.1 V for inverter 6X, the delay overlap occurs at stage 20. We can easily minimize the process variability effects by using larger transistor sizes. For a given technology, the DOS chart is an important design criterion for assessing the impact of the process variability.

3 New Timing Failure at Low Voltage and Its Characteristics

Timing skews introduced by process variability can cause various circuit failures [1, 8]. Setup and hold time requirements can be violated, and pulse width can diminish. In traditional SPICE corner simulations, all transistors are fast or slow concurrently. In real silicon, the variability causes same circuit to have different timing, even when the identical designs are adjacent to each other on the same chip. Once the timing skew exceeds design margin, timing failures will occur. One common problem is hold-time failure. It is well-known that hold-time failures cannot be easily fixed. Therefore, it is very important to understand how the variability-induced timing skews affect the circuit functionality.

The variability causes statistical timing failures in self-timing circuits. In digital design, many circuits, such as memories, register-file arrays and pulsed-latch arrays, pulse generator circuits are used. Therefore, we study a pulse generator as an example. A pulse generator circuit is shown in Fig. 9. For simplicity, we study the minimum pulse width problem.

The timing margin for the pulse generator is defined as the pulse width minus slew. When the timing margin is zero, the pulse waveform becomes triangular in

shape. When the timing margin is larger, the pulse waveform becomes more rectangular. Figure 10 shows the timing margin at different voltages and traditional fixed process corners. The vertical axis is timing margin, and the horizontal axis is voltage. It is clear that the fast, fast (*ff*) process corner is the worst-case process corner. The reason is simple, when transistor speed is faster, the data arrives earlier. At a given supply voltage, the slow, slow (*ss*) corner has a better margin then the *ff* corner. As voltage decreases, the mean timing margins for both corners improve.

Next, we will look at the worst-case, 3-sigma timing margin in Monte Carlo simulation. The 3-sigma timing margin is quite different from the traditional timing margin. As the voltage decreases, the 3-sigma margin decreases, as shown in Fig. 11. Moreover, there is a cross-over voltage where the margin of the *ss* corner with the local mismatch becomes worse than the margin of the mean *ff* corner at high voltage. Simply speaking, as the voltage decreases, the variability effects become dominant, and an additional worst-case condition, *statistical worst-case*, occurs for the timing margin. Note that the timing margin is very sensitive to

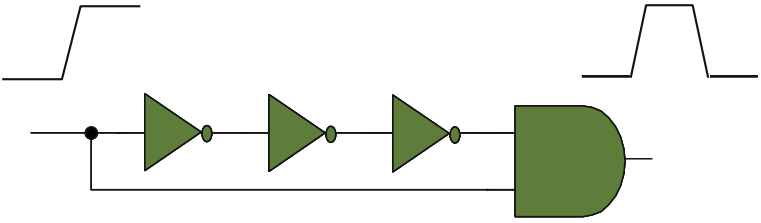


Fig. 9 Pulse generator circuit

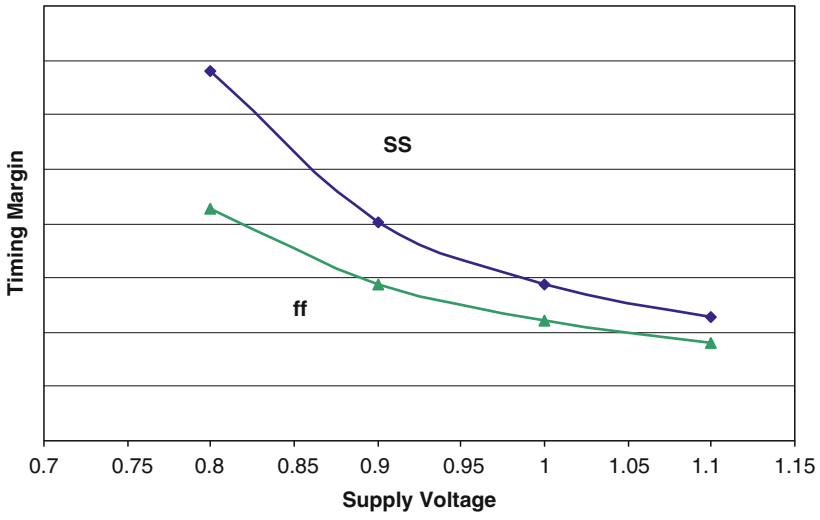


Fig. 10 Timing margin (fixed corners) versus supply voltage

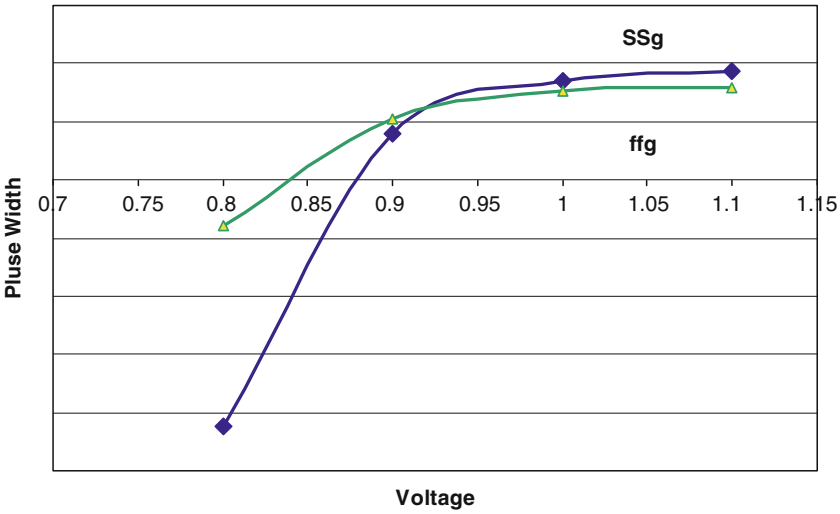


Fig. 11 3-sigma worst case timing margin versus supply voltage

voltage when the supply voltage is low. This sensitivity is not linear but more quadratic like.

The two worst-case conditions represent two entirely different failure modes. The high Vdd and *ff* corner timing failure is a majority circuit failure, while the low Vdd and *ss* corner failure is a statistical failure. When voltage is low, the majority of the circuits will work fine, except that the 3-sigma distribution of transistors is a problem. If low Vdd design robustness is not fully considered, yield may suffer. This failure is often overlooked by designers and unfortunately the burden is usually placed on process control engineers.

What is the worst-case process, voltage and temperature (PVT) condition for variability-induced timing failures? We now know that the worst-case voltage is low. The worst-case process corner is design-dependent. We don't know the answer to the worst case temperature for the statistical timing margin. There are two reasons. First, in low-power technology, a temperature reversal effect [5] may occur. Second, the temperature effect is not modeled accurately in the Monte Carlo SPICE models available.

Variability-induced statistical failures cannot be detected using fixed corners. Fixed corners assume that all transistors are faster or slower simultaneously, but failures occur only when one path becomes faster and the other path becomes slower. To detect these low-Vdd statistical failures, a statistical analysis technique such as Monte Carlo or statistical static timing analysis (SSTA) is needed. Ideally, a one-to-one correspondence exists between the traditional fixed corner margin and the variability-induced statistical margin. So, we need to meet only the fixed corner margin to cover the statistical failure margin. Unfortunately, the statistical failure margin varies greatly with different circuits, transistors sizes, etc. Therefore, designers must verify these margins separately.

To improve the timing margin, we increased the transistor size. Figure 12 shows the improved 3-sigma timing margin. The cross-over voltage decreases from 0.93v to 0.83 V. This cross-over voltage is named as *statistical failure voltage* (V_{sf}). When the V_{sf} is lower than the operating Vdd, the worst-case margin still occurs at the traditional worst-case condition. When the V_{sf} is higher than the V_{dd}, statistical failure timing margin is worse than the traditional worst-case margin. In this case, these two worst-case margins must both be verified. The V_{sf} depends mainly on the following factors:

Transistor threshold V_t
Circuit topology
Transistor size

Figures 10 and 11 shows difference voltage characteristics, including different cross-over voltages. The voltage characteristics of a given design determine the lowest V_{dd} level for voltage scaling. Even though in this work we used a simple pulse generator as example, the characteristics of statistical timing failure are basically ubiquitous. For any design with race conditions, a given V_{sf} exists. When the V_{sf} is higher than the specified lowest operating voltage, the effects of process variability begin to dominate and a statistical worst-case margin must be verified. Therefore, it is very important for designers to know the V_{sf} for a given circuit.

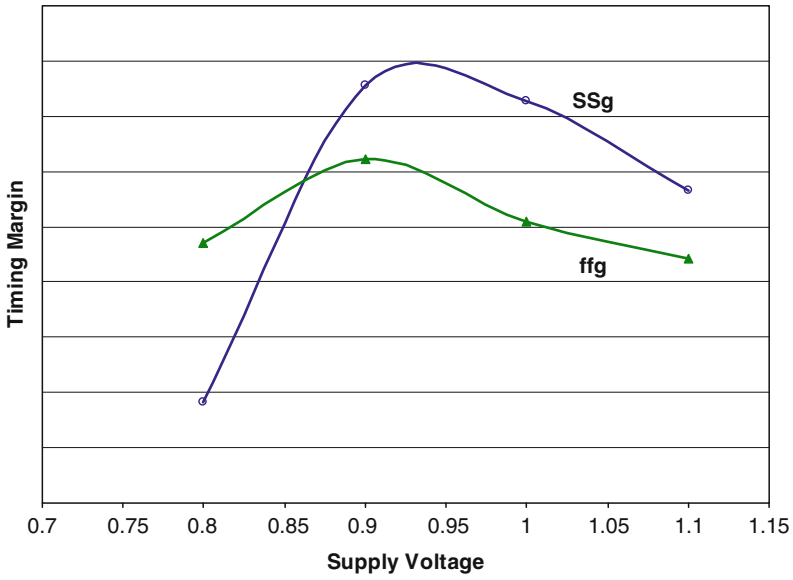


Fig. 12 Improved timing margin versus supply voltage curves

4 Variability-Resilient Design Methodology

We now have a better understanding of the variability impact on timing variation. Due to the increasing variability in 65 nm technologies and beyond, circuits may suffer yield loss. To improve yield, designers must verify the circuit timing margin not only under the traditional worst-case condition but also under the statistical worst case condition. The statistical worst-case condition is frequently different from the traditional worst-case condition.

A statistical failure-resilient design and verification methodology for low-power design includes the following steps:

1. Generating a DOS chart for a given technology, and checking the design against the chart to determine if the process variability effects are a potential problem.
2. Verifying the design margin in a Monte Carlo simulator and finding the Vsf.
3. Comparing the V_{sf} with the lowest operating voltage.
4. If the V_{sf} is higher than the lowest V_{dd}, investigating whether the process variability sensitivity can be improved by changing the circuit topology and transistor sizing. If not, then verifying the design in Monte Carlo simulations under the statistical worst case condition.

To provide an adequate margin for a design, designers must verify it according to the specification and also know the following:

1. The actual worst-case design parameter values
2. The sensitivities of the design margin with respect to the critical design parameters

When the margin is sensitive to the supply voltage, the designer must guarantee that the design margin is adequate at the real worst-case voltage. For example, even though the lowest supply voltage may be designed at a certain level, on-chip DC drop and power/ground voltage fluctuation may exist. For the best design margin, we must consider the actual worst-case voltage. This way the design will remain truly robust during volume production, resulting in a higher yield.

5 Challenge in Low-Power Designs

Lowering the supply voltage is the best way to reduce power [6]. Nevertheless, lower voltages make circuits sensitive to process variability. To reduce dynamic and leakage power, small transistors should be used, but as we know, small transistors have large process variability. If power consumption and layout density are not concerns, using larger transistor sizes and higher supply voltages will eliminate the problem of process variability-induced timing failures. However, if we want to reduce power by using lower supply voltages, reduce die size by using smaller transistors, and reduce leakage by using high V_t transistors, we must thoroughly consider the effects of process variability. Otherwise, major yield

problems will occur. For low-power CMOS designs, process variability is the major challenge for circuits using nanometer manufacturing technology.

For a given technology and circuit, a watershed-like supply voltage level exists. Below this level, the circuit becomes sensitive to process variability. Once the circuit is variability-sensitive, statistical circuit failures may occur. Many circuit techniques can no longer provide the design advantages they used to offer. For example, global variability can be successfully tracked by replica circuits. However, the intra-die random variability cannot be duplicated. The only way to counter the effects of the variability is by appropriately margining the design. When voltage scales down, the variability effects increase. As a result, the margin must increase. Therefore, voltage scaling can be very expensive. The advantage of using small transistors for low-leakage and low-power diminishes quickly if the variability is not given adequate and careful consideration.

The variability-induced timing failure is a statistical failure. It manifests itself as yield loss. Hence, a statistical analysis method must be used. The run time of Monte Carlo simulations is much longer than single-point SPICE simulations. Also, the effectiveness of Monte Carlo simulations decreases rapidly for large circuits. Therefore, the analysis must be performed using a divide and conquer approach, which significantly increases the complexity of the circuit design, verification, and optimization.

To ensure a high-yield design, it is crucial to choose an appropriate transistor threshold and circuit operation voltage. Most importantly, circuit designers must fully understand the impact of the variability on circuits and product yield. It is the designer's responsibility to ensure a high circuit yield in silicon. The variability will persist due to its inherent physical nature. It is also equally important that EDA companies develop tools to assist designers in combating the challenges presented by the variability.

6 Conclusion

Intro-die random process variability-induced timing failure is one of the biggest challenges in today's low-power CMOS designs. For the first time, this work examined the effects of the variability on circuit functionality, yield, and voltage scaling. A simple yet effective method for quantifying the effects of the variability on digital timing, the DOS chart, was introduced. The variability causes timing skews which in turn cause circuit failures and yield loss at low voltages. This new failure mode is statistical in nature and behaves differently from traditional timing failures. It has a different worst-case corner that requires further statistical analysis. We presented a variability-resilient design and verification methodology that deals with this new failure.

References

1. C. Visweswariah, Death, taxes and failing chips, in *Proceedings of Design Automation Conference*, June 2003, pp. 343–347
2. F.N. Najm, N. Menezes, Statistical timing analysis based on a timing yield model, in *Proceedings of the Design Automation Conference*, June 2004, pp. 460–465
3. D. Agarwal, V. Blaauw, S. Zolotov, S. Vruduhula, Computation and refinement of statistical bounds on circuit delay, in *Proceedings of the Design Automation Conference*, June 2003, pp. 348–353
4. P.S. Zuchowski, P.A. Habitz, J.D. Hayes, J.H. Oppold, Process and environmental variability impacts on ASIC timing, in *Proceedings of the International Conference on Computer-Aided Design*, Nov 2004, pp. 336–342
5. S. Borker, et. al., Parameter variability and impact on circuits and microarchitecture, in *Proceedings of the IEEE/ACM International Design Automation Conference*, June 2003, pp. 338–342
6. T.C. Chen, Where CMOS is going: trendy vs. real technology, *IEEE Solid State Circuits Society Newsletter*, **20**(3), Sept 2006
7. Y. Huang et al. Efficient diagnosis for multiple intermittent scan chain hole-time faults, in *Proceedings of the 12th Asian Test Symposium (ATS'03)*, 2003
8. Amitava Majumdar et. al., Hold-time validation on silicon and the relevance of hazards in timing analysis, in *Proceedings of the Design Automation Conference*, 2006, pp. 326–339
9. Enrico Malavasi, et al. Impact analysis of process variability on clock skew, in *Proceedings of the International Symposium on Quality Electronic Design*, 2002

How Does Inverse Temperature Dependence Affect Timing Sign-Off

Sean H. Wu, Alexander Tetelbaum, and Li-C. Wang

1 Introduction

As processing technology migrates into sub-90 nm region, design performance can be affected by factors that were considered secondary before. One of such factors is the Inversed Temperature Dependence (ITD) effect [1]. When a circuit is operating in low voltage, the propagation delay of a cell may decrease as the temperature increases [4]. The reason behind ITD effect is due to the temperature effect on the threshold voltage, V_{TH}. As supply voltage (V_{DD}) scaled, the value of $|V_{GS} - V_{TH}|$, the absolute difference between transistor gate to source voltage and threshold voltage, decreases. The smaller $|V_{GS} - V_{TH}|$ makes saturation current more sensitive to change in V_{TH}, which decreases as the increase of temperature. The smaller V_{TH} incurs more current that makes the device switching faster. On the other hand, transition delay is also proportional to the electron mobility, which decreases as the temperature rises. Hence the device performance depends on the racing condition of electron mobility and V_{TH} together as temperature rises. Traditionally, timing is signed off at two extreme temperature corners, one representing the best case and the other representing the worst case. With ITD, the highest sign-off temperature can no longer guarantee the worst case, and vice versa. This poses a serious problem to the timing sign-off methodology, i.e. it is possible that the worst-case temperature occurs at some intermediate point and finding this point can be quite difficult. Due to the goal of having low power design, modern designs are implemented by

S.H. Wu (✉)
LSI Corporation, USA
Department of ECE, University of California, Santa Barbara, USA
e-mail: sean.hsi.wu@gmail.com

A. Tetelbaum,
LSI Corporation, USA

L. C. Wang
University of California, Santa Barbara, USA

A. Amara et al. (eds.), *Emerging Technologies and Circuits*,
Lecture Notes in Electrical Engineering 66, DOI 10.1007/978-90-481-9379-0_13,

standard cells with high V_{TH} extensively. Coupling with the ITD effect, it is necessary to understand the impact on sign-off methodology.

The work done by Park et al was one of the earliest research indicate temperature inversion effect in low voltage device [2]. Kumar and Kursun showed how temperature increases can speed up cell transition at the worst case [1]. Although increasing VDD can make a design less sensitive to temperature variation, it may not be a feasible solution since the power overhead is undesirable. Gerousis's study discussed design and modeling challenge in 90 nm and beyond [3]. In order to account for the possibility that worst case delay could occur at either the highest or lowest temperature, timing-sign off on four corners is suggested. The work by Lasbouygues et al discussed how to use non-linear factor to account for performance impact from temperature and voltage variation [4]. Dasdan and Hom discussed how the fastest and slowest delay of a path can be bounded with the presence of inverse temperature dependence [5]. However, in order to bound the path delay, every cell needs to be re-characterized by sweep through every voltage and temperature to find the extreme cases of delay. To eliminate design performance's dependence on temperature, couple works had studied and suggested setting optimum supply voltage where circuit performance does not vary with temperature [6, 7]. However, this is not achievable with aggressive voltage scaling. Testing for temperature sensitive failure was presented by Long [8]. None of previous works discuss the relationship between change of delay and temperature can be non-monotonic. In this work, we will present how ITD interact with the change in manufacturing process, supply voltage and temperature. The rest of the paper is organized as follows. In Section 2, the experiment and result will be discussed. Section 3 discusses the current practice on dealing ITD. We will conclude the paper and talk about future work in Section 4.

2 Experimental Study

For each cell, seven different output loading capacitance and seven different input slews are simulated with several Process/Voltage/Temperature, known as PVT corner libraries. In each three different PV corners, best process/voltage, nominal process/voltage, and worst process/voltage are used. Two different temperature settings are in our interest, -40° and 125°. Since the ITD is closely tight to the value of V_{TH}, cells made by different V_{TH} are also studied. We use the typical TSMC 65 nm device model file. The output loading and input ramp time settings are same as the ones found in the look-up table in the cell libraries. There are more than 8,900 cell, load, and input slew combinations simulated for each library and PV corner combination. The first thing we look at how likely ITD occurs across the cells.

Figure 1 shows how delay is impacted by temperature increases at the two sign-off temperatures at worst case process and voltage corner on standard V_{TH} device library. The negative value on the x-axis means the delay decrease percentage due

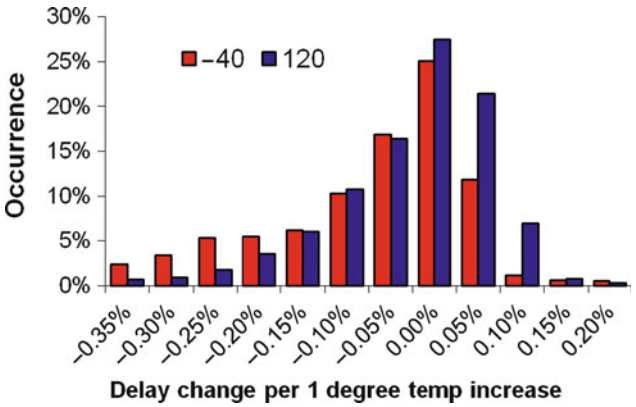


Fig. 1 ITD occurrence

Table 1 The possibility of ITD occurs

	High V_{TH} (%)	Standard V_{TH} (%)	Low V_{TH} (%)
Worst corner	79.03	78.31	47.16
Best corner	27.01	30.81	9.19

to the increase of temperature. More than 50% of the cells become faster as the temperature increases from the two temperatures corners. In Table 1, three different threshold voltage libraries were studied at the two corners, the best PV and worst PV corners. First, the occurrence of ITD is more likely found in devices with higher threshold voltage. Moreover, the best corner has fewer cells affected by IDT in comparison to worst corner. The reason for both trends is due to the gap between supply voltage and threshold voltage, i.e. the best corner has a larger $|V_{GS} - V_{TH}|$ and therefore it is less affected by ITD. In general, ITD can be found on average about 22% in the best PV corner and 65% in the worst PV corner.

2.1 Temperature Versus Delay Relationship

Figure 2 shows a typical linear relationship between temperature and delay of a cell. This is what to expect normally when using the two extreme temperature corners for timing sign-off. The minimum and maximum delays are found at the corner temperatures. In contrast, Fig. 3 shows another example where the relationship is non-monotonic. The maximum delay of this cell occurs at -30° and is 11% worsen than the -40° corner delay. This illustrates the risk of current timing sign-off methodology as the sign-off corners might not adequately represent the worst case.

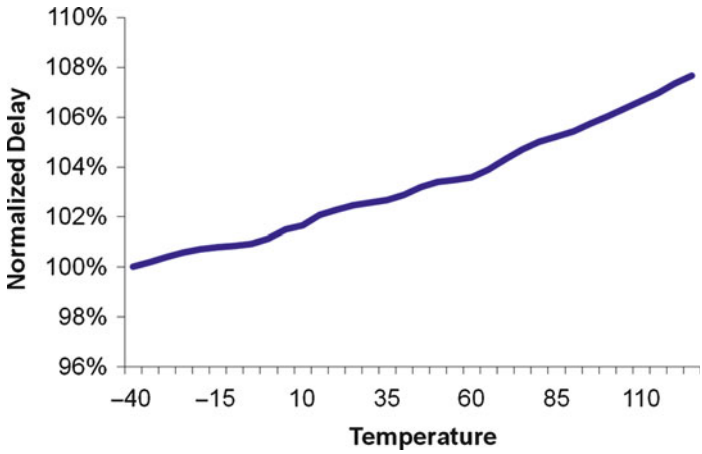


Fig. 2 Expected delay versus temperature relationship

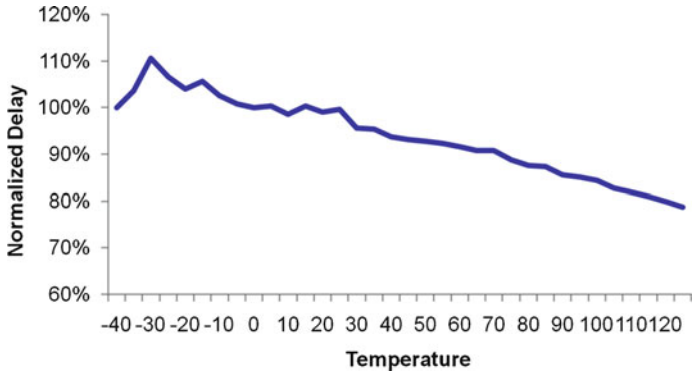


Fig. 3 Undesired non-monotonic relationship between delay versus temperature

2.2 Impact of V_{TH}

It is shown that threshold voltage has a direct impact on determining whether a cell is under ITD or not. It is interesting to see how the change of V_{TH} would impact the cell behavior. Therefore, a more detailed study is done on varying the V_{TH} and on a wider temperature range. The result is shown in Fig. 4. Again, the y-axis is the normalized cell delay, which is normalized to the delay. The x-axis is the temperature setting.

The high V_{TH} cell shows a very obvious concave upward behavior. As the threshold voltage is lowered, the curve is less bent. With a low threshold voltage, the cell delay almost does not show any ITD effect. On the other hand, if we take a

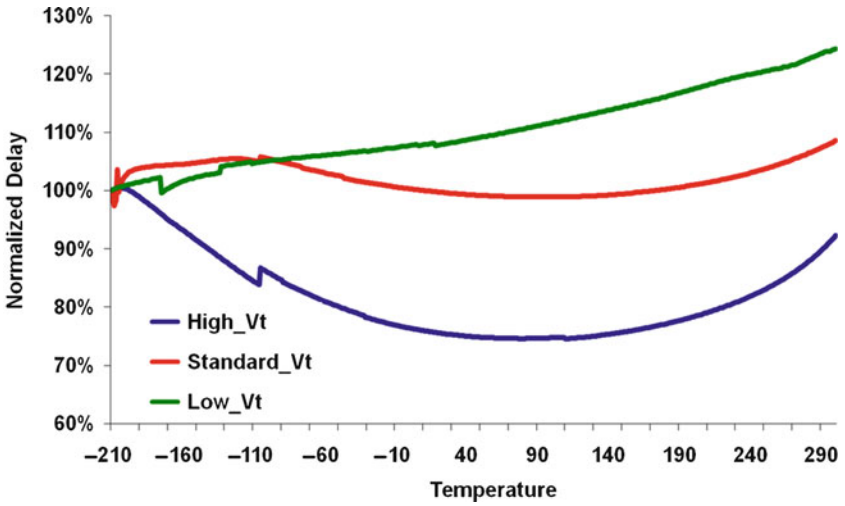


Fig. 4 Threshold voltage's impact on falling transition delay

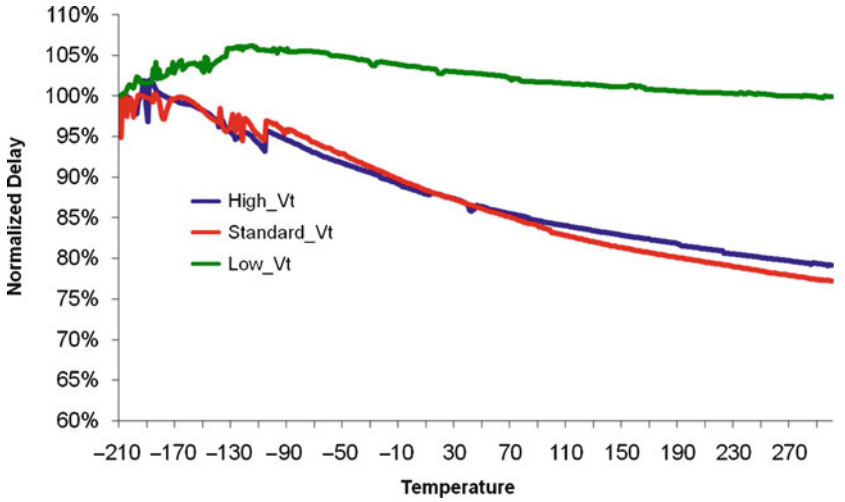


Fig. 5 Threshold voltage's impact on rising transition delay

look at a look at the opposite transition delay of the same cell; Fig. 5 shows a very different result. Although lower V_{TH} is less vulnerable to ITD, in this case all three curves clearly show ITD effect. It is obvious that low V_{TH} cells are more resistant against ITD, but the usage of low V_{TH} cells is limited because of the low power requirements posted on designs today.

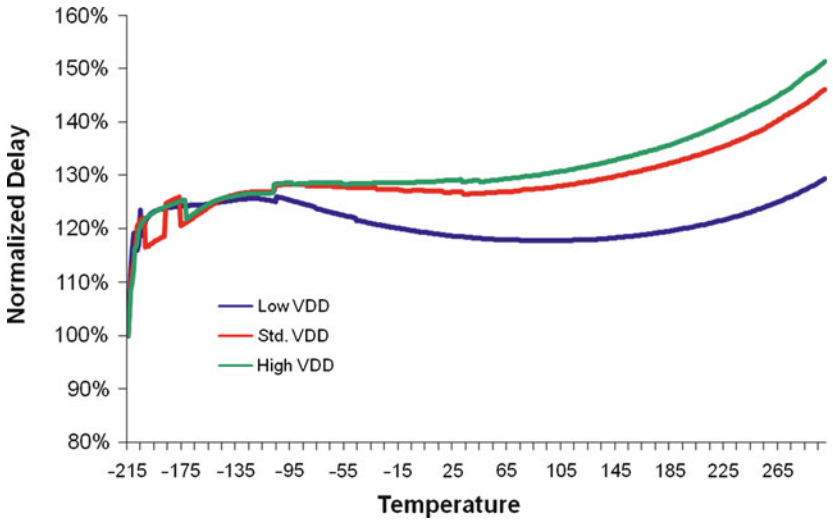


Fig. 6 Supply voltage's impact on transition delay

2.3 *Impact of Supply Voltage*

The level of supply voltage (V_{DD}) has a dramatic impact on ITD. This is observed again in Fig. 6. Three different supply voltages, low or worse case V_{DD}, standard or nominal case V_{DD}, and high or best case V_{DD}, are used to examine the effect of ITD. The ITD is evident in the middle section of the low V_{DD} curve and become less visible in the high V_{DD} curse. Using higher V_{DD} is more desirable to avoid ITD and non-monotonic delay arcs. On the other hand, low V_{DD} would cause the performance to be less predictable versus the change of temperature.

It becomes a tradeoff between the desire to achieve low power and the desire to guarantee worst-case timing in timing sign-off, the tradeoff remains unclear unless we can model ITD effect in detail. As expected, a higher V_{DD} is desired to over the non-monotonic effect. Unfortunately, with the desire of low power design, it is more likely for us to see delay arcs that are more non-monotonic.

2.4 *Impact of Process Variation*

ITD can also be process-variation dependent. For example, Fig. 7, show the result based on a two-input NAND with V_{DD} is set at the low value and using standard V_{TH} MOSFET device is used on the *slow*, *typical*, and *fast* corners. The ITD effect at the slow process corner is more significant. At the fast corner, the ITD effect almost disappears. Due to the interaction between process variation and

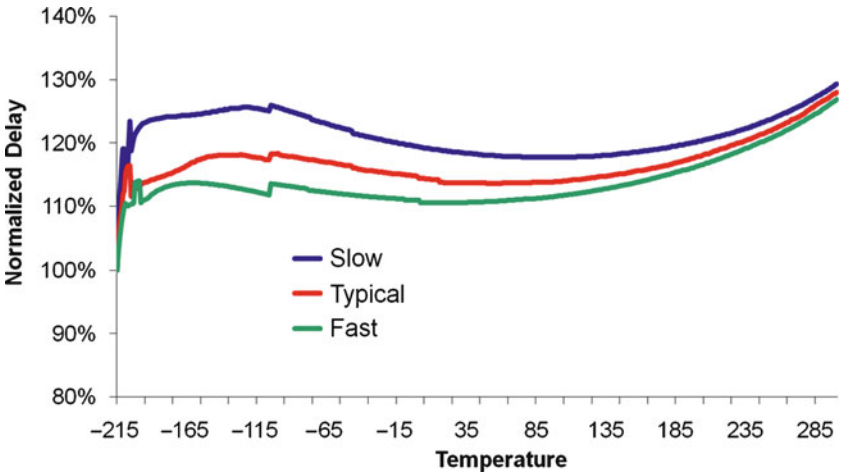


Fig. 7 Process variation's impact on transition delay

ITD, some fabricated samples would fail during testing. This does not affect sign-off methodology as long as the methodology ensures no path would violate timing constraints systematically due to ITD. Failures due to random process variation should be solved in testing or process control.

2.5 Random Cells and Loading

Figure 8 illustrates the effect of ITD on different cells with different amount of output loading. The y-axis indicates the amount of delay exceeding maximum delay obtaining from 120° to −40° corner. Seven different cells and load combinations are presented. It is noticeable that maximum delay occurs at temperatures other than the corner temperatures. The amounts of delay exceeding corner temperature are all less than 1.5%, but occurring at different temperatures. Implying to path delay, it becomes difficult to determine at which temperature the maximum delay would occur because individual cells may reach peak delay at different temperatures. On the other hand, ITD might not affect path delay as much as individual cell delay. We will try to illustrate the ITD impact on path in the next section.

2.6 ITD Impact on Path

As discussed the in the previous section, ITD could have significant impact on individual cells. The next question we are interested is what ITD's impact on a path is. Therefore, a simple path made by the inverter cell that is impacted by ITD the

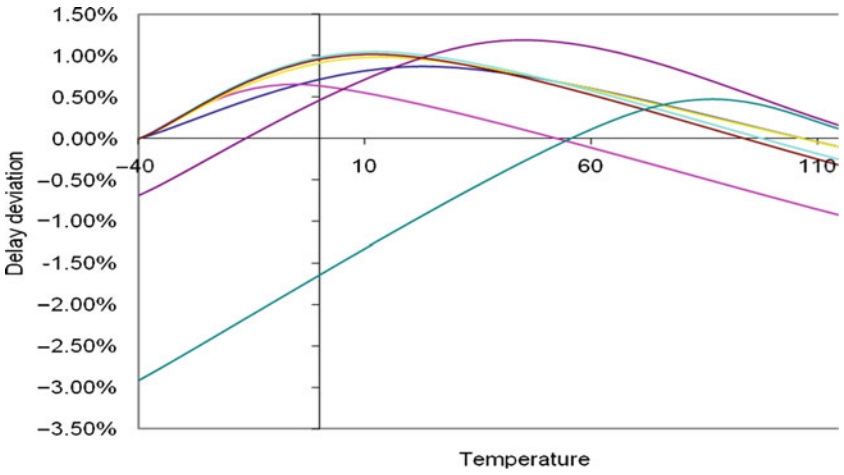


Fig. 8 Delay versus temperature on seven different cells and out loading combinations

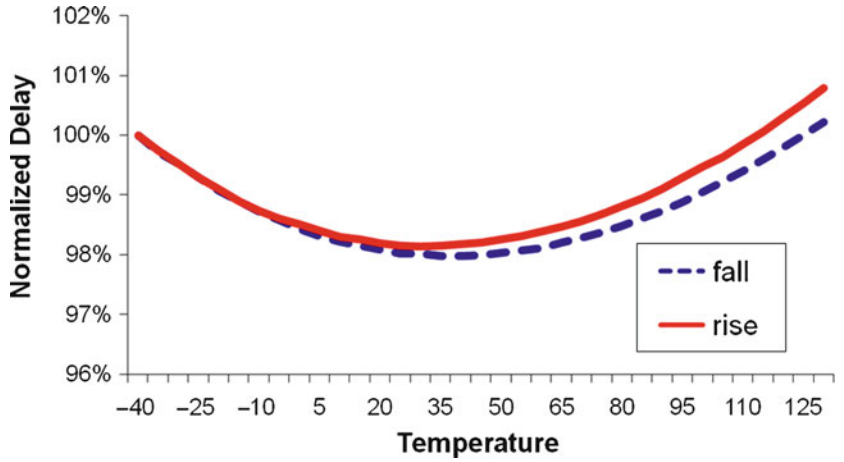


Fig. 9 Performance versus temperature on one single path

most is put under study. In Fig. 9, it shows the delay of rising and falling transition on different temperature of a ring oscillator is made by the cell that ITD affects the most. Although ITD's impact on individual cell's delay is larger than 10% as showed earlier, the minimum path delay is only 2% faster than the cold temperature corner. The result implies that due to the change in input signal ramp time and output loading when cells are put into paths, ITD characteristic changes. And the impact on the path is not necessarily equal to the impact seen on the cells. On the other hand, it also shows the complexity of modeling ITD in detail is high due to its interaction with many different factors.

3 Potential Solution

In order to guard against the non-monotonic behavior in cell delay with respect to temperature variation, several solutions are possible. The most naive solution is to do timing sign-off at multiple intermediate temperatures. This is usually not acceptable because it may dramatically increase the cost of timing sign-off. An alternative solution is to add an extra margin to guardband the ITD effect and continue to use the current sign-off flow. Adding margin is usually the last resort because it increases the design cost. Moreover, deciding a reliable margin for ITD is still challenging. The margin shall not be determined by individual cell because it might be over pessimistic; arbitrary amount of margin cannot be justified.

One method used by designers is to *swap* suspicious cells that are likely to be impacted ITD with cells with lower threshold voltage devices. By swapping, the ITD can be compensated as shown in Fig. 10, in which the non-monotonic delay arc disappears completely. The advantage of the method is that it requires no alternation in layout and routing since the cell footprint remains the same. The tradeoff of using low V_{TH} cells is higher leakage. In addition, the swapping can be arbitrary without any guarantee on avoiding the ITD effect. And this practice is only applied only when the path timing violates the constraint at low temperature corner but passes in the high temperature corner. By swapping with low V_{TH} cells, it improves the ITD impact at the low temperature corner and also speeds up at the high temperature corner. So the path would satisfy the timing constraints.

However, in the typical sign-off flow, the timing of a path is only estimated at the cold and hot temperature corners. By only looking at the result at $-40°$ and $125°$ in Fig. 9, it would be assumed that this is a typical path since it operates faster at the cold temperature and slower at the hot temperature corner. A timing analysis done at an intermediate temperature is required to identify the path has a non-monotonic relationship between timing and temperature. Of course, the more timing analysis at different temperatures would increase the observability of such paths. Although it is

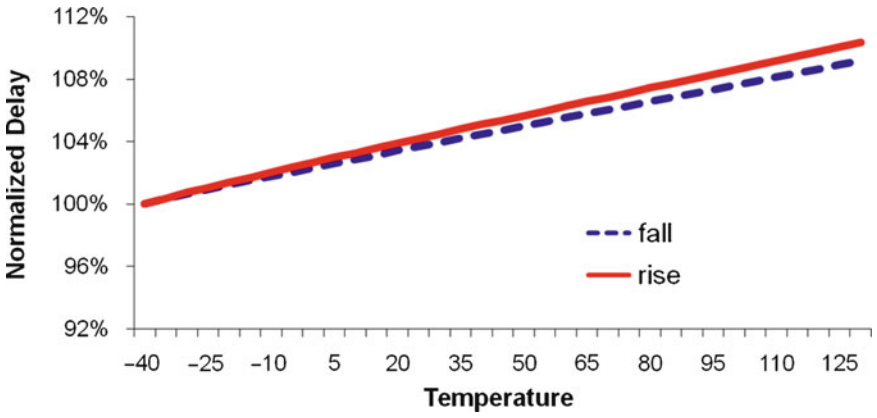


Fig. 10 Performance versus temperature on one single path after cell swapping

not feasible to do timing analysis at multiple temperatures of full design due to the cost, such practice can be limited to the selected critical paths. Our experiment result also indicates that paths that consist of large amount of high V_{TH} cells are more likely affected by ITD. This provides a general guideline on picking paths for further analysis. As of in typical ASIC design, the number of potential timing violated paths are generally a lot less than the total amount of paths.

If one tries to model ITD effect in detail, this can mean a significant increase in cost for library characterization. As shown above, the ITD effect can depend on many things and hence, the complexity to model the ITD effect of a cell can be high. If one tries to incorporate such a model in a static timing analyzer (STA), this may mean an increase in run time. Unless ITD effect is proved to significantly impact timing yield, adding cost to the current characterization and STA is usually not acceptable. Adding margin might not be an acceptable option. Hence this leaves a design team with the dilemma that on one hand, the existence of ITD effect is known and on the other hand, its impact to timing yield is hard to measure. Without a way to quantify the impact, it becomes difficult to justify having a new sign-off flow.

4 Conclusion

In this work, we demonstrate that *inverse temperature dependence* is a complicated issue as it depends on many different factors such as manufacturing process, supply voltage, type of cells, etc. It makes ITD very difficult to be modeled in the tools. The relationship between temperature and path timing might be non-monotonic due to ITD, which threaten the effectiveness of the timing sign-off flow. Adding margin is not feasible as it makes design process more difficult. In order to detect the critical paths affected by ITD in pre-silicon stage, timing analysis at additional temperatures is required on selected paths. Once such affected paths are identified, *swapping* is applied on suspicious cells that are affected by ITD to low V_{TH} devices.

One can think that ITD is just one of the many un-modeled timing effects based on a pre-silicon sign-off flow. Un-modeled effects can be due to (1) their impact to timing yield is unclear or (2) it is too costly to model and simulate them. The ITD may not need to be fixed in the pre-silicon stage. The future work is in direction of using post-silicon information as an extension to the pre-silicon timing sign-off flow and discuss how to feed back silicon information to determine where to fix a design and where to improve a model.

References

1. R. Kumar, V. Kursun, Reversed temperature-dependent propagation delay characteristics in nanometer CMOS circuits, *IEEE Transaction on Circuit and Systems* **53**(10) Oct 2006
2. C. Park et al., Reversal of temperature dependence of integrated circuits operating at very low voltages, *IEDM Conference*, 1995

3. V. Gerousis, Design and modeling challenges for 90 nm and 50 nm, *Custom Integrated Circuits Conference*, 2003
4. B. Lasbouygues et al., Temperature- and voltage-aware timing analysis. IEEE Trans. Comp-Aided Design Integr. Circuits Syst. **26**(4) (Apr 2007)
5. A. Dasdan, I. Hom, Handling inverted temperature dependence in static timing analysis. ACM Trans. Design Automat. Electronic Syst. **11**(2) (Apr 2006)
6. A. Bellaouar, A.F.M.I. Elmasry, K. Itoh, Supply voltage scaling for temperature insensitivity CMOS circuit operation. IEEE Trans. Circuit Syst II **45**(3), 415–417 (March 1998)
7. K. Kanda, K. Nose, H. Kawaguchi, T. Sakurai, Design impact of positive temperature dependence of drain current in Sub 1V CMOS VLSI's. IEEE JSSC **36**(10), 1559–1564 (2001)
8. E. Long et al., Detection of temperature sensitive defects using ZTC, *IEEE VLSI Test Symposium*, 2004, pp. 185–192

CMOS Logic Gates Leakage Modeling Under Statistical Process Variations

Carmelo D'Agostino, Philippe Flatresse, Edith Beigne, and Marc Belleville

1 Introduction

In the history of semiconductor, transistor scaling has been used as a primary method to improve IC performance. However in the nanometer regime, aggressive scaling is reaching its limits and the control of semiconductor manufacturing process is becoming increasingly difficult. Variations in manufacturing process have grown, and variations in device parameters have grown even more, resulting in wider distributions which, in turn, could result in yield loss [1]. Another scaling consequence has been a drastic increase of leakage currents. Leakage has become a major contributor to the total IC power, reducing battery life during stand-by operations in portable applications. Furthermore, this is worsen by the very large impact of variations on device leakage. In the nanometer regime the four major contributors to transistors leakage are: the subthreshold leakage (Ids), the gate leakage (Igd), the reverse-biased drain and source substrate junction band to band tunneling (Ibtbt), and the gate induced drain leakage (Igidl). Each of those leakage currents becomes significant in nano-scaled devices tightening the constraints of nowadays digital designs [2].

Today, fitting within the power budget is as important for designers as achieving maximum performance: instead of targeting only the absolute maximum performance, designers need to maximize the performance considering the given power budget [3, 4]. Furthermore, since the leakage currents in a device depend mainly on the transistor geometry and the threshold voltage, statistical variation of those parameters leads to a significant spread of the total leakage [5].

C. D'Agostino (✉) and P. Flatresse
STMicroelectronics Crolles, FTM/DAIS, France
e-mail: carmelo.dagostino@st.com

E. Beigne, and M. Belleville
CEA-LETI Grenoble, MINATEC, France

A. Amara et al. (eds.), *Emerging Technologies and Circuits*,
Lecture Notes in Electrical Engineering 66, DOI 10.1007/978-90-481-9379-0_14,

Although Monte-Carlo analyses are accurate in estimating the leakage distributions, they considerably increase the simulation time and hence the design cycle. Therefore, an analytical statistical estimation and modeling methodology of the total leakage current, considering the effects of process parameter variations, is needed for designing CMOS circuits in the nano-meter regime [6].

Numerous statistical analysis techniques have been proposed recently for estimating the Probability Density Function (PDF) of circuit delay and leakage power considering Process, Voltage and Temperature (P-V-T) variations [7–9]. Rao et al. [7] have proposed a statistical leakage analysis method for modeling the impact of gate length variations on subthreshold leakage current. However, this analysis uses an empiric relation between the current and the process parameter. Moreover the methodology cannot be easily extended to other process variations such as oxide thickness or dopants concentration since that would beforehand require an empirical and invertible relationship between the leakage and the considered fluctuating parameter. Dadgour et al. [8] presented a statistical framework to estimate statistical parameters of current leakage while considering variability in process, temperature and voltage. Nevertheless BTBT leakage current is neglected and the shape of the resulting statistical distribution is not determined.

Although the previously proposed techniques predict the circuit delay or leakage with a good accuracy, their reliability with respect to manufacturing strongly depends on the correctness of the empiric relations used for the gate delays and power. Moreover, most of past approaches approximate the leakage variation as log-normal shaped distribution. Such approximations can lead to inaccurate results in nano-scaled devices, especially because the leakage current has an extremely complex dependency on process variations. Li et al. [9] recently proposed an approach that does not rely on log-normal approximation, but it requires a preliminary standard cell library characterization (based on SPICE simulations) of cell leakage current as a function of parameter variations.

The goal of the proposed methodology is to obtain a time-efficient and accurate estimation of the PDF of the leakage current of a logic gate. Using this methodology, time consuming Monte-Carlo based simulations on high-complexity structures can be avoided. Figure 1 describes the three main phases of the current methodology implementation:

- Probability analytical relationships are included in standard transistor models (BSIM/PSP).
- Monte Carlo simulations are performed on a set of single transistors: the process parameters whose variations impact the leakage spread are identified. This step is performed only one time for each technology node. Since the simulations are performed on low-complexity devices, the computational time is very short.
- Parametric simulations are performed on the logic gates, sweeping, with a discrete step, the variation range of the identified process parameters. Accuracy of the result and computation time can be selected setting the right number of parametric simulations.

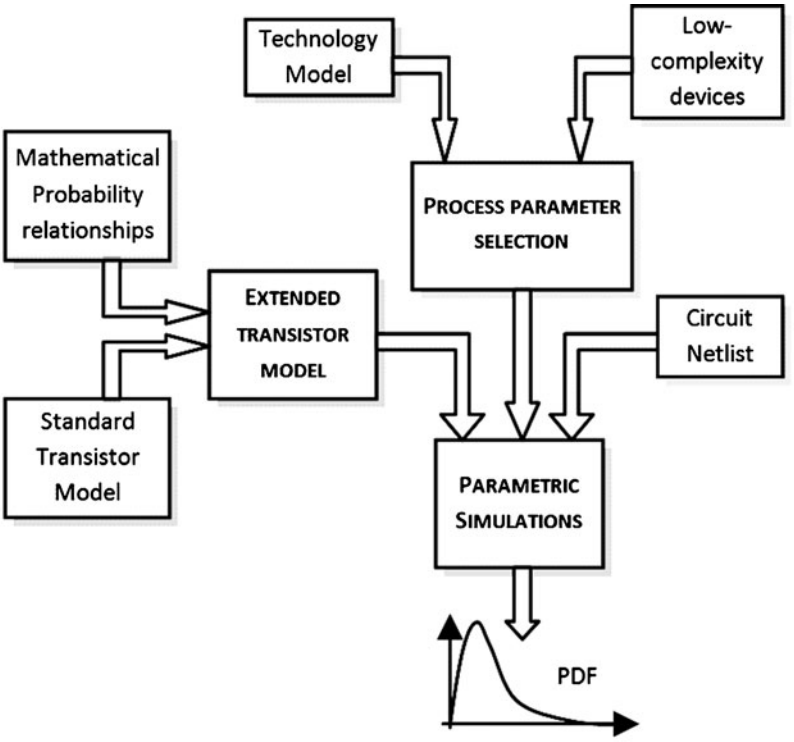


Fig. 1 Main steps of the proposed analytic leakage estimation methodology

2 Analytical Model of Current Leakage Under Process Variations

Previous works report that each transistor leakage component depends on more than one parameter and has different sensitivity with respect to those parameters variation [10]. The proposed general framework can be easily used to model the total leakage current with respect to different process parameters variations (L_{gate}, T_{ox}, V_{th0}, ...). Thus, the proposed methodology can be extended to estimate the overall impact of a number of different variation sources on the total leakage current of a logic gate. For clarity of purpose, in the following sections, the methodology is illustrated analyzing the variability of the leakage current with respect to the drawn channel length. To consider the impact of another process parameter on the variability of the leakage current, it is enough to substitute in the following formulas the gate length with the preferred parameter.

2.1 Mathematical Problem Formulation

Given the PDF of $L = f_x(L)$, the dependence between the leakage current and the gate length $I = h(L)$, and its inverse function $l = h^{-1}(L) = g(I)$, the PDF of I can be expressed using the expression in Eq. 1 [11]

$$PDF(I) = f_y(I) = \frac{f_x(g(I))}{h'(L)} \tag{1}$$

where $h'(L)$ is the first derivative of the function $h(L)$.

To compute the PDF of the leakage current, it is essential that: (a) the function g is a closed-form expression and (b) the function h is differentiable over the given range of currents [7]. Unfortunately, the complexity of the relationship between the leakage current and the channel length does not allow the derivation of $g(I)$ to satisfy those two conditions. Rao et al. [7] proposed an approximate exponential fit for the function $h(L)$ that maintains the general form of the BSIM3 model. However this requires a set of empiric fitting parameters for each technology node and for each transistor configuration.

In the novel approach proposed in this paper, the complete BSIM4 equation set has been used without any approximate equation and without any empiric fitting parameter. Exploiting the intrinsic calculations of electrical simulators, the PDF of the total leakage current has been expressed directly as function of L, i.e. $PDF\left(I_{(L)}\right)$. Hence, the function $L = g(I)$ is not used any more.

From a mathematical point of view, assuming that the drawn gate length has a Gaussian distribution with a fixed mean, μ_L, and standard deviation, σ_L, the PDF of I can be wrote as in Eq. 2.

$$PDF(I) = f_y(I) = \frac{f_x(L)}{h'(L)} \; f_y(I) = \left(\frac{1}{h'(L)}\right) \cdot \left(\frac{1}{\sigma_L\sqrt{2\pi}}\right) \cdot \exp\left(\frac{-(L-\mu_L)^2}{2\sigma_L^2}\right) \tag{2}$$

Finally, to calculate the mean and standard deviation (SD) of the leakage current distribution, it is sufficient to perform a numerical integration of $f_y(I)$ over the given range of leakage currents.

In the following sections is detailed the computation of the function $f_y(I)$, starting from the expressions of $h(L)$, $h'(L)$, and the PDF of I. The approach for a single device is presented in Section 2.2, and its application to the analysis of a complete logic gate is detailed in Section 2.3.

2.2 Leakage Distribution of a Single Transistor

The overall leakage in a nano-scaled device can be modeled as the summation of the four main leakage sources (Eq. 3).

$$I_{total} = I_{BTBT} + I_{sub} + I_{gate} + I_{gidl} \tag{3}$$

where I_{BTBT} is the reverse-bias source and drain p-n junction leakage current, I_{sub} is the sub-threshold leakage, I_{gate} represents the oxide tunneling current and the gate current due to hot carrier injection, and I_{gidl} is the drain leakage inducted by the high electric fields presents between drain and gate.

Hence, for leakage calculation, the device can be modeled as a sum of independent voltage controlled current sources, where each source describes one leakage contribution (Fig. 2) [12, 13].

Consequently, the first derivative of the total leakage current as function of the technological parameter (i.e. the gate length), $h'(L)$, can be calculated as a sum of independent partial derivatives (Eq. 4).

$$\frac{\vartheta I_{total}}{\vartheta L} = \frac{\vartheta I_{BTBT}}{\vartheta L} + \frac{\vartheta I_{sub}}{\vartheta L} + \frac{\vartheta I_{gate}}{\vartheta L} + \frac{\vartheta I_{gidl}}{\vartheta L} \tag{4}$$

Moreover, any approximated relation between the leakage components and the gate length is avoided since Eq. 2 can be directly enclosed into BSIM4 transistor model code (general for NMOS and PMOS transistors). The main effort remains into the analytical expression of the first derivatives of the leakage components as function of the considered technological parameter.

The enlarged Verilog-A BSIM4 transistor code is then used into an ELDO simulation. Parametric simulations are performed sweeping the technological parameter (e.g. L) from its theoretical minimum value ($\mu_L - 3\sigma_L$) to its theoretical maximum ($\mu_L + 3\sigma_L$). In the same time, other SPICE parameters are fixed using

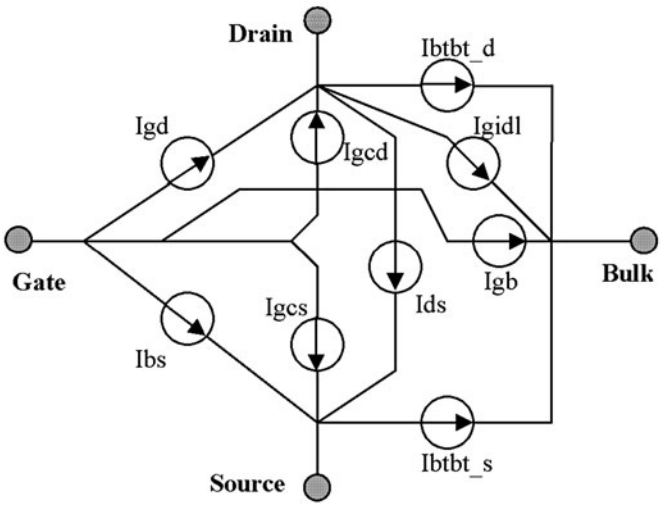


Fig. 2 Deep-submicrometer transistor model as a combination of voltage-controlled current sources

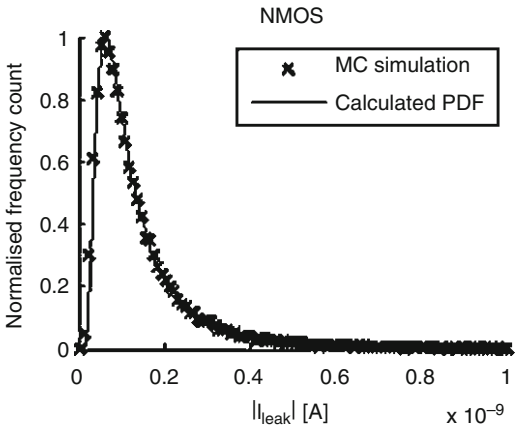


Fig. 3 Comparison of the Monte Carlo PDF and the analytical PDF found with Eq. 2 of the total leakage current for NMOS transistor, W/L[nm] = 200/60 (PMOS in the inset WL[nm] = 280/60–65 nm STMicroelectronics technology

standard technology model cards. The result of the simulation is a series of probability values relative to each possible leakage currents in the range defined by the variation of L. These values represent, in other words, the PDF of I. A comparison between the $PDF(I)$ obtained with the analytical formula and 50,000 Eldo Monte-Carlo simulations is shown in Fig. 3. The plots of the PDF's, including the tail portion, are matching and have a lognormal-like shape (due to the exponential dependency of leakage current on gate length).

2.3 Leakage Distribution of Logic Gates

In this section, the approach developed to estimate the leakage current distribution of individual transistors is extended to consider complete logic gates. The analytical expression of current as function of gate length in stacks of multiple transistors is quite complex, and cannot be expressed in an efficient analytical way without recurring to simplifying approximations or iterative algorithms [7–9].

An alternative approach is to approximate the PDF of the different leakage currents as lognormal distributions, and to calculate the total current as a sum of correlated (or uncorrelated) lognormals [14]. This methodology has however two main draw-backs: (a) the PDFs are not exactly lognormal distributions since their expression does not come directly from an exponential formula; (b) an exact closed-form expression for the sum of lognormal distribution does not exist (only approximation methods are present in literature for summation of lognormal random variables [15]). To overcome these disadvantages each leakage current (e.g. each current going through the logic gate terminals connected to the supply rail) should be accounted as a generic Random Variable (RV). Hence, considering, for instance, two of these independent random variables, X and Y, with density functions $f_x(x)$ and $f_y(y)$ respectively, their sum, $Z = X + Y$, is a RV defined by the convolution of $f_x(x)$ and $f_y(y)$[11] (Eq. 5).

$$PDF(Z) = (f * g)(z) = \int_{I_{\min}}^{I_{\max}} f_x(z-y) f_y(y) dy \tag{5}$$

where $I_{\min}$and $I_{\max}$ represent respectively the minimum and maximum leakage current resulting from the $\pm 3\sigma$ variation of the technological parameter around its mean value.

However, since the exact analytic expression of $f_x(x)$ and $f_y(y)$ is unknown, this integral has to be computed numerically. Consequently, considering the great number of leakage currents present in an actual complex logic cell, this approach could be too expensive in terms of computational time.

The approach presented in this paper is quite different and no approximation or intensive calculations are needed. In fact, the methodology employed to determine the PDF of a single transistor is sufficiently robust to provide also the analytical expression of the PDF of the total leakage of a complete logic gate. In such case, the estimation of the PDF of the leakage current, I_{total}, is the sum of all the currents impacting the leakage of the logic gate, and its derivative is the sum of each single derivative. For example, considering a two input Nand gate with one input connected to the positive supply rail and the other to ground (Fig. 4), the total leakage current is determined by the sum of the gate currents of transistors P1 and N1 and the source and substrate currents of transistors P1 and P2 (Eq. 6).

$$I_{total} = I_{gateP1} + I_{gateN1} + I_{subP1} + I_{subP2} + I_{dsP1} + I_{dsP2} \tag{6}$$

Since all those currents are statistically independent, the derivative of the total current (as function of the gate length) can be simply calculated as a sum of independent derivatives (Eq. 7).

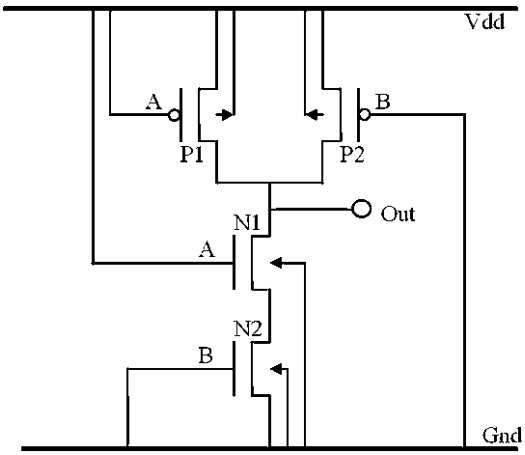


Fig. 4 Leakage currents in a two input Nand gate

$$\frac{\vartheta I_{total}}{\vartheta L} = \frac{\vartheta I_{gateP1}}{\vartheta L} + \frac{\vartheta I_{gateN1}}{\vartheta L} + \frac{\vartheta I_{subP1}}{\vartheta L} + \frac{\vartheta I_{subP2}}{\vartheta L} + \frac{\vartheta I_{dsP2}}{\vartheta L} + \frac{\vartheta I_{dsP2}}{\vartheta L} \tag{7}$$

The same result (with a difference in sign) can be obtained considering all the current flowing through the terminals connected to ground (gate, source and substrate of N2, substrate of N1 and gate of P2).

The expression of $\vartheta I_{total}/\vartheta L$can be directly used to analytically determine the PDF of the leakage current as function of the process parameter. Assuming that the drawn gate length has a Gaussian distribution with a fixed mean, the PDF(I) can be calculated as Eq. 8.

$$PDF(I) = f_y(I) = \left(\frac{1}{\vartheta I_{total}/\vartheta L}\right) \cdot \left(\frac{1}{\sigma_L\sqrt{2\pi}}\right) \cdot \exp\left(\frac{-(L-\mu_L)^2}{2\sigma_L^2}\right) \tag{8}$$

Furthermore, the methodology is completely compatible with all possible input combinations, since the biasing conditions for each transistor are analytically determined by the electrical simulator. It is important to note that, thanks to this approach, PMOS and NMOS transistors composing the logic gate can have dissimilar gate lengths without any loss in term of accuracy. Figure 5 validates the outlined leakage analysis methodology by Monte-Carlo comparison: the curves coming from analytical calculation and from the Monte-Carlo simulation match in shape and in statistical parameter in all the possible inputs combination. The drawn gate length is assumed to be normally distributed with $\pm 3\sigma$ variation around its mean.

3 Results

In this section, first the results obtained from the analytical approach, outlined in the previous sections, are compared to Monte-Carlo simulations for some individual standard cells. Then the difference between deterministic corner-based analysis and statistical analysis is shown. In the analysis that follows the mean and variance of the two approaches are compared.

Table 1 compares the analytical approach to Monte-Carlo simulations for two logic gates (considering all possible input vectors) implemented in 90 nm STMicroelectronics technology. The drawn gate length is assumed to be normally distributed with 3σ variation around its mean. The table shows that the error in estimating the mean leakage current is typically lower than 4%. The error in estimating the standard deviation is slightly bigger and on average is around 20%.

Table 2 compares the analytical approach to Monte-Carlo simulations for the same two logic gates in 65 nm technology, considering all possible input vectors. Also in this case the drawn gate length is assumed to be normally distributed with 3σ variation around its mean, being 60 nm. The table shows that the error in

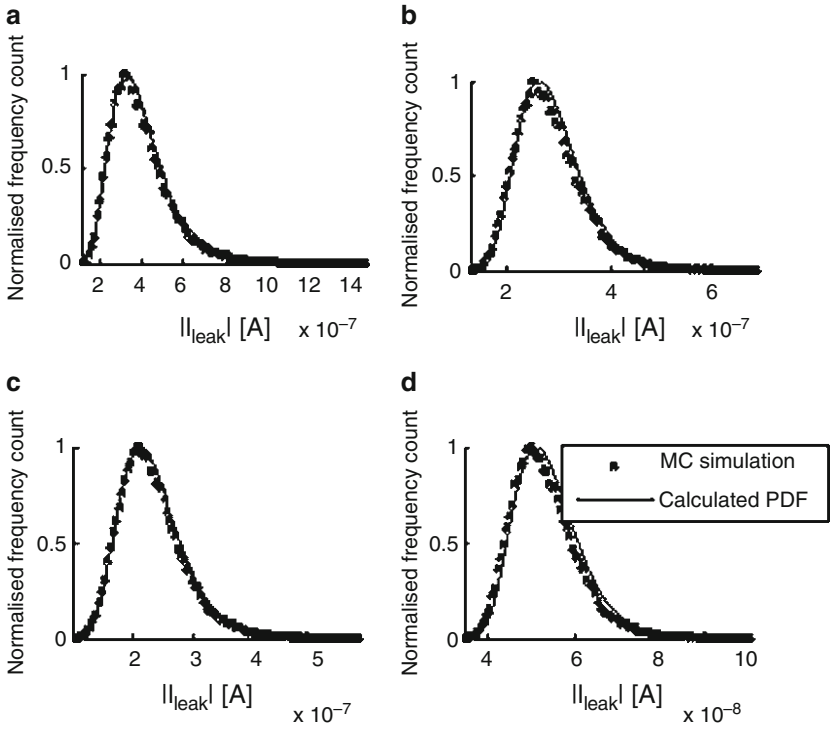


Fig. 5 Comparison of the Monte Carlo PDF and the analytical PDF from Eq. 2 of the total leakage current for a NOR logic gate – 65 nm STMicroelectronics technology. (**a**) Inputs: A = 0, B = 0. (**b**) Inputs: A = 0, B = 1. (c) Inputs: A = 1, B = 0. (d) Inputs: A = 1 B = 1

estimating the mean leakage current is typically smaller than 1% and shows an error of about 7% only for one particular input configuration. The error in the estimation of standard deviation is extremely small on average (around 1%).

Using the proposed parametric simulation strategy, the number of circuit simulation has been reduced by a factor of a hundred compared to standard Monte-Carlo simulations. This factor can be increased further compromising on the accuracy of the results.

Table 3 compares the median current leakage of the three logic gates estimated with the standard corner based approach and the new analytical statistical method. As can be seen, the traditional worst corner approach significantly overestimates the leakage since all devices composing the cell are assumed to be operating at the pessimistic corner point. In particular in 90 nm technology the worst case corner value is about six times bigger than the statistical leakage estimation. Furthermore, it is important to notice that the statistical estimation is very close to the nominal leakage value.

Table 1 Comparison of the analytical approach with Monte Carlo simulation for two different logic gates –90 nm STMicroelectronics technology. To simulate the worst leakage situation, temperature has been set to 125°C and Vdd to Vdd_{Nom}+10% for the two gates

Logic gate	Input vector	Mean [A] MC	Mean [A] analytic	E%	SD[A] MC	SD[A] analytic	E%
NAND 2	0 0	9.33e–12	9.40e–10	0.68	4.13e–10	3.65e–10	11.67
	0 1	5.77e–9	5.55e–9	3.78	3.17e–9	2.46e–9	22.39
	1 0	7.72e–9	7.37e–9	4.53	4.43e–9	3.30e–9	25.26
	1 1	2.10e–8	2.02e–8	3.78	1.18e–8	9.26e–9	21.73
NOR 2	0 0	1.66e–8	1.60e–8	3.13	9.44e–9	7.18e–9	24.04
	0 1	3.41e–8	3.31e–8	2.91	1.89e–8	1.51e–8	20.00
	1 0	2.58e–8	2.51e–8	2.50	1.39e–8	1.14e–8	18.23
	1 1	4.29e–9	4.28e–9	0.27	1.83e–9	1.59e–9	13.17

Table 2 Comparison of the analytical approach with Monte Carlo simulation for two different logic gates – 65 nm STMicroelectronics technology. To simulate the worst leakage situation, temperature has been set to 125°C and Vdd to Vdd_{Nom}+10% for the two gates

Logic gate	Input vector	Mean [A] MC	Mean [A] Analytic	E%	SD[A] MC	SD[A] analytic	E%
NAND 2	0 0	3.94e–8	4.24e–8	7.55	7.82e–9	8.04e–9	2.85
	0 1	1.52e–7	1.53e–7	0.86	5.43e–8	5.36e–8	1.35
	1 0	2.00e–7	2.02e–7	0.92	6.90e–8	6.82e–8	1.18
	1 1	4.99e–7	5.00e–7	0.16	1.18e–7	1.17e–7	1.07
NOR 2	0 0	3.91e–7	3.93e–7	0.59	1.40e–7	1.38e–7	1.43
	0 1	2.82e–7	2.84e–7	0.66	6.34e–8	6.29e–8	0.79
	1 0	2.31e–7	2.31e–7	0.19	5.33e–8	5.28e–8	1.02
	1 1	5.30e–8	5.42e–8	2.24	7.70e–9	7.89e–9	2.52

Table 3 Comparison of the corner base leakage analysis and the statistical methodology results – 90nm and 65nm STMicroelectronics technology. To simulate the worst leakage situation, temperature has been set to 125°C and Vdd to Vdd_{Nom}+10% for the two gates

Tech	Logic gate	Analytic estim. [A]	Worst corner Est. [A]	Ratio WC/ An.	Typ corner estim. [A]	Ratio Typ/ An
	XOR3	1.33e–7	2.84e–6	2.1	1.39e–6	1.0
65 nm	NAND2	2.03e–7	4.43e–7	1.9	3.06e–7	1.4
	NOR2	2.18e–7	4.67e–7	2.2	2.29e–7	1.0
	XOR3	1.91e–6	1.20e–5	6.3	2.18e–6	1.1
90 nm	NAND2	8.60e–8	5.29e–7	5.6	9.80e–8	1.1
	NOR2	1.70e–7	1.01e–6	5.4	1.92e–7	1.1

4 Conclusions

In this paper a novel methodology to estimate distributions of the leakage current of logic cell in presence of process statistical variations has been presented. It has been illustrated for single transistors and then applied to complex logic gates. The proposed framework has been verified by means of Monte-Carlo based simulations using an effective gate length of 90 and 65 nm. The results show that the methodology is accurate in estimating the overall mean and standard deviation of the total leakage current. It has been shown that using the analytical approach the pessimism introduced by corner based analysis can be significantly reduced while saving on the computational effort required for Monte-Carlo simulations.

References

1. S.R. Stg, J. Srivatsava, R.N. Tondamuthuru, Process variability analysis in DSM through statistical simulations and its implications to design methodologies, in *9th International Symposium on Quality Electronic Design*, Mar 2008, pp. 325–329
2. A. Agarwal, S. Mukhopadhyay, A. Raychowdhury, K. Roy, C.H. Kim, Leakage power analysis and reduction for nanoscale circuits. IEEE Micro **26**(2), 68–80 (March–April 2006)
3. B. Nikolic, Design in the power-limited scaling regime. IEEE Trans. Electron Devices **55**(1), 71–83 (Jan 2008)
4. S. Bhardwaj, S. Vrudhula, Y. Cao, LOTUS: Leakage optimization under timing uncertainty for standard-cell designs, in *IEEE Proceedings of the International Symposium on Quality Electronic Design*, 2006
5. S. Borkarm, T. Kamik, S. Narendra, J. Tschnz, A. Kershavarzi, V. De, Parameter variations and impact on circuits and microarchitecture, in *Proceedings of the IEEE on Design Automation*, 2003
6. B.H. Calhoun, Y. Cao, X. Li, K. Mai, L.T. Pileggi, R.A. Rutenbar, K.L. Shepard, Digital circuit design challenges and opportunities in the era of nanoscale CMOS, Proc. IEEE **96**(2), Feb 2008
7. R. Rao, A. Srivastava, D. Blaauw, D. Sylvester, "Statistical Estimation of Leakage Current Considering Inter- and Intra Die Process Variation", in *IEEE Proceedings of the International Symposium on Low Power Electronics and Design*, 2003
8. H. F. Dadgour, S.C. Linm, K. Banerjee, "A statistical Framework for Estimation o Full-Chip Leakage-Power Distribution under Parameter Variations", in *IEEE Transactions on Electron Devices*, vol. 54, no. 11, Nov. 2007
9. X. Li, J. Le, L.T. Pileggi, "Projection-Based Statistical Analysis of Full-Chip Leakage Power with Non-Log-Normal Distributions", in *Proceedings of the 43th conference on Design Automation*, 2006
10. A. Raychowdhury, S. Mukhopadhyay, K. Roy, "Modeling and Estimation of Leakage in Sub-90nm Devices", in *IEEE Transactions on VLSI Design*, 2004
11. A. Papoulis, *Probability, Random Variables, and Stochastic Processes* (McGraw-Hill, New York, 1991)
12. S. Mukhopadhyay, K. Roy, "Accurate Modeling of Transistor Stacks to Effectively Reduce Total Standby Leakage in Nano-scale CMOS Circuits", in *IEEE International Symposium on VLSI Circuits*, 2003

13. K. Roy, S. Mukhopadhyay, H. Mahmoodi-Meimand, "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS circuits", in *Proceedings of the IEEE*, vol. 91, no. 2, Feb. 2003
14. H. Chang, S. S. Sapatnekar, "Full-chip Analysis of Leakage Power Under Process Variations, Including Spatial Correlation", in *Proceedings of the 42th Conference on Design Automation*, 2005
15. H. Nie, S. Chen, "Lognormal Sum Approximation with Type IV Pearson Distribution", in *IEEE Communication Letters*, vol. 11, no. 10, Oct. 2007

On-Chip Circuit Technique for Measuring Jitter and Skew with Picosecond Resolution

K.A. Jenkins, Z. Xu, A.P. Jose, and K.L. Shepard

1 Jitter and Skew in VLSI Circuits

When VLSI circuits operate at multi-GHz frequencies, the demand for precise clock generation and distribution become more and more stringent. Inevitably, clock edges develop timing jitter, which is the deviation of the timing of the clock edges from their ideal values. Jitter arises from power supply noise on the circuits which distribute the clock, and from the phase-locked loops (PLLs) which generate a high frequency clock by multiplying a very stable lower frequency signal. Since computation requires that certain logical operations are completed within each clock cycle, a certain amount of jitter is assumed in the design of a chip, by including it in the clock budget. Jitter which is too large can impact the allowed clock budget, or even cause data transmission and computation errors. The measurement of jitter is required for high-speed circuits, to determine if timing specifications and margins are met. In contemporary circuits, where clock periods may be as small as 250 or 300 ps, rms jitter may be required to be as small as a few picoseconds. Conventional measurement of such small timing delays requires driving the signal of interest off-chip with high fidelity off chip drivers. Measurement is performed with a high performance oscilloscope or similar instrument, limiting characterization to only a few samples.

Performing jitter measurement with an in-situ measurement circuit clearly has the advantages of eliminating the drivers and permitting measurements of multiple internal nodes. Furthermore, in-situ measurements can be made during ordinary circuit operation, eliminating the need for special test conditions and fixtures. There are several published methods by which on-chip jitter measurements can be achieved, but no demonstration of a simple, compact design with high temporal

K.A. Jenkins (✉)
IBM T.J. Watson Research Center Yorktown Heights, NY 10598, USA
e-mail: jenkinsk@us.ibm.com

Z. Xu, A.P. Jose, and K.L. Shepard
Columbia University, USA

A. Amara et al. (eds.), *Emerging Technologies and Circuits*,
Lecture Notes in Electrical Engineering 66, DOI 10.1007/978-90-481-9379-0_15,

resolution [1–5]. To address this need, jitter measurement circuits with sub-ps resolution have been developed and are described here [6, 7]. The technique used in these circuits has a very simple design which measures the jitter of a signal of interest with respect to a reference signal, or jitter of one clock edge with respect to the preceding edge. The example circuits, implemented as described in this chapter, reproduce the jitter measured with a conventional oscilloscope, and have a measured resolution limit of better than 0.4 ps rms. In contrast to previous methods, the circuit technique uses a single latch, and is thereby small enough that it can easily be placed in multiple locations on a complex chip. The circuit is operated and controlled entirely with low frequency digital input and output signals.

In addition to jitter, clock signals may exhibit skew due to mis-design or across-chip process or voltage variation. Skew refers to the difference of the *average* arrival time of different signals, generally in different locations of a chip. Skew can be detrimental, in that clock signals may not be sufficiently synchronized in different areas of the chip. Hence, similar to jitter, some skew is assumed in the clock budget, but if larger than expected, it can lead to errors or reduced system speed. Skew of tens of ps or less is almost impossible to measure by sending the signals off-chip, because of probe and cable delays, unless multiple signals are multiplexed into a single off-chip driver.

The most common previous measurement of jitter and skew has been the time-to-digital converter (TDC) method based on a tapped delay line. This method uses a delay chain with tapped with multiple latches which evaluate whether or not a data signal has arrived before a clock signal. The concept is illustrated in Fig. 1. A reference clock is launched simultaneously to the clock inputs of a series of latches, while the signal is delayed between each latch, by, for example, an inverter. The latches for which the signal arrive prior to the clock register a "1", while the remainder register "0" (after accounting for the stage-to-stage inversion of the delay

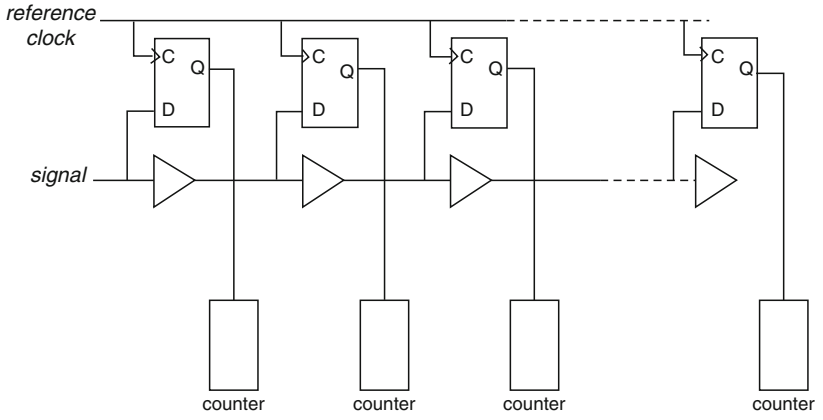


Fig. 1 Tapped delay line for time-to-digital conversion (TDC)

chain). A digital code is thereby generated which represents the arrival time of the signal. Jitter of a signal can be obtained by reading out this digital signal repeatedly, and determining the statistical distribution of the arrival time. Skew can be determined by using this TDC circuit in multiple locations on a chip and comparing the average encoded digital value [8].

Clearly, the time resolution of the tapped delay line TDC is limited by the delay of a single stage in the delay line. An improvement of this resolution is achieved using a vernier technique, as illustrated in Fig. 2 [1]. In this technique, the clock is also delayed by a delay chain which has a slightly different delay per stage (t1) from that of the primary delay chain (t2). This enhances the resolution by the usual vernier technique, so a resolution of (t2–t1) is achieved, although it also reduces the dynamic range of measurement.

Even with this improvement, the TDC has the limitation that the stage-to-stage delay is assumed to be constant, which is unlikely to be true in modern CMOS technologies, unless very wide devices (which are more uniform that minimum size devices) are used for the delay elements. Furthermore, the TDC is sensitive to power supply voltage noise since delay depends on voltage. In fact, it has been shown that a TDC circuit intended for skew measurement can take advantage of this sensitivity to measure power supply noise [9]. Therefore power supply noise and signal jitter and skew may be indistinguishable. Finally, the TDC requires many latches to capture the range of values expected. The measurement circuit may be larger than the PLL used to generate the signal being measured.

The circuits described in this work addresses all of the issues raised above. Instead of multiple latches, the technique uses a single latch. The difference in stage delay is eliminated by the use of the single latch and an on-chip calibration method,

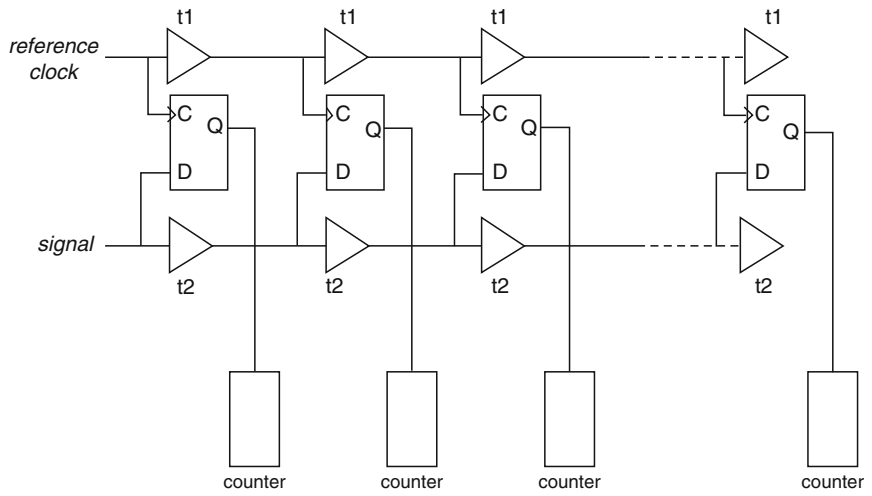


Fig. 2 Tapped delay line for time-to-digital conversion using vernier delay

and power supply noise sensitivity is minimized by the use of two delay chains, so that the effect of power supply noise is common to both, and therefore negligible.

The principle of the measurement circuit is explained by the simplified drawing in Fig. 3. A signal with jitter drives the data input of an edge-triggered latch, or flip-flop, via a delay chain of fixed time delay. A signal without jitter, ie, a reference clock, drives the clock input of the latch via a delay chain with a variable delay. (The clock and data might, in fact, be the same signal delayed by a period, as explained below.) The latch registers those data signals which arrive before the clock signal. Figure 4 illustrates this concept. The data signal has a time distribution, or jitter, represented here as Gaussian. Depending on the timing of the reference clock with respect to that distribution, different numbers of data events are latched. Thus the latch effectively integrates all the signals which arrive before the clock, indicated by the shaded areas. By systematically varying the delay of the clock so that it sweeps through the distribution, a cumulative distribution function (CDF) of the data arrival time is generated. The original distribution can be determined by differentiating the CDF. Although illustrated in Fig. 4 with a Gaussian jitter distribution, the principle is quite general, and applies to any distribution of edges.

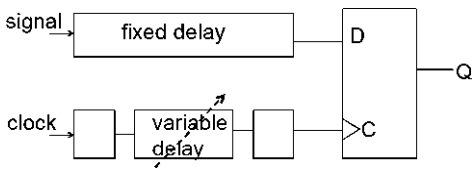


Fig. 3 Illustration of the single latch jitter measurement technique

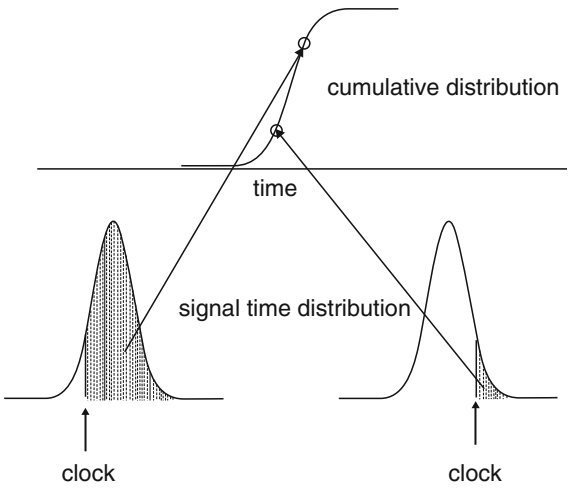


Fig. 4 Illustration of the generation of a cumulative distribution by varying the clock arrival time at the latch

This circuit technique can be used to measure long-term jitter, period jitter, and skew. The performance has been demonstrated in stand-alone designs, as well as in a circuit incorporated in a novel clock distribution network, where it also provides the measurements of skew necessary to demonstrate an active-deskewing technique.

This basic method can be used for a variety of measurements, including:

- Long-term (or tracking) jitter
- Long-term jitter of random data
- Long-term jitter of multiplied or divided clocks
- Period jitter
- n-Cycle jitter
- Point-to-point skew or phase error

2 Circuit Design

A more complete schematic of one implementation of this circuit is shown in Fig. 5. The two delay chains and the flip-flop are surrounded by additional circuit elements to complete the measurement. The variable delay chain uses the same number of delay elements as the fixed delay chain, with variable delay elements inserted in the middle of this chain. This ensures that signals entering and leaving the variable delay chain are not affected by the variable elements except for their delay, so that, for example, the clock and data signals present at the latch input always have the same amplitude, independent of the delay. Also included in this diagram are

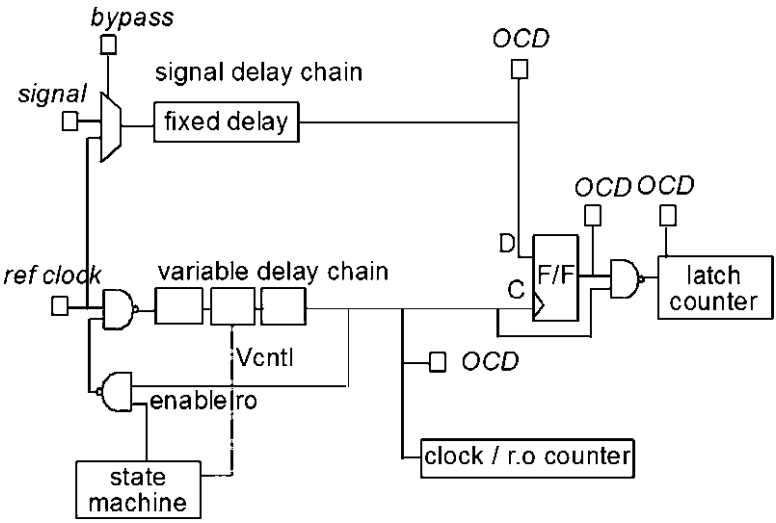


Fig. 5 Block diagram of long-term jitter measurement circuit, including observation points

indications of observation points, or off-chip drivers (OCDs) which are incuded for circuit verification in the evaluation prototype, but which are not required for circuit operation.

On-chip counters are used to accumulate counts used for on-chip measurement. The output of the latch is NANDed with the reference clock, essentially causing a latch reset for each clock, so that a pulse is created whenever the latch is set high. For any given value of the delay, the latch counter counts those pulses which occur when the signal arrives before the clock, and the reference counter counts the total number of clocks. The ratio of latch counts to reference counts represents one value of the CDF. The delay is swept through an appropriate range of values in order to generate the entire CDF. The latches and counters are operated at the reference clock rate, but the delay increment and read-out operations are independent of the clock, and no synchronization is required.

Time-calibration of the variable delay chain is also performed with on-chip counters. The variable delay elements can be constructed from many different circuits. In the examples presented here, the delay elements are voltage controlled inverters and current-controlled differential buffers. Regardless of the delay control mechanism, a key technique to calibrate the delay is configuring the variable delay path as a ring oscillator by simply setting the “ro enable” and ”ref clock” inputs to the NANDs to the “high” state. In the first example described below, the delay was controlled by voltage which was applied by an off-chip source, but an on-chip voltage generated by a resistor divider network, operating from a fixed voltage, would work just as well. The ring oscillator enables a delay-to-control calibration. The frequency of the ring oscillator is determined by counting the pulses in an on-chip counter for the same time as the reference counter is counting the reference clock. This relates the ring oscillator frequency to the reference clock frequency, which is always known.

$$f_{RO} = f_{ref} \frac{ROcounts}{refcounts},$$

Where f_{RO} is the ring oscillator frequency, f_{ref} is the reference clock frequency, and *ROcounts* and *refcounts* are the values in those counters when measured for the same time interval. The delay per volt is given by the derivative of the ring oscillator period with respect to voltage. Since only the derivative is used, the additional NANDs needed to convert the delay chain to a ring oscillator do not contribute to the delay calibration. This is independent of the nature of the variable delay elements. As a result of the calibration, jitter measurements with a correct time scale are obtained with no additional adjustments. This delay chain method eliminates any concern about stage-to-stage delay variation, since the entire chain is used, and is calibrated. Furthermore, it is not required that the delay be a linear or even uniform function of the control, although resolution may be degraded by severe non-uniformity.

The circuit has been demonstrated in three designs. Two were stand-alone pad cage design for evaluation and characterization, using different CMOS technologies. In these designs, signals were applied though high bandwidth probes to test the

Table 1 Key features of the three implementations of the jitter and skew measurement circuit

Design	Technology (μm)	Function	Clock period (ps)	Delay elements
Stand-alone	0.13	Long-term jitter only	Any	Inverters
Stand-alone	0.09	Long-term or period jitter	900	Inverters
Embedded	0.18	Period jitter and skew	500	Differential buffers

response of the measurement circuits to known jitter, and also to measure the impact of power supply noise. The circuit was also incorporated at several test points in a clock distribution network, where it was used to measure skew as well as jitter. Key features of these designs are listed in Table 1.

A differential (sense-amplifier) flip-flop [10] was used in all three implementations of this circuit because it has a very small metastability window and setup time. Latch metastability can determine the resolution limit of this timing circuit, since it results in a distribution of counts, rather than a unique state. The simulated metastability window of the latch is less than 0.1 ps, consistent with the high resolution demonstrated by measurement below. This particular flip-flop is also quite insensitive to power supply noise. Simulations of a 1.0 V latch, show that the setup time of the latch stays relatively constant over a wide range of power supply variation, thus introducing little or no false jitter in a noisy environment.

3 Jitter Measurement with On-Chip Circuit

3.1 Jitter Definitions

Jitter can characterized by several different figures of merit. For many purposes, it is only necessary to consider long-term jitter and period jitter, which are distinguished in Fig. 6. Long-term jitter, in which a signal's jitter is compared to a reference signal, shown in Fig. 6a, describes how well a clock tracks a reference. Although the illustration shows the situation that the clock with jitter has the same frequency as the reference, in most cases the clock will actually be multiplied by a phase-locked loop to give a higher frequency. This type of jitter pertains to the case of several sub-circuits or chips being synchronized to a single clock. Period jitter is illustrated in Fig. 6b, in which the variation of the period of the signal is compared on successive cycles. This type of jitter is important, for example, in establishing latch hold time requirements, and for determining timing requirements for I/O data circuits where there is no accompanying clock. Period jitter is generally smaller than long-term jitter, making measurement more demanding.

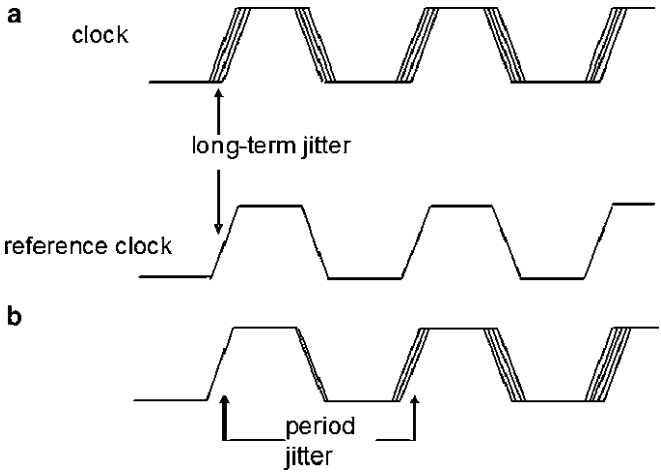


Fig. 6 Definition of long-term jitter (**a**) and period jitter (**b**)

3.2 Long-Term Jitter

Measurement examples of long-term jitter with the 0.13 μm design are shown here. The measurement of jitter with this circuit requires a calibration of the delay chain, as explained earlier. A typical calibration is shown in Fig 7, in which one half of the reciprocal of the measured frequency is graphed as a function of the control voltage. The factor of a half arises because the period of a ring oscillator corresponds to two loops around the ring, one for the rising edge and one for the falling edge. A delay sensitivity of about 0.5 ps/mV is obtained in this example. The sensitivity is determined by the number of variable stages used, and their voltage sensitivity, so a delay chain with selectable sensitivity could easily be designed. Only the slope of the calibration is used, as explained above; the extra delay of the NANDs in the ring oscillator loop is of no consequence. For large excursions of the control voltage, the delay can be seen on a high-bandwidth oscilloscope using the observation points of the test circuit, and has served to corroborate the ring oscillator calibration.

An example of the measurement is shown in Figs. 8 and 9. An external pulse generator creates two signals, a reference clock and a signal of the same frequency, but with jitter added. Figure 8 shows the jitter as measured with a sampling oscilloscope in the conventional manner. The reference clock signal triggers the oscilloscope, and the signal with jitter is applied to the oscilloscope input. The rising edge is displayed, and by selecting a measurement level, a histogram of the timing arrival at that level is displayed.

Figure 9a shows the CDF measured when the same signals are applied to the on-chip circuit. The CDF is shown as a function of control voltage of the delay line, (Vcntl). The jitter distribution is recovered by differentiating the CDF with respect to this control voltage and applying the voltage-to-time calibration of Fig. 7. The result of these two operations is shown in Fig. 9b. The expected jitter spectrum is

obtained. To estimate the standard deviation of the distribution, the data are fit to a Gaussian function which is superimposed on the measured data. The on-chip jitter circuit gives a standard deviation of 3.2 ps, compared with 4.1 ps with the oscilloscope. The slightly larger value with the oscilloscope is due to its larger trigger jitter, compared to that of the on-chip circuit.

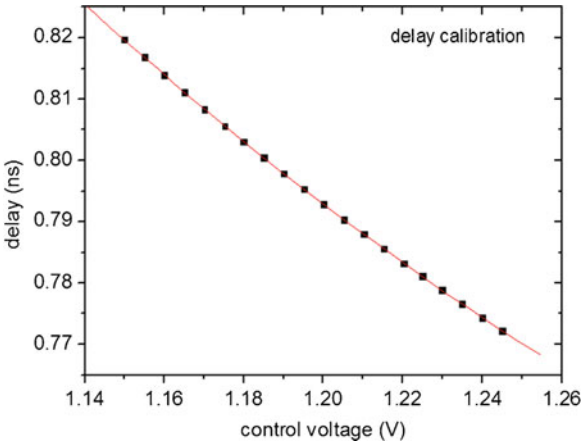


Fig. 7 Example of delay chain calibration

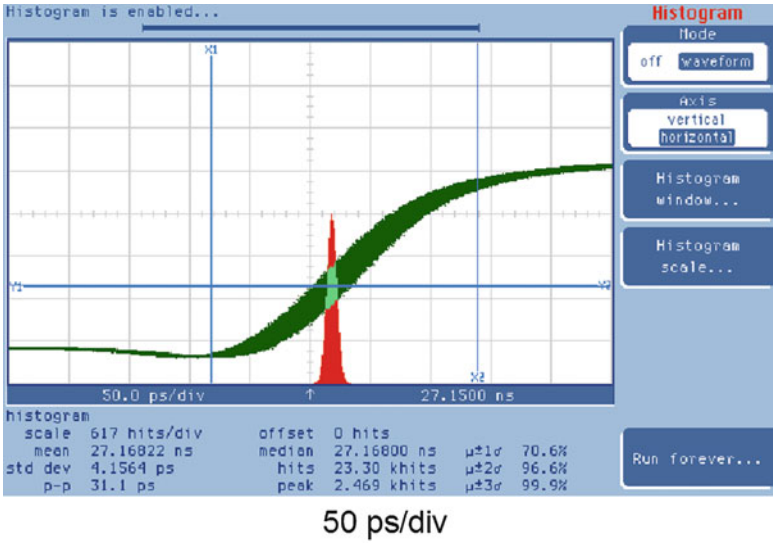


Fig. 8 Oscilloscope screen showing measured jitter signal which is applied to the on-chip circuit

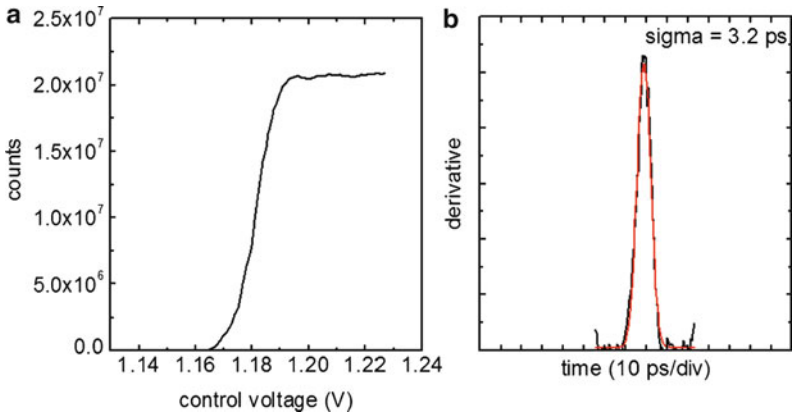


Fig. 9 Example of measured CDF (**a**) and (**b**) calibrated jitter obtained from the CDF, to be compared with Fig. 8

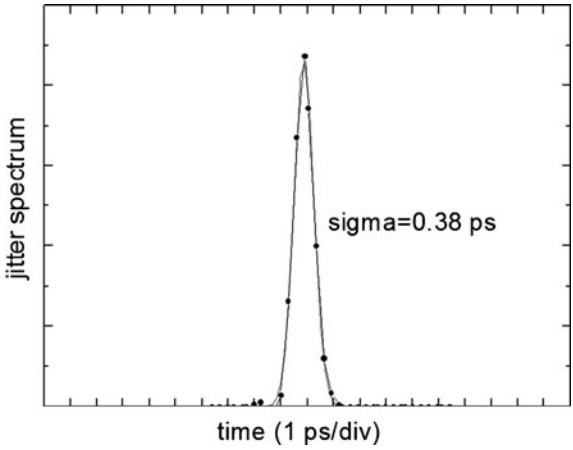


Fig. 10 Example of resolution measurement of the on-chip jitter measurement circuit

The intrinsic measurement resolution of the on-chip circuit is determined by applying a signal with no added jitter. Referring to Fig. 5, this is achieved by switching the 'bypass' control so that one signal, the reference clock, goes to both the data and clock delay chains and latch. Then sweeping the delay as in the usual measurement, the circuit measures its intrinsic resolution, that is, the apparent jitter measured when the clock and data signals are identical, that is there is no jitter on the measurement signal. This is equivalent to measuring the jitter on the triggering signal of an oscilloscope. This measurement is shown in Fig. 10. This version of the circuit shows a long-term jitter resolution of less than 0.4 ps rms. This resolution is thought to be dominated by the metastability window of the latch, but the small

value indicates that, in general, a flip-flop is an excellent timing discriminator, more precise than might be expected from the cutoff frequency, or rise time, of the FETs. This resolution is better than most commercial high-performance oscilloscopes available today.

3.3 Period Jitter

The circuit described above can only measure jitter with respect to a reference, that is, long-term jitter. However, it is a fairly simple matter to modify the design to measure period jitter. The idea is to have the signal of interest drive two delay lines, one fixed, and one variable as before. For this measurement, though, the one delay line, say, the fixed one, is longer than the other by approximate the period of the signal to be measured. In this way, the signal going to the data input of the latch arrives one cycle later than the signal going to the clock input of the latch. Thus the jitter will be measured as in Fig. 6b.

An implementation of this concept is shown in Fig. 11. In this design, switches are added which make it possible to switch in the extra delay to enable period jitter, or leave it configured as in the design of Fig. 5, used for long-term jitter. There is also a switch ("bypass") at the input which routes a single signal (in this case, it is called the reference clock) to both delay chains. A circuit for measuring period jitter is more restrictive than the earlier one, in that the signal period must be decided

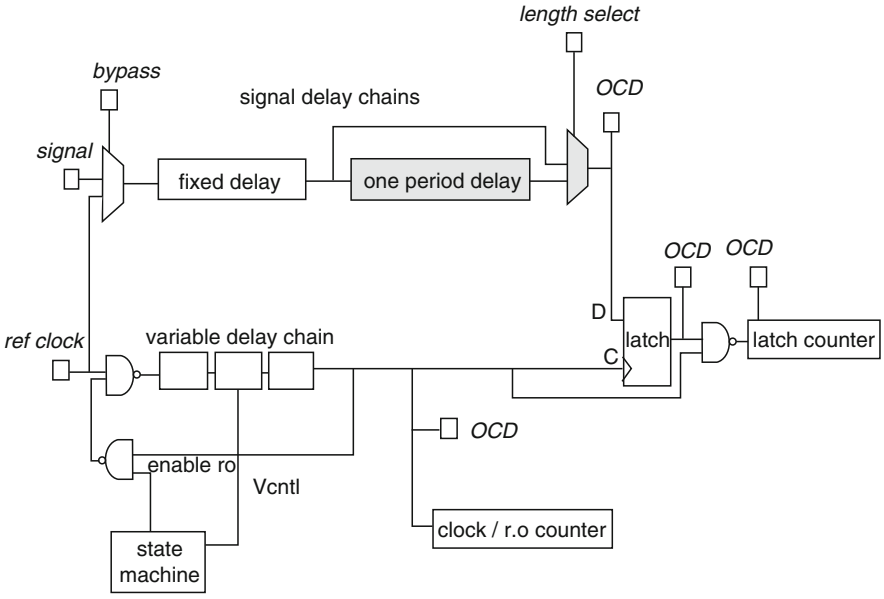


Fig. 11 Block diagram of circuit modified to measure period jitter as well as long-term jitter

prior to the design, in order the insert the correct amount of delay. It is possible, of course, to design a circuit with a variety of cycle delay blocks, which can be inserted or removed, as needed, to cover a range of frequencies.

A demonstration of period jitter measurement is done with the 90 nm version of the stand-alone circuit described in Fig. 11, where the design target was a period of about 900 ps, or a frequency of about 1.1 GHz. As before, the circuit is evaluated by applying a signal of known jitter through the high frequency probes to the circuit, and sweeping through the variable delay to generate a CDF. The calibration procedure is exactly the same as before. For comparison, the applied input jitter is measured conventionally with an oscilloscope.

For this example, a complex jitter distribution is created by modulating the phase delay control of a pulse generator with a sinusoidal signal, resulting in the a histograms with two pealks shown in Fig. 12. (It is noted that the jitter of the falling edge, which indicates the impact of jitter on duty cycle, can be measured at half the target frequency, as in this particular example. Hence in this example, the jitter is measured on the falling edge of a 540 MHz clock.)

The on-chip jitter measurement clearly reproduces the applied jitter as measured on the oscilloscope. It also demonstrates, by the correct spacing of the two peaks, that the time calibration of the delay chain is accurate.

Similar to the example from long-term jitter described above, the resolution for period jitter can be measured by applying a signal with no jitter, ie, well below 1 ps rms. Measurements show that the on-chip resolution in this case is about 1 ps rms, not as good as the 0.4 ps shown previously. This is due to power supply noise generated in the off-chip-drivers used in the demonstration test circuit. In a totally on-chip circuit without off-chip drivers, this noise would not be present.

However, power supply noise in normal operation is a potential problem in the period jitter measurement. Because of the extra path delay used to compare cycle-delayed edges, the circuit is susceptible to jitter generated by power supply noise

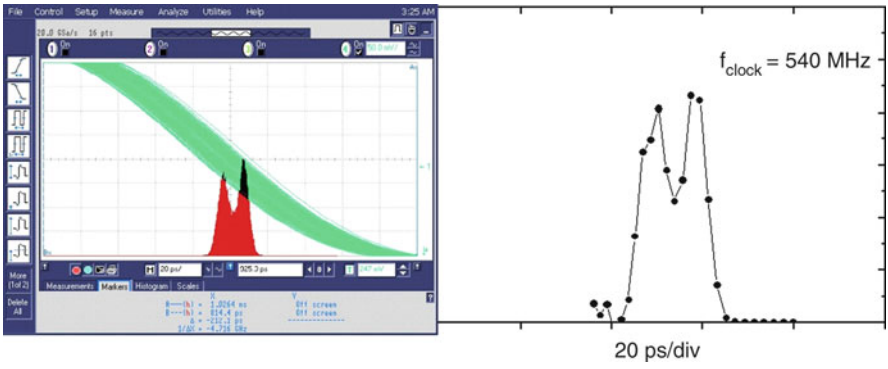


Fig. 12 Oscilloscope screen showing complex period jitter applied to test circuit (*left*) and resulting on-chip measurement (*right*)

affecting the delay line with the added delay segment. Power supply noise will increase or decrease the delayed signal from its true value, introducing jitter at the input to the latch. In the test circuit, such noise can be intentionally applied directly in order to observe the impact. As an example, when Gaussian noise of 30 to 100 mV is applied to the delay chain voltage, there is a broadening of the resolution spectrum from 1.0 to 2.3 ps rms, when the signal frequency is 1.1 GHz. While this is a concern, it is noted that in the stand-alone circuit, simple CMOS inverters are used for the delay chains, making them particularly sensitive to power supply voltage. A better design of this measurement technique, as shown below, uses delay elements with high power supply rejection ratio (PSRR). As mentioned above, the flip-flop is quite insensitive to power supply noise, and is not a source of apparent jitter.

4 Clock Skew Measurement with On-Chip Circuit

The difference in average arrival time of clock at various locations on a chip is called clock skew. Some skew is tolerated, or even expected, depending on the design of the clock network, but it must nonetheless be known accurately. The jitter measurement circuit, appropriately configured, can give an accurate measurement of clock skew. For this measurement, the multiple copies of on-chip jitter measurement circuits are placed at various points on the clock tree where skew is to be measured. However, instead of each circuit having its own variable delay line, the multiple instances of the measurement circuits use a single common variable delay chain. This method of skew measurement is illustrated in the block diagram of Fig. 13. The delay is varied in the usual way, and each latch registers counts according to the arrival time of its respective signal. In this way, the timing difference, or skew, between test points is given by the difference of the means of the generated distributions of those points. Use of a single variable delay line is analogous to using a common time base on an oscilloscope which displays waveforms from multiple inputs, showing the time displacement as offsets of the waveforms appearing on the screen.

This measurement capability is demonstrated in a 0.18 μm, 1.8 V design [7]. In order to measure period jitter of a novel 2 GHz resonant-clock distribution network, jitter measurement circuits are included on the test chip [11]. The clock circuit is designed for very low period jitter, hence the need for a high-resolution on-chip measurement. In order to avoid sensitivity to power supply noise, the delay elements in this implementation use differential buffers, with the variable elements controlled by their tail currents and pFET gate voltage [12].

An example of measured clock skew is shown in Fig. 14. In this figure, the cumulative jitter distribution generated by the on-chip circuit is measured at two points on the clock tree. Since the jitter distribution at each point is about the same, the CDFs have the same shape, and it is sufficient to measure just the shift of the midpoint of the CDF, eliminating the need for differentiating. In this example, the shift is about 16 ps. Clearly the on-chip circuit has more than enough accuracy to

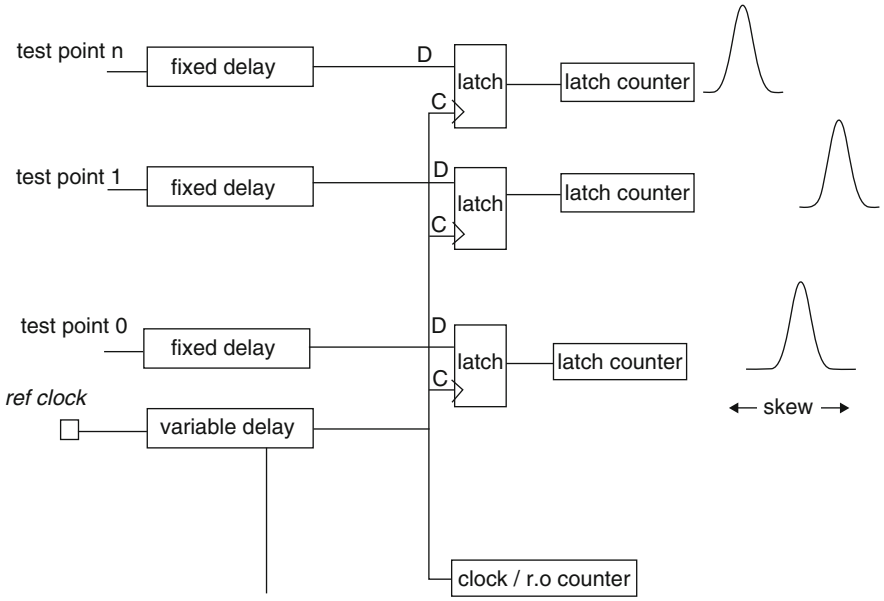


Fig. 13 Block diagram of use of the jitter measuring circuit to measure clock skew between various test points

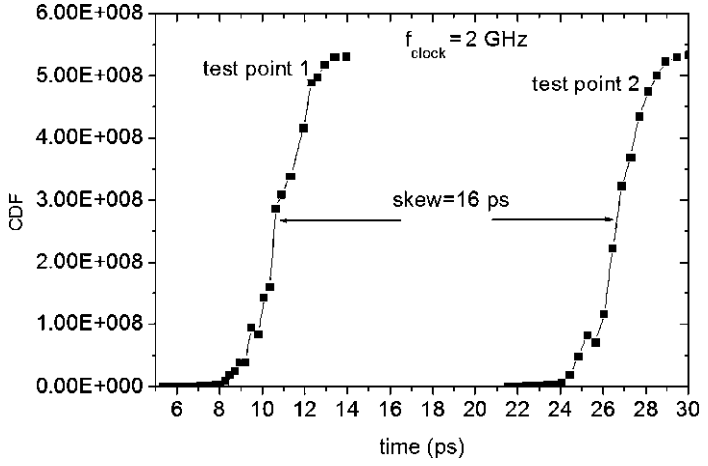


Fig. 14 Example of measuring clock skew between two test points using only the cumulative distributions (CDFs)

measure skew of this magnitude. The accuracy of skew measurement is limited by the jitter resolution, which is on the order of 1 ps. Hence, the same circuit can be used to measure long-term jitter, period jitter, and clock skew, with ps or sub-ps accuracy.

5 Summary

An on-chip jitter and skew measurement circuit has been presented. By using two delay lines and a single latch, the circuit measures the cumulative probability distribution due to jitter. By differentiating and using an on-chip delay calibration, the original distribution is recovered. In addition, use of a common time base for multiple placements of the circuit enables measurement of clock skew. The circuit has been demonstrated in stand-alone test sites for characterization, and has been used to measure the jitter and skew of an experimental clock network. It has a demonstrated resolution of less than 1 ps up to 2 GHz clock frequency.

Acknowledgements The author would like to thank his collaborators, A. P. Jose, K. L. Shepard, Z. Xu, and D. F. Heidel, for contributions to this work, and D. Beisser for physical design of some of the test circuits.

References

1. P. Dudek, S. Szczepanski, J. Hatfield, A high-resolution CMOS time-to-digital converter utilizing a vernier delay line. IEEE J. Solid-State Circ. **35**, 240–247 (2000)
2. S. Sunter, A. Roy, BIST for phase-locked loops in digital applications, *International Test Conference*, 1999, pp. 532–540
3. A.M. Frisch, T.H. Rinderkneckt, US Patent No. 6295315
4. T. Xia, J.-C. Lo, Time-to-voltage converter for on-chip jitter measurement. IEEE Trans. Instru. Measure. **52**, 1738–1748 (2003)
5. M. Ishida, K. Ichiyama, T.J. Yamaguchi, M. Soma, M. Suda, T. Okayasu, D. Watanabe, K. Yamamoto, A programmable on-chip picosecond jitter-measurement circuit without a reference-clock input, *ISSCC Digest of Technical Papers 2005*, 2005, pp. 512–513
6. K.A. Jenkins, A.P. Jose, D. Heidel, An on-chip jitter measurement circuit with sub-picosecond resolution, *Proceedings of ESSCIRC*, 2005, pp. 157–160
7. K.A. Jenkins, K.L. Shepard, Z. Xu, On-chip circuit for measururing period jitter and skew of clock distribution network, *Digest of Custom Integrated Circuits Conference*, 2007, pp. 157–160
8. P. Restle, R.L. Franch, N.K. James, W.V. Hupott, T.M. Skergan, S.C. Wilson, N.S. Schwartz, J.G. Clabes, Timing Uncertainty measurements on the Power5 Microprocessors, *ISSCC Digest of Technical Papers,* 2004, pp. 1–2
9. R. Franch, P. Restle, N. James, W. Huott, J. Friedrich, R. Dixon, J. Weitzel, K. Van Goor, G. Salem, On-chip timing uncertainty measurements on IBM microprocessors, *International Test Conference*, 2007, pp. 1–7
10. J. Montanaro, R.T. Witek, K. Anne, A.J. Black, E.M. Cooper, D.W. Dobberpuhl, P.M. Donahue, J. Eno, W. Hoeppner, D. Kruckemyer, T.H. Lee, P.C.M. Lin, L. Madden, D. Murray, M.H. Pearce, S. Santhanam, K.J. Snyder, R. Stehpany, S.C.A. Thierauf, 160-MHz, 32-b, 0.5-W CMOS RISC microprocessor. IEEE J. Solid-State Circ. **31**, 1703–1714 (1996)
11. Z. Xu, K.L. Shepard, Low-Jitter active deskewing through injection-locked resonant clocking, *Digest of Custom Integrated Circuits Conference*, 2007, pp. 9–12
12. J.G. Maneatis, Low-jitter process-independent DLL and PLL based on self-biased techniques. IEEE J. Solid-State Circ. **31**, 1723–1732 (1996)

Part VI
Analog and Mixed Signal

DC–DC Converter Technologies for On-Chip Distributed Power Supply Systems – 3D Stacking and Hybrid Operation

Makoto Takamiya, Koichi Onizuka, Koichi Ishida and Takayasu Sakurai

1 Introduction

Recently System-on-a-Chip (SoC) and System-in-a-Package (SiP) are getting more and more interest as major integration technologies. They are often used to integrate various types of circuit blocks from processors and memories to analog circuits. Each block demonstrates a different optimal supply voltage (VDD) and the difference tends to increase as the technology scales down. For example, memory and analog circuits tend to prefer higher voltage compared with logic blocks. Figure 1 shows the VDD trends for precision analog/RF, performance analog/RF, high performance logic and lowpower logic with the design rule trends according to the International Technology Roadmap for Semiconductors (ITRS) 2006 update [1]. Multiple-VDD implementation is therefore essential in lowpower and high-performance systems.

On the other hand, the power line integrity issue, including IR drop and voltage noise has become obvious in recent years. To improve the power integrity issue, it is valuable to decrease the input current of a package. To supply power whose amount is the same as before, the package input voltage must be increased and down converted at the vicinity of the load circuit.

To realize the multi-VDD implementation and solve the power integrity issue as well, the distributed on-chip power supply circuits are useful. The concept of the distributed power supply is shown in Fig. 2. High voltage is distributed by a main power grid in a package and is then converted to the lower voltages at the vicinity of the target blocks by distributed on-chip voltage converters. By doing so, the input

M. Takamiya (✉)
VLSI Design and Education Center and Institute of Industrial Science
University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo, Japan, 153-8505
e-mail: mtaka@iis.u-tokyo.ac.jp

K. Onizuka, K. Ishida and T. Sakurai
Institute of Industrial Science, University of Tokyo, Tokyo, Japan

A. Amara et al. (eds.), *Emerging Technologies and Circuits*,
Lecture Notes in Electrical Engineering 66, DOI 10.1007/978-90-481-9379-0_16,

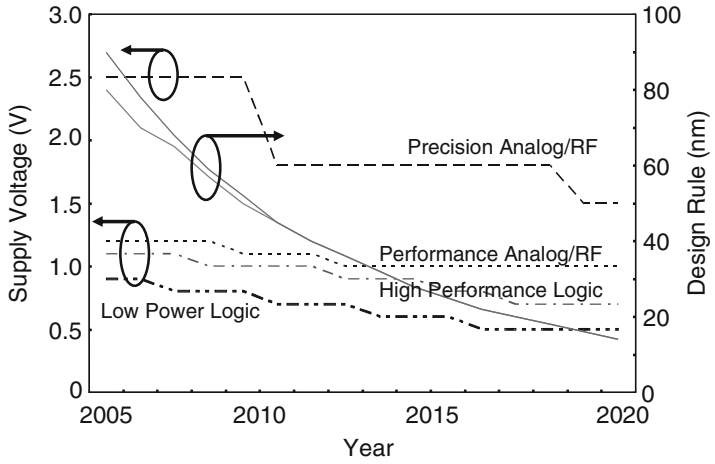


Fig. 1 Supply voltage trend with design rule trend according to ITRS

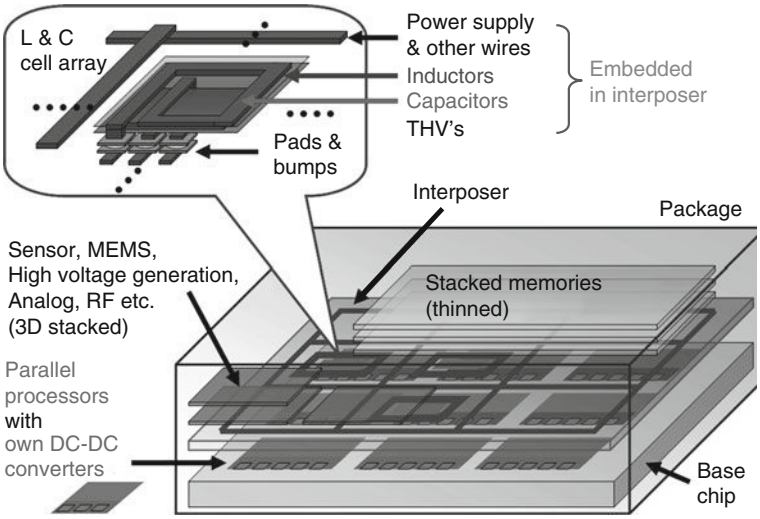


Fig. 2 Concept of distributed power supply system

current to the package is reduced. This approach reduces cost and power integrity issues.

In the first half of this paper, 3D-structured buck converters are presented with some theoretical analyses. The 3D-stacking scheme is one of the key technologies to realize high efficiency, compact and cost effective on-chip distributed power supply systems.

Moreover, the power gating scheme has become one of the most important technologies in recent years to cutoff the leakage current when a load logic circuit is in a standby mode [2]. A buck converter with built-in power gating scheme is presented as well in the second half of this paper. Hybrid operation of a buck converter and a linear regulator successfully removes the conventional power gating switches from the logic circuit.

2 3D-Structured Buck Converter

2.1 Theoretical Analysis of Power Efficiency

Buck converters are good candidates for high efficiency DC–DC conversion in a distributed power supply system. In order to implement on-chip buck converters for on-chip power supply systems, the two following conditions should be met. First, all elements of the converter must be integrated at least in a package. Secondly, the implementation cost per one voltage domain must be minimized while maintaining high power efficiency. To maximize the power efficiency, the power loss must be minimized. The major components of power loss are classified into three parts: dynamic switching loss of the switching transistors, resistive loss of the switching transistors and the resistive loss of the inductor as shown in Fig. 3 [3].

C_B and R_B are described as follows when W_P, W_N, C_P, C_N, R_P, and R_N represent gate widths, parasitic capacitances per unit gate length, and parasitic series resistances per unit gate length of high side PMOS and low side NMOS transistors. D is the duty cycle which equals to V_{OUT}/V_{IN}.

$$C_B = W_P C_P + W_N C_N, \tag{1}$$

$$R_B = \frac{R_P}{W_P} D + \frac{R_N}{W_N}(1 - D). \tag{2}$$

Here, effective values of C_P and C_N (Ceff) are expressed as $1.3C_G + C_J$ where C_G and C_J indicate the gate and junction capacitance per unit width [3]. By minimizing C_B at constant R_B, the optimal width ratio of transistors $\alpha = W_P/W_N$ is determined as

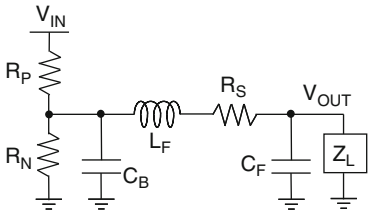


Fig. 3 Circuit model of a simple buck converter including major power-loss components

$$\alpha = \sqrt{\frac{DR_P C_N}{(1-D)R_N C_P}}. \tag{3}$$

By using the total width $W_{TOTAL} = W_P + W_N$, R_B and C_B are written as

$$R_B = \frac{R_0}{W_{TOTAL}};\ R_0 = (1+\alpha)\left[\frac{DR_P}{\alpha} + (1-D)R_N\right], \tag{4}$$

$$C_B = W_{TOTAL} C_0; \quad C_0 = \frac{\alpha C_P + C_N}{1+\alpha}. \tag{5}$$

There is a current ripple I_R in the output filter inductor.

$$I_R = \frac{V_{IN} D(1-D)}{2fL_F}. \tag{6}$$

Here, the maximum and minimum inductor currents are $I_L + I_R$ and $I_L - I_R$ in a continuous mode, respectively. From formula (6), effective inductor resistance R_I and effective current Irms are described as follows.

$$R_I = \frac{V_{IN} D(1-D)}{2f\tau_L I_R}. \tag{7}$$

$$I_{rms}^2 = \left(I_L^2 + \frac{I_R^2}{3}\right). \tag{8}$$

Here, τ_L is the figure of merit described as follows when RS represents series resistance of the filter inductor.

$$\tau_L = \frac{L_F}{R_S}. \tag{9}$$

The total major power loss of a buck converter is written as follows when P_{CAP}, P_{RES}, and P_{IND} indicate capacitive loss and resistive loss of switching transistors, and resistive loss of filter inductor respectively.

$$P_{LOSS} = P_{CAP} + P_{RES} + P_{IND}. \tag{10}$$

P_{CAP}, P_{RES}, P_{IND} are written as formulas (11)–(13).

$$P_{CAP} = C_B V_{IN}^2 f = W_{TOTAL} C_0 V_{IN}^2 f. \tag{11}$$

$$P_{RES} = R_B I_{rms}^2 = \frac{R_0}{W_{TOTAL}} \left(I_L^2 + \frac{I_R^2}{3} \right). \tag{12}$$

$$P_{IND} = R_I I_{rms}^2 = \frac{V_{IN} D(1-D)}{2f\tau_L I_R} \left(I_L^2 + \frac{I_R^2}{3} \right). \tag{13}$$

The total power efficiency is defined as follows.

$$\eta = \frac{P_{OUT}}{P_{OUT} + P_{LOSS}} = \frac{V_{IN} D I_L}{V_{IN} D I_L + P_{LOSS}} = \frac{1}{1 + P_{LOSS}/V_{IN} D I_L}. \tag{14}$$

To maximize the power efficiency, $P_{LOSS}/V_{IN}DI_L$ must be minimized. $P_{LOSS}/V_{IN}DI_L$ is minimized when $P_{CAP} = P_{RES} = P_{IND}$. As a result, optimal f and W_{TOTAL} are derived as follows.

$$f = \frac{\sqrt[3]{C_0^2 D^2 (3-3D)^2 (I_R^6 + 3I_L^6)}}{3C_0 \sqrt[3]{R_0 I_R^2 \tau_L^2 (2I_R^2 + 6I_L^2)^2}}. \tag{15}$$

$$W_{TOTAL} = \frac{\sqrt[3]{I_R R_0^2 \tau_L (2I_R^2 + 6I_L^2)}}{V_{IN} \sqrt[3]{3C_0 D(1-D)}}. \tag{16}$$

By substituting formulas (15) and (16), $P_{LOSS}/V_{IN}DI_L$ is simplified as follows.

$$\frac{P_{LOSS}}{V_{IN} D I_L} = \frac{\sqrt[3]{3R_0 C_0 (1-D)(I_R^2 + 3I_L^2)^2}}{I_L \sqrt[3]{2\tau_L I_R D^2}}. \tag{17}$$

By differentiating formula (17) with respect to I_R, following conditions are derived to minimize $P_{LOSS}/V_{IN}DI_L$.

$$I_R = I_L, \tag{18}$$

$$W_{TOTAL} = R_0 \frac{I_L}{V_{IN}} \sqrt[3]{\frac{8\tau_L}{3R_0 C_0 D(1-D)}}, \tag{19}$$

$$f = \sqrt[3]{\frac{D^2 (1-D)^2}{3R_0 C_0 \tau_L^2}}. \tag{20}$$

A buck converter can be designed optimally by employing formulas above, when performances of transistors and inductor, load current I_L, and duty cycle D

Table 1 Optimized parameters in two cases

130 nm CMOS VIN = 1.2 V, Tl = 100 ns	WP (mm)	WN (mm)	L (nH)	f (MHz)
(I) D = 0.3, IL = 10 mA	0.5	0.4	180	70
(II) D = 0.7, IL = 100 mA	7.8	2.8	20	63

are fixed. The required output filter capacitance C_F is calculated by formula (21) when Vripple denotes output voltage ripple.

$$C_F = \frac{(1-D)V_{OUT}}{16L_F f^2 V_{ripple}}. \tag{21}$$

To verify the accuracy of the circuit model and the optimization theory, two cases of simulations (I) D = 0.3, I_L = 10 mA, and (II) D = 0.7, I_L = 100 mA were performed in 130 nm CMOS. In both cases, values of C_{gtot} and C_{db} in 130 nm CMOS spice model were simply applied to C_G and C_J. Here, the transistor parameters were calculated as R_P = 1,700 Ω/μm, R_N = 440 Ω/μm, C_P = 2.930 fF/μm, and C_N = 2.362 fF/μm. By using the parameters, theoretical power efficiency optimizations were performed as shown in Table 1.

Figures 4 and 5 show the simulation results of power efficiency dependence on (a) filter inductance L_F, (b) switching frequency f, (c) total transistor width W_{TOTAL}, and (d) width ratio of transistors α, with theoretical optimization results in case (I) and (II) respectively. In the simulations, parameters except swept one are set to theoretically optimal values. As will be appreciated from the results, the accuracy of the circuit model and the optimization theory are high enough for a simple buck converter. Furthermore, power efficiency dependence on α is relatively small compared with other parameters.

2.2 *Si-CMOS + Si-CMOS Implementation*

When the buck converter is designed optimally, the total power loss is simplified as follows.

$$P_{loss} = V_{in} I_L \sqrt[3]{24\frac{R_0 C_0}{\tau_L} D(1-D)} \propto \sqrt[3]{\frac{R_0 C_0}{\tau_L}}, \tag{22}$$

where

$$R_0 C_0 = \left(\sqrt{DR_P C_P} + \sqrt{(1-D)R_N C_N}\right)^2. \tag{23}$$

As seen from (22), the smaller R_0C_0 and the larger τ_L are better for higher power efficiency. R_0C_0 is shown to be roughly proportional to the product of effective

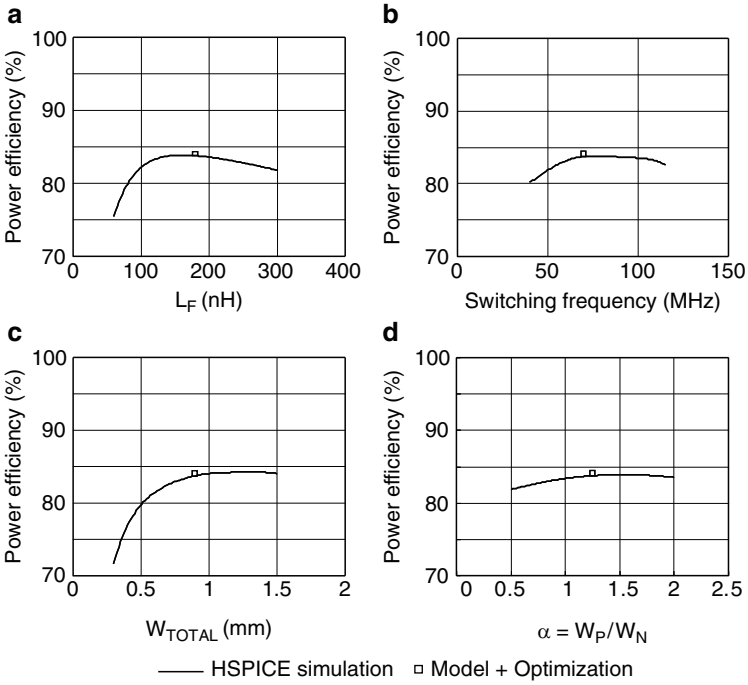


Fig. 4 Power efficiency dependence on (**a**) inductance, (**b**) switching frequency, (**c**) total transistor width, and (**d**) width ratio of transistors in case (I)

conduction resistance R_T and effective switched capacitance C_{eff} per unit gate width of switching transistors. Here, C_{eff} is expressed as $1.3C_G + C_J$ where C_G and C_J indicate the gate and junction capacitance per unit width [3]. As technology scales, R_T and C_{eff} scale as $1/k^{0.7}$ and $1/k$ where k denotes scaling factor [4]. k doubles when technology scales down by half. Therefore, it is better from the power efficiency point of view to use the more advanced technology.

One may argue, however, that the maximum V_{IN} can be lower in case of transistors with the smallest line width. When V_{IN} is higher than the maximum VDD for a certain technology, the switching transistors must be cascoded to relax the voltage over-stress. If two transistors are cascoded, R_T increases by a factor of 1.2 and C_{eff} decreases by a factor of 0.4 [3]. As a result, even though we have to use cascaded structure, R_0C_0 decreases as technology scales. Thus, it can be said that it would be better to use the most scaled transistors for high conversion efficiency.

Figure 6 shows the calculation results of maximum power efficiencies dependence on duty cycle (voltage conversion ratio) and process technology when $\tau_L = 100$ ns. The maximum power efficiencies are independent of load current I_L and input voltage V_{IN}.

Table 2 shows the transistor parameters of 350, 180, 130 and 90 nm CMOS process technologies. R_P, R_N, C_P and C_N basically decrease as the technology scales

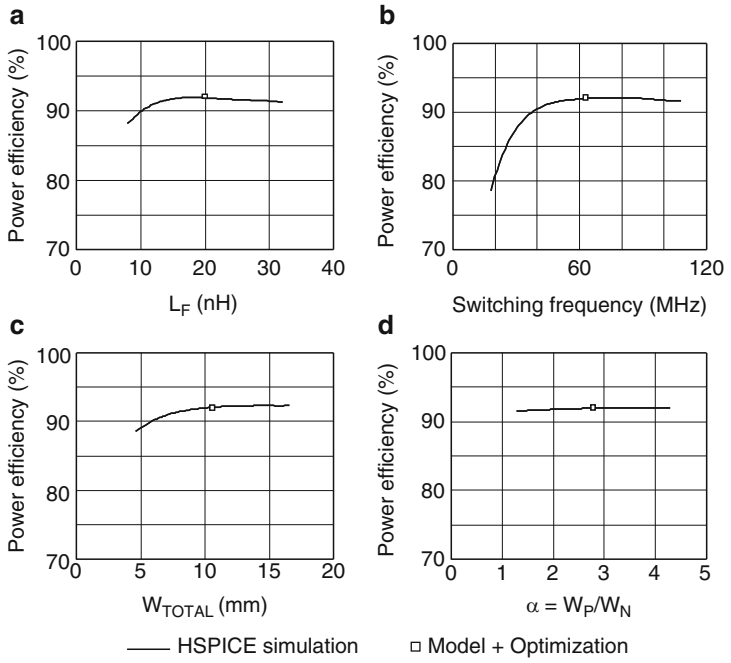


Fig. 5 Power efficiency dependence on (**a**) inductance, (**b**) switching frequency, (**c**) total transistor width, and (**d**) width ratio of transistors in case (II)

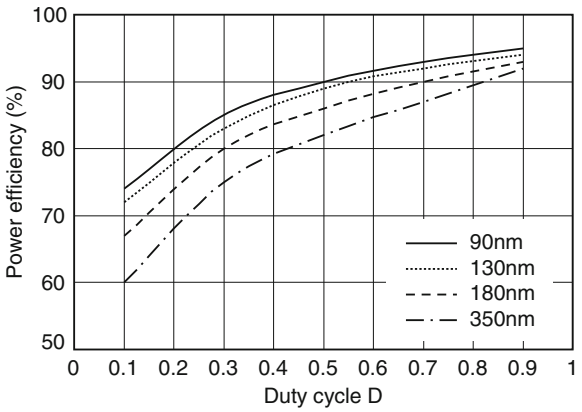


Fig. 6 Maximum power efficiency dependence on D and process technology

down. As shown in Fig. 6, the smaller the process technology is, the higher power efficiency a buck converter performs when the value of τ_L is fixed. Especially in the range in which D is small, smaller process technology is advantageous.

On the other hand, τL is mainly determined by the thickness of the metal wire as described in (9) and its dependence on process technology is small. Thus the inductor is not necessarily fabricated by using the most advanced technology which is expensive. In addition to the inductor, fabrication cost of on-chip MOS capacitor per unit capacitance increases as technology scales. Thus it is reasonable to implement active elements and output filter on separate die whose process technologies are different.

Figure 7 shows the basic concept of the stacked-chip implementation of a buck converter. The lower chip fabricated in the advanced technology contains the controller and switching transistors of a buck converter and target circuits. The upper chip fabricated in conventional and cheap process technology contains LC filter elements such as L's and C's. By stacking two chips face-to-face and connecting them via metal bumps, a buck converter for on-chip distributed power supply systems can be fabricated in a well balanced manner for best cost and power tradeoff.

Table 2 Transistor parameters of 350, 180, 130 and 90 nm

Process technology (nm)	VDD (V)	RP (Ω/μm)	RN (Ω/μm)	CP (fF/μm)	CN (fF/μm)
350	3.3	5,215	1,870	3.434	3.319
180	1.8	3,130	915	2.760	2.576
130	1.2	1,700	440	2.930	2.362
90	1.0	1,776	426	2.020	2.079

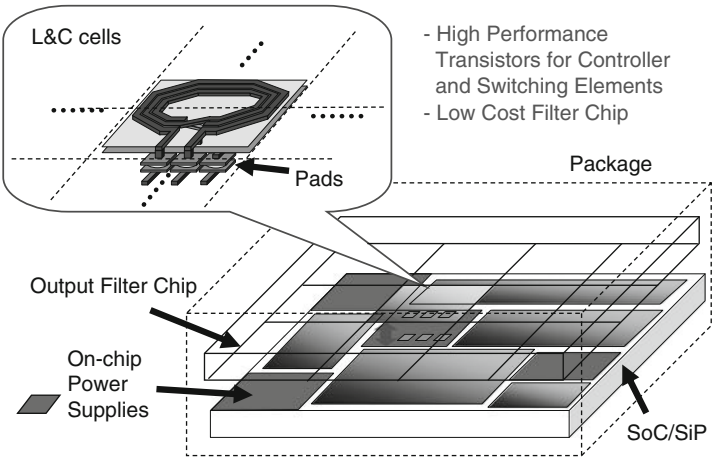


Fig. 7 Concept of the stacked-chip implementation

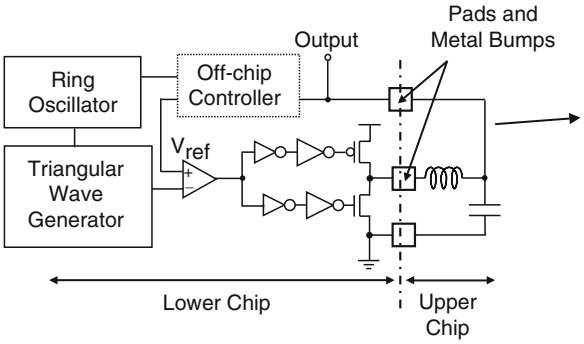


Fig. 8 Diagram of the test chip for stacked implementation

To demonstrate the feasibility of the stacked-chip buck converter, an on-chip buck converter is designed in 0.35-μm CMOS for upper and lower chips. The lower chip could be manufactured by 90-nm or more advanced technology for the higher efficiency but this test chip is to show the feasibility of the stacked-chip approach. Figure 8 shows the circuit diagram of the buck converter. Parameters are optimized for the power efficiency as described here in below.

Drivabilities of the tapered buffers are set high enough and the inverter sizes are calculated to minimize the on-on overlap time of switching transistors. The outer diameter of the filter equals to that of the filter inductor d_{OUT}, which is set at 2 × 2 mm by assuming that 10 mm-square chip can have 25 voltage domains. 6.8 × 6.9 mm chip with seven voltage domains has already been presented [5]. τ_L degrades as the outer diameter of the inductor shrinks, however, the power efficiency can be kept high because R_0C_0 decreases as technology scales. The inductance is estimated by a simple formula from [6]. Narrow metal-to-metal spacing and wide metal wire are preferable in this application. τ_L is a function of d_{IN}/d_{OUT} ratio, which can be calculated using the inductance formula and the sheet resistance. The inductance and the parasitic resistance are roughly proportional to n2 for a fixed d_{IN}/d_{OUT} ratio, n being the number of turns, considering that the space is negligibly narrow compared with the line width. The normalized τ_L curve in Fig. 9 is therefore independent of n. From the calculation result, d_{IN} is decided to be about $0.5d_{OUT}$.

Here, n is decided as 3 to maximize the power efficiency. As a result, the calculated inductance of this work is 22nH when the sheet resistance is about 0.02 Ω/□. The open space at the center of the inductor is filled with a MOS capacitor for the output filter. Area efficiency is more important than linearity for the filter capacitor, because the output voltage does not change dynamically in a normal operation. From that aspect, MOS capacitor is more suitable than any other types of on-chip capacitors like Metal-Insulator-Metal (MIM) capacitor or poly-silicon capacitor. The obtained capacitance is about 1 nF. Under those conditions, the power efficiency dependence on the output current and the switching frequency for $V_{IN} = 3.3$ V and $V_{OUT} = 2.3$ V is plotted in Fig. 10 by modifying the duty cycle definition in [7–11

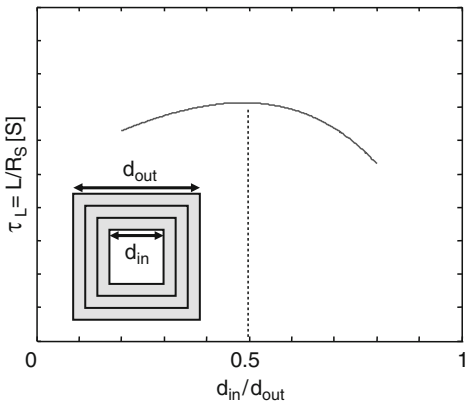


Fig. 9 τL dependence on dIN/dOUT for square inductor

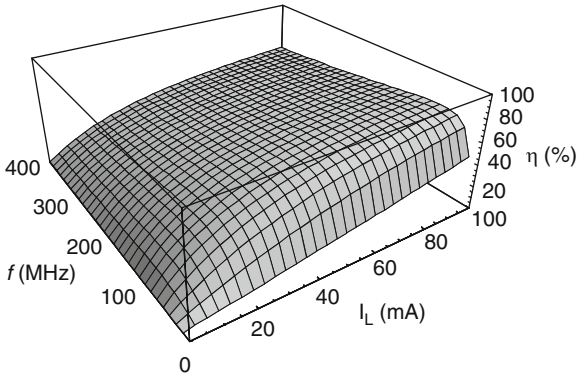


Fig. 10 Estimated power efficiency dependence on load current and switching frequency

D is redefined in this work as follows to take the voltage drop caused by the filter inductor into account, when it is simply set as $D = V_{OUT}/V_{IN}$ in [6–10]. By doing so, the estimation accuracy of several values is improved.

$$D = \frac{V_{OUT} + R_S I_L}{V_{IN}}. \tag{24}$$

Here, I_L denotes the load current which equals to the DC part of the inductor current.

The output voltage ripple ratio is described as follows when f and C_F denote switching frequency and output filter capacitance, respectively.

$$\frac{V_{ripple}}{V_{OUT}} = \frac{1 - D}{16 L_F C_F f^2}. \tag{25}$$

It is impossible to choose the switching frequency under 100 MHz in this case because the output voltage ripple goes up above 10% for the chosen values. The gate width of the high-side and the low-side transistors are designed to be 1,000 μm and 500 μm under the load current condition of 60 mA using the optimal gate width formulas in [6–10].

The test buck converter with the stacked-chip implementation was fabricated and measured. Figure 11 shows the chip microphotograph of the output filter on the upper chip.

Figures 12 and 13 show the measurement setup and its cross-sectional diagram. The pad size and the effective bump diameter of this experimental setup are 200 × 200 μm and 150 μm, respectively. Micro bumps whose diameter is 30 μm and whose resistance is as low as 14 mΩ/bump have been realized in industry environments [11] and can be used instead for further smaller area.

The output waveform in Fig. 14 shows $V_{OUT} = 1.86$ V at $I_L = 60$ mA. The measured voltage ripple is smaller than ±10%, which is comparable to the result of the more expensive solution in [12]. Figure 15 shows the simulated and measured

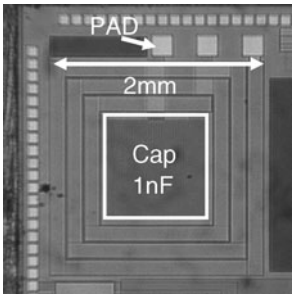


Fig. 11 Chip microphotograph of output filter

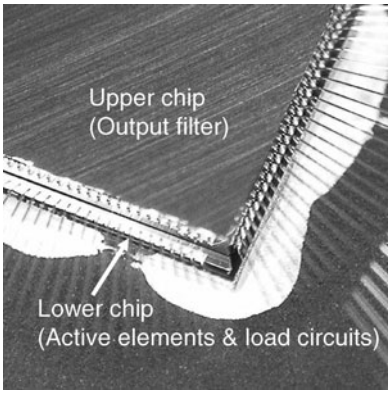


Fig. 12 Measurement setup of test chip

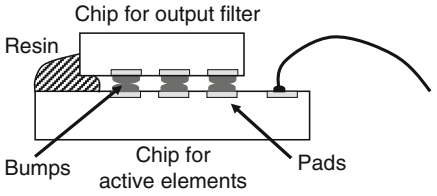


Fig. 13 Cross section diagram of measurement setup

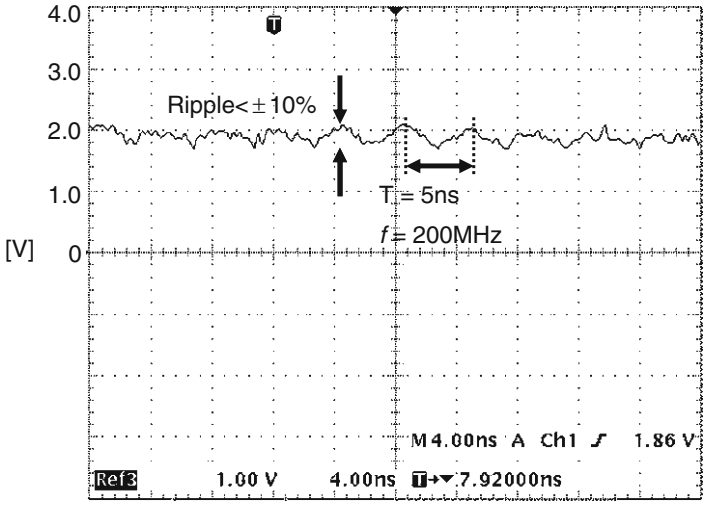


Fig. 14 Voltage waveform for VOUT = 1.86 V and IL = 60 mA

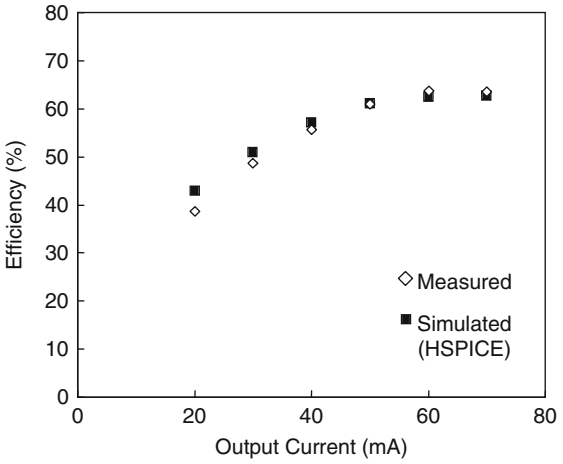


Fig. 15 Simulated and measured power efficiency for VOUT = 2.3 V and f = 200 MHz

power efficiency with $V_{OUT} = 2.3$ V for an output current range from 20 to 70 mA. The maximum efficiency of 62% is achieved for 70 mA output current. The measurement results compare well with the HSPICE simulation results. The simulation considers all the parasitic elements including the inductor parasitic resistance of 2.5 Ω, and the inductor input-to-ground capacitance of 25 pF.

2.3 Si-CMOS + Interposer Implementation

In order to further increase the efficiency, it is effective to use inductors whose τ_L is higher than that of the previous inductors. A thin-film inductor surrounded by magnetic core material as proposed in [13] can be a solution but is expensive. Implementing the inductor on a glass epoxy interposer as shown in Fig. 16 is an effective yet inexpensive solution.

The thickness of the metal wire on an interposer is generally thicker than 10 μm and the inductor on the interposer shows a much lower resistance than that on a silicon chip where the metal thickness is usually less than 1 μm. A capacitor on the upper chip made with a conventional technology is connected to the lower chip manufactured by an advanced technology like 90-nm CMOS, through the through-hole vias in the interposer. By this implementation, high efficiency and low cost are achieved at the same time. The maximum power efficiency is derived as follows from formula (14).

$$\eta = \frac{P_{OUT}}{P_{OUT} + P_{LOSS}} = \frac{V_{IN} D I_L}{V_{IN} D I_L + V_{IN} I_L \sqrt[3]{24 \frac{R_0 C_0}{\tau_L} D(1-D)}} = \frac{1}{1 + \alpha {\tau_L}^{-1/3}}. \quad (26)$$

Here, P_{OUT} indicates the output power. α is a function of R_0C_0 and D, and can be a constant when the process technology and D are fixed. Therefore, simply higher τ_L directly results in higher power efficiency. Figure 17 shows the calculation results of

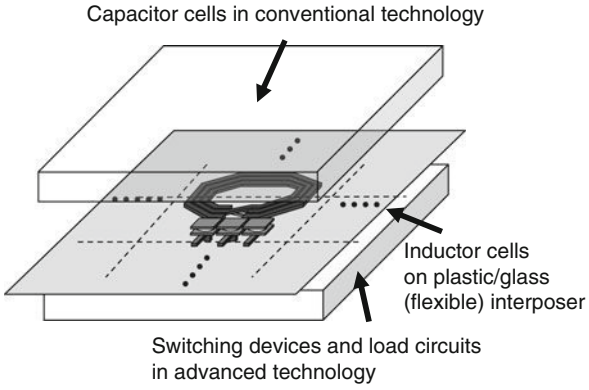


Fig. 16 Another stacked-chip implementation to gain high τL

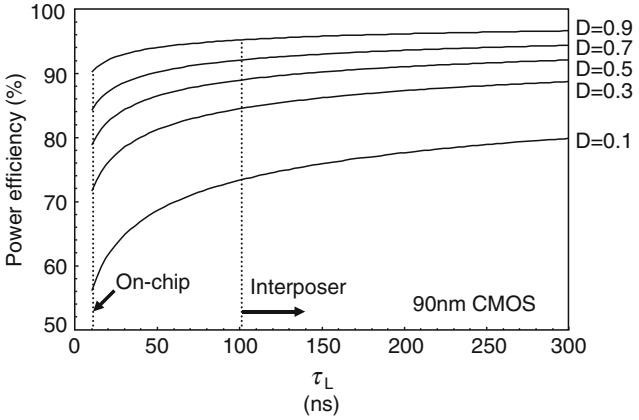


Fig. 17 Power efficiency dependence on τL in several duty cycles

the maximum power efficiency dependence on τ_L in several duty cycles. Especially when D is small, it is valuable to use an inductor whose τL is large. Moreover, the power efficiency simply increases as τ_L increases for any values of D.

Generally the value of τ_L is larger than 100 ns in case of inductors implemented on an interposer, when it is around 10 ns in case of on-chip inductors.

The structure shown in Fig. 16 is assembled using a newly introduced interposer and the same lower and upper chips presented in the first half of this section. In this implementation, only a capacitor is used on the upper chip. Figure 18 shows an inductor array on generic Flame Resistant 4 (FR-4) glass epoxy interposer with two metal layers. The circled inductor in the array, which achieved the minimum metal spacing in the trial manufacture, is used for the measurement. The metal thickness on the interposer is 30 μm, the substrate thickness is 100 μm, and the diameter of the through-hole via is 100 μm. This implementation increases τ_L by 30 times compared with the case of an on-chip inductor.The outer diameter of the inductor is increased by 10% to achieve the same value of on-chip inductance because the minimum spacing of metal lands on glass epoxy is larger than that of on-chip interconnects. The permittivity of the glass epoxy is generally more than four times higher than SiO2, however, the parasitic capacitance between both sides of the interposer can be negligible.

That is because the substrate thickness is large enough compared with the line width. S-parameters of the fabricated inductor were measured using a test element group and a network analyzer. Figure 19 shows the equivalent model to extract the inductance and the parasitic series resistance of the inductor from the measured S-parameters. Figure 20 shows the extracted characteristics of the inductor. Measured inductance and τL were 18 nH and 100 ns at 200 MHz, respectively. Since the characteristics under 200 MHz are important in this application, the estimation results of the parameters are sufficiently high. In the high frequency region over 200 MHz, the parasitic resistance increases rapidly because of skin effect. The skin depth of the metal wire at 500 MHz is assumed to be smaller than 3 μm.

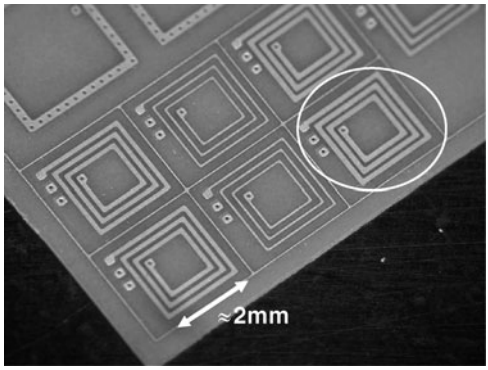


Fig. 18 Manufactured inductor array on glass epoxy interposer

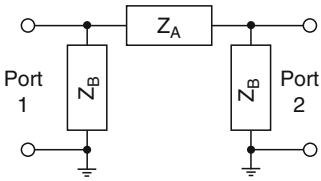


Fig. 19 Equivalent model for inductance and resistance extraction

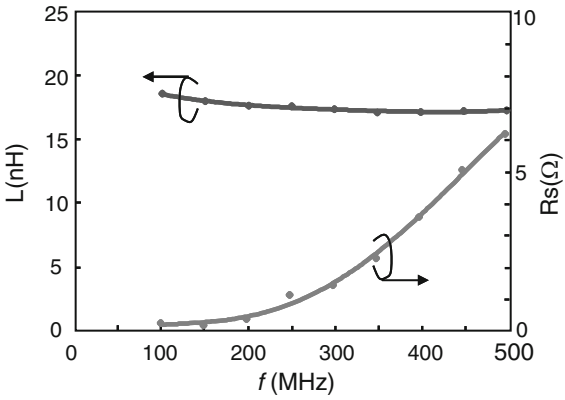


Fig. 20 Measured inductance and resistance of fabricated inductor

The inductance decreases gradually as frequency increases, because of the current concentration caused by skin effect.

Figure 21 shows the cross section of the stacked chips. Figure 22 shows the comparison of measured power efficiency between two types of implementations of "two chips" and "two chips + interposer" for $V_{IN} = 3.3$ V and $V_{OUT} = 2.3$ V.

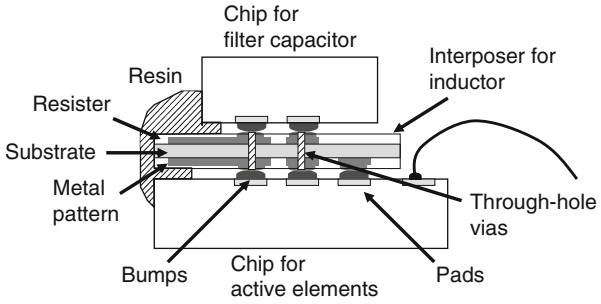
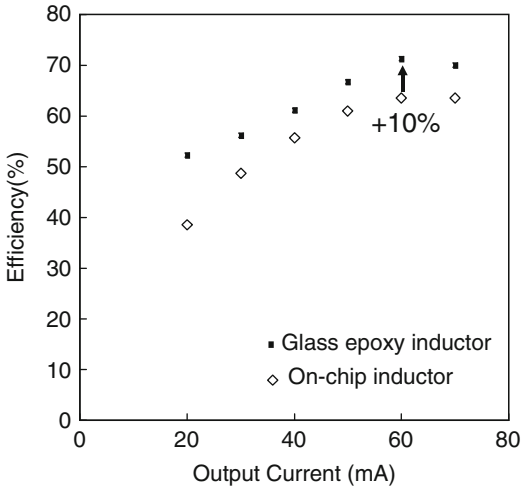


Fig. 21 Cross-sectional diagram of measurement setup

Fig. 22 Efficiency comparison between two types of implementations



The power efficiency with the glass epoxy inductor is improved by 5–14% depending on the output current compared with the on-chip implementation. The maximum power efficiency of 71.3% is achieved at an output current of 60 mA. The possible reason that the efficiency does not improve the most at 60 mA, is that the switching transistors are not changed optimally according to the τ_L characteristic of newly implemented inductor.

2.4 W-CSP Implementation

Wafer level chip size package (W-CSP) technology has become one of the popular packaging technologies for compact and cost-effective implementation in recent years. They are usually used for portable equipments.

In general, W-CSP technology provides additional metal lines as a post-fabrication process over conventional silicon LSI's. The metal lines are thicker but whose processing accuracy is lower than those of silicon LSI's. Filter inductors for a buck converter do not require the processing accuracy but prefer thicker metal lines as described in previous sections. W-CSP technology therefore comes to one of the good candidates for filter inductor implementation. Two types of buck converters shown in Fig. 23 were designed for power efficiency comparison between on-chip implementation and W-CSP implementation of the filter inductor. Components other than the output filter inductor are completely the same. Several design values where chosen to the described values according the optimization theory when the outer diameter of the filter inductor was set to 2 mm. The two types of chips were fabricated in 0.15 μm SOI-CMOS process and whose microphotographs are shown in Fig. 24. Values of τ_L are 10 and 500 ns for on-chip and W-CSP inductors respectively.

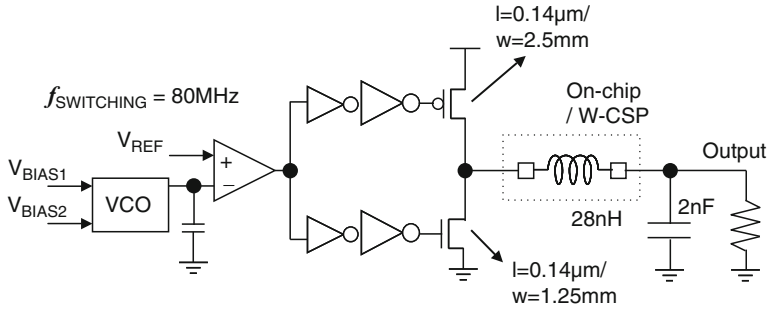


Fig. 23 Test circuit for efficiency comparison of Si-CMOS and W-CSP implementation

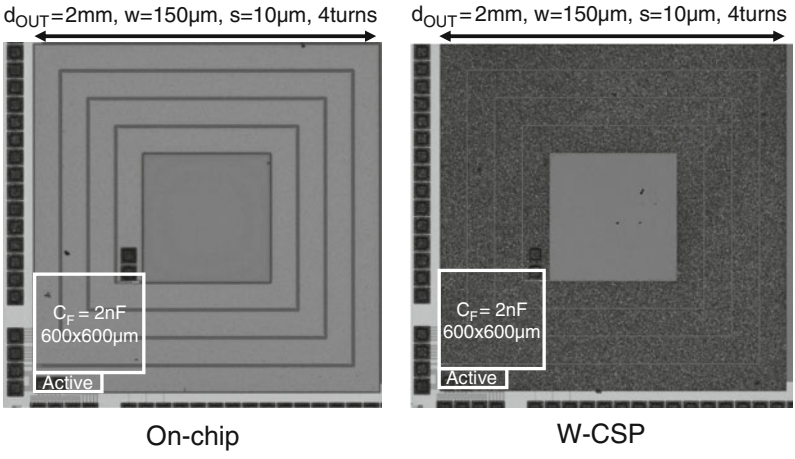


Fig. 24 Chip microphotographs

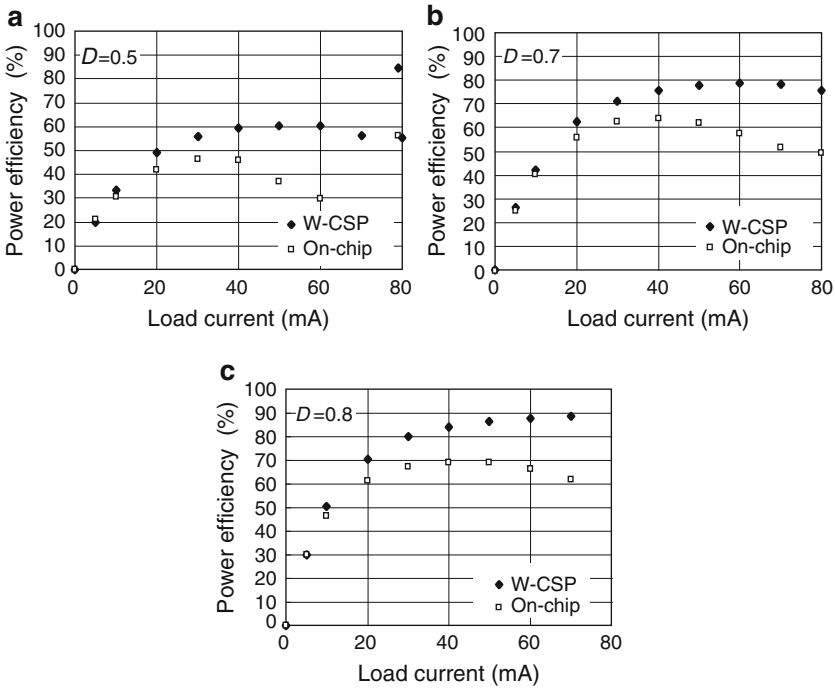


Fig. 25 Comparison results of measured power efficiency for D = **(a)** 0.5, **(b)** 0.7 and **(c)** 0.8

Figure 25 shows the measured results of power efficiency for D = 0.5, 0.7 and 0.8 when load currents are swept from 0 to 80 mA. VDD was fixed to 1.0 V which is the standard value of the process. The maximum measured power efficiencies are 61%, 79% and 89% for the W-CSP implementation, while 46%, 62% and 71% for the on-chip implementation. The W-CSP implementation improves the power efficiency by more than 10% in average, compared with the conventional on-chip type implementation.

3 Hybrid Operation of Linear Regulator and Buck Converter

3.1 Quick Wake-Up and Power Gating

Efficient voltage regulation with fine-grained adaptive supply voltage control and integrated power gating are essential components for power-constrained, high-performance microprocessors and digital signal processors. Distributed on-chip fast-transient DC-DC converters are required for fine-grained power management of individual block voltage control [14]. A hybrid buck converter with a linear transient accelerator architecture is proposed for fast wake-up and high power efficiency.

Figure 26 shows the proposed on-chip hybrid DC–DC converter architecture with a buck converter, linear transient accelerator, and stable transition enabler realizing smooth transitions among buck, linear, and hybrid regulation modes. High-speed voltage transition is achieved with the linear transient accelerator, while improved steady-state power efficiency is achieved by the buck converter. The voltage transient time of a single buck converter is larger than a linear transient accelerator due to the time constant limitation and the ringing issue of the LC filter.

Figure 27 describes the operation mode sequence of the proposed on-chip DC-DC converter. In the stand-by mode, two PMOS transistors M1 and M2 in Fig. 26 act as leakage cut-off switches. Since M1 and M2 are made with I/O transistors which have high V_{TH} and thick gate oxide, stand-by leakage is completely cut-off without power-gating switch inserted in series to the load circuit block. The proposed leakage cut-off mechanism which is integrated in the power supply unit solves the issues of the conventional power gating approach such as extra delay and area overhead. When a wake-up signal is issued, the linear transient accelerator alone is activated and transistor M1 supplies the load current. Once the output voltage is settled, the buck converter starts to operate as a hybrid mode. Then, the linear regulator starts to gradually fade out while the buck converter is getting to operate in full power. Finally the buck converter alone operates to achieve high efficiency. Circuit stability in the hybrid mode can be guaranteed by proper design, which was shown in [15], although the purpose of [14] is to cancel a ripple of a buck converter with a linear regulator and it does not act as a transient accelerator being

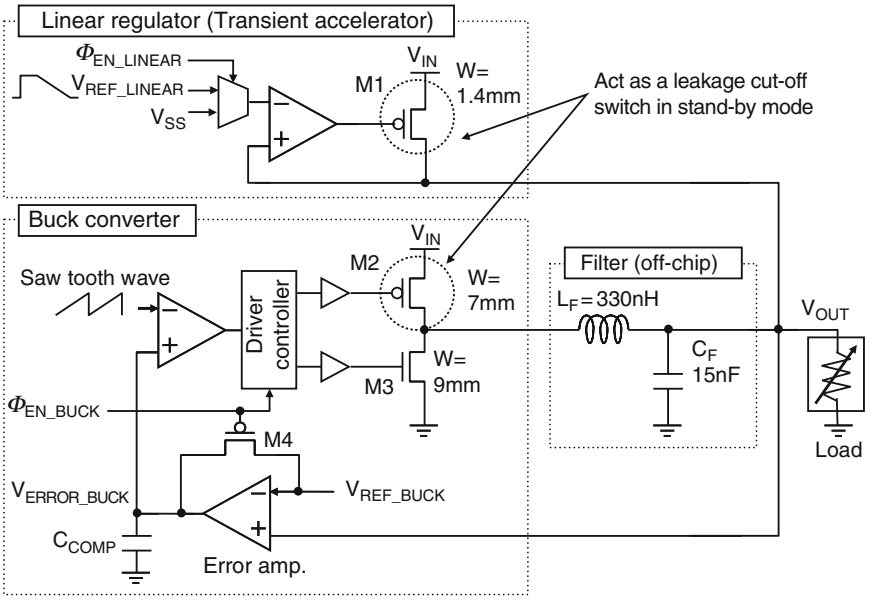


Fig. 26 Proposed on-chip DC-DC converter realizing stable transition among linear, hybrid and buck modes

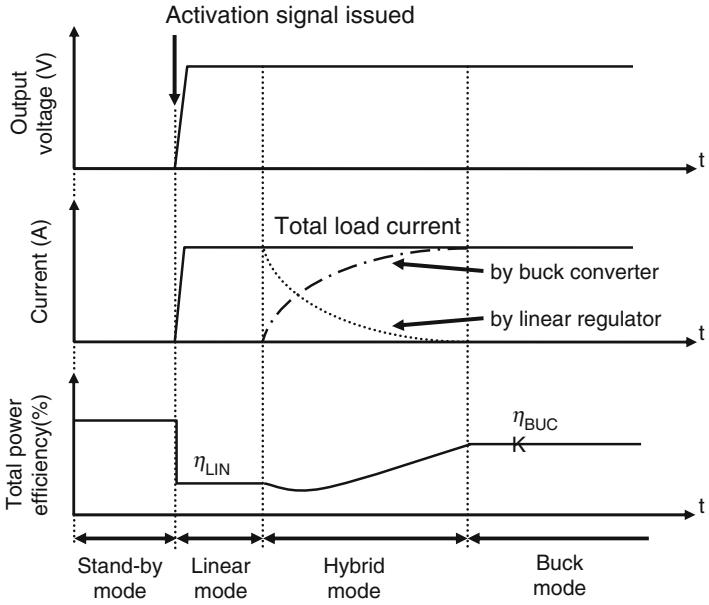


Fig. 27 Operation mode sequence of the proposed on-chip DC-DC converter

different from this paper. Refs [16–19] also describe hybrid systems of a linear regulator and a buck converter to achieve higher efficiency over wide range of load current level, but does not use the linear regulator as an accelerator.

Figure 28 shows the simulation results of mode transitions for the load current I_{LOAD} of 6 and 60 mA with and without the proper buck converter starting technique proposed in this paper. In this paper, an additional transistor M4 is placed to set the output of the error amplifier at V_{ERROR_BUCK} at the starting phase of the buck converter. By doing so, the starting duty ratio of the buck converter is fixed at the level when the output current equals to zero because the linear regulator is supplying the whole current at the start. Without M4 and this special care, voltage ringing as high as 1.3 V occurs as shown in Fig. 28a, which impacts negatively to the quick mode transition and the reliability of the load circuits. When the linear converter fades out, if the linear regulator dies very rapidly by decreasing V_{REF_LINEAR} at the rate of -1 V/ns as shown in Fig. 28c, a large voltage droop of 250 mV will be generated. This droop can be suppressed to 30 mV by gradually turning off the linear regulator by decreasing V_{REF_LINEAR} at the rate of -2×10^{-5} V/ns as shown in Fig. 28d. Since the slower transition decreases the droop monotonically, the slope of V_{REF_LINEAR} can be designed taking trade-off of the droop and total power efficiency. Longer transition consumes more power by the linear regulator. Thus design care should be taken at the start of the buck converter operation and the stopping of the linear regulator using the methodology proposed above.

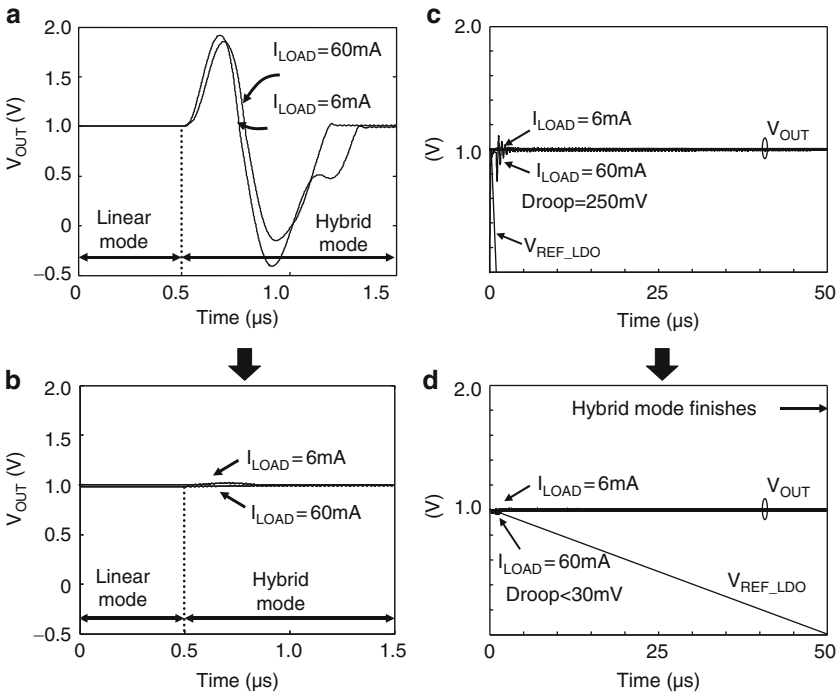


Fig. 28 Simulation results on transitions from linear mode to hybrid mode, and hybrid mode to buck mode

3.2 Measured Results

Figure 29 shows the measured waveforms in mode transition from stand-by mode to linear mode. In the stand-by mode, two orders of magnitude smaller leakage current is expected compared with the conventional power-gating due to the elimination of gate tunneling leakage at 65 nm technology by using I/O transistors. The measured leakage current in the design was 100 μA range. I/O transistors are necessary due to higer input voltage. The voltage transition from $V_{DDL} = 0$ V to $V_{DDL} = 1.0$ V is measured to be 80 ns. Without the linear accelerator, the buck converter alone shows the wake-up time more than 400 ns. Thus more than five times acceleration is achieved. As shown in the right side of Fig. 29, the controller logic of the buck converter turns both of the switching PMOS and NMOS off in the standby mode and in the linear mode. When the linear regulator turns on, the wake-up time in the linear mode is determined by the accelerator size and the load capacitor. As soon as the output voltage is settled in the linear mode, the buck converter can start the operation.

Figure 30 shows the measurement results of the transition from the linear mode to the hybrid operation mode in two different time scales. Stable mode transitions without ringing issue as shown in Fig. 28 are observed in both cases.

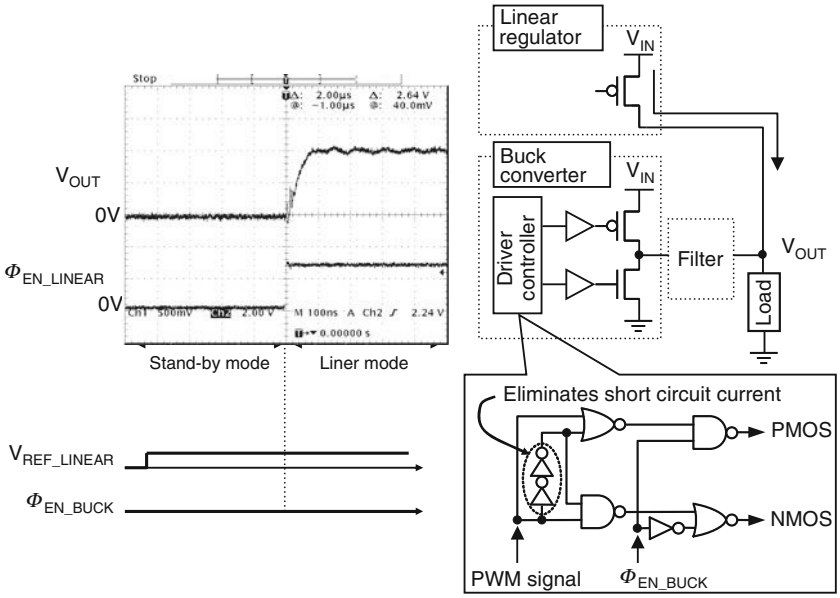


Fig. 29 Transition from stand-by mode to linear mode

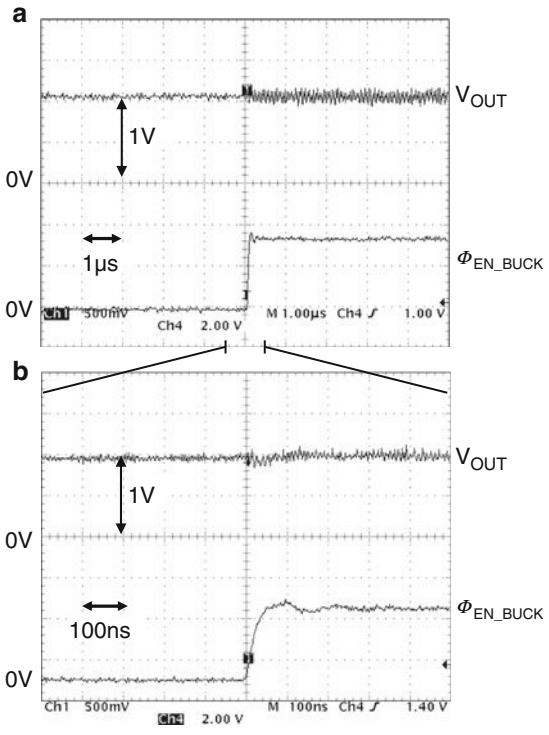


Fig. 30 Transition from linear mode to hybrid mode

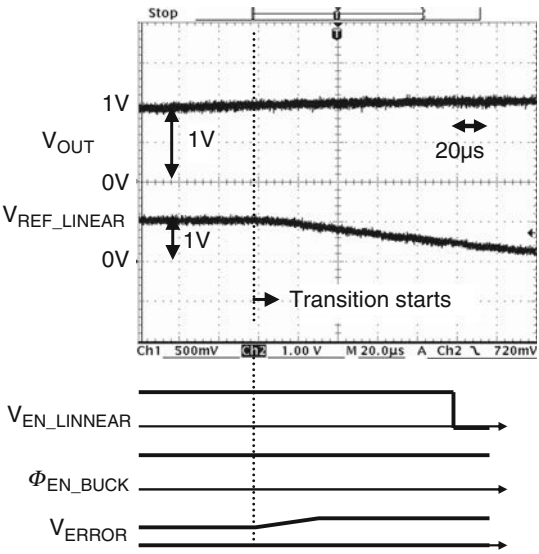


Fig. 31 Transition from hybrid mode to buck mode

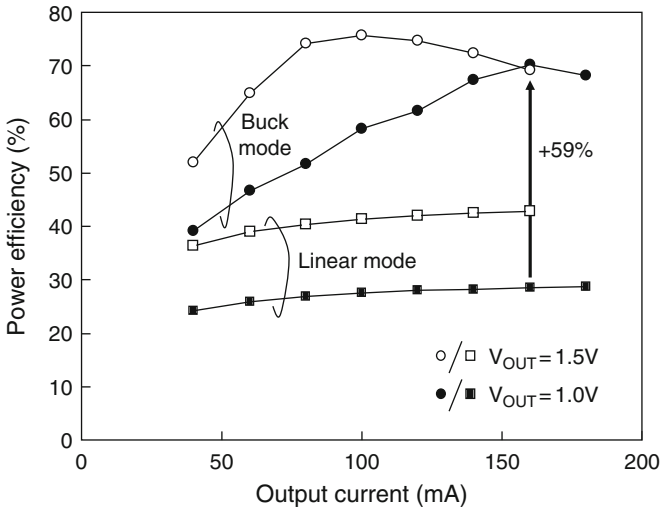


Fig. 32 Measured power efficiency in linear and buck modes

Figure 31 shows the measurement results of the transition from the hybrid mode to the single operation mode of the buck converter for load current of 60 mA where the linear regulator is gradually fade out. In the measurement, V_{REF_LINEAR} was decreased with the slope of -8.3×10^{-6} V/s even slower than the simulation in Fig. 28. By using the slow transition, no visible voltage droop is observed.

Figure 32 shows the measured power efficiency in linear and buck modes for $V_{OUT} = 1$ V and 1.5 V. 59% higher stead-state power efficiency in buck mode is observed compared with the linear mode at $V_{OUT} = 1$ V and $I_{LOAD} = 160$ mA. Figure 33 shows the measurement setup of the proposed converter. An off-chip inductor and a capacitor were mounted on a circuit board. Power lines were connected by multiple bonding wires. Figure 34 shows the microphotograph of the fabricated converter. The layout area of the linear regulator was 100 × 150 μm while that of buck converter was 350 × 400 μm.

Table 3 summarizes the measured performance of the proposed converter.

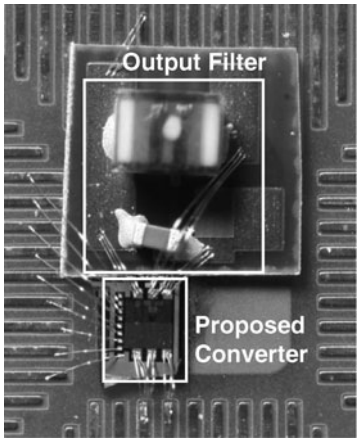


Fig. 33 Measurement setup

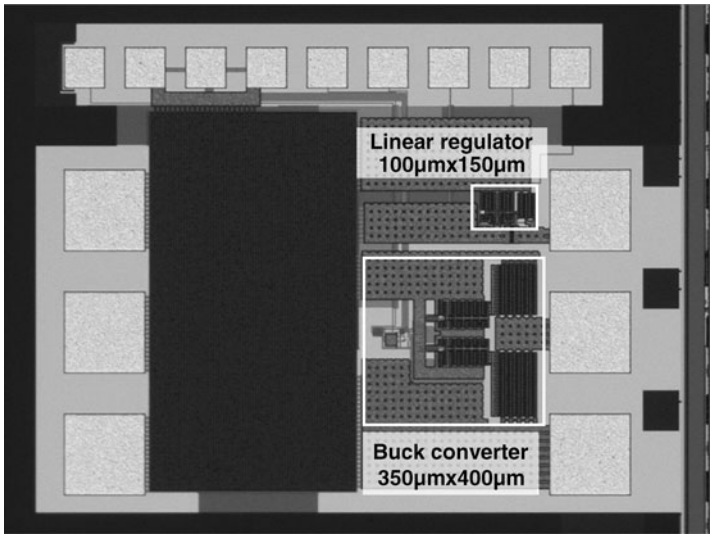


Fig. 34 Microphotograph of proposed converter

Table 3 Performance summary

Process	180-nm CMOS
Input voltage	2–3.3 V (HVDD, HVT transistors)
Output voltage transient time	80 ns (VOUT = 0 → 1 V, linear mode)
Switching frequency (buck mode)	50 MHz (LF = 330 nH, CF = 15 nF)
Power efficiency η	75.8% (buck mode) 41.3% (linear mode) 43.3% (hybrid mode) at VIN = 3.3 V, VOUT = 1.5 V, ILOAD = 100 mA
Area	100 × 150 μm (linear regulator) 350 × 400 μm (buck converter)

4 Conclusions

Two types of buck converters are presented in this paper to realize distributed power supply systems. The first one is the 3D-structured high-efficiency buck converter. The measured results show the maximum power efficiency of 71.3% with a planar filter inductor implemented on a FR-4 interposer. A filter inductor implemented by W-CSP technology also improves the power efficiency by more than 10% compared with an on-chip inductor. The second one is the buck converter with hybrid operation scheme to realize built-in power gating. The measured results show 80 ns fast wake-up time from standby-mode, which is 5× faster than the conventional buck converter. The standby current of 100 μA range and stable transients during converter mode transitions are measured as well.

Acknowledgement Authors would like to thank the chip fabrication program of VLSI Design and Education Center (VDEC), the University of Tokyo in collaboration with Rohm Corporation and Toppan Printing Corporation. This research was also partly supported by OKI Electric Industry.

References

1. The International Technology Roadmap for Semiconductor 2006 Update [Online]. Available: http://www.itrs.net/Links/2006Update/2006UpdateFinal.htm
2. H. Kawaguchi, K. Nose, T. Sakurai, A super cut-off CMOS (SCCMOS) scheme for 0.5-V supply voltage with picoampere stand-by current. *IEEE J Solid-State Circ.* **35**(10), 1498–1501 (2000)
3. G. Schrom, P. Hazucha, F. Paillet, D.S. Gardner, S.T. Moon, T. Karnik, Optimal design of monolithic integrated DC-DC converters, *IEEE International Conference on IC Design and Technology*, 2006, pp. 65–67
4. T. Sakurai, R. Newton, Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas. IEEE J. Solid-State Circ. **25**(2), 584–594 (1990)

5. G. Uvieghara, M.-C. Kuo, J. Arceo, J. Cheung, J. Lee, X. Niu, R. Sankuratri, M. Severson, O. Arias, Y. Chang, S. King, K.-C. Lai, Y. Tian, S. Varadarajan, J. Wang, K. Yen, L. Yuan, N. Chen, D. Hsu, D. Lisk, S. Khan, A. Fahim, C.-L. Wang, J. Dejaco, Z. Mansour, M. Sani, A highly-integrated 3G CDMA2000 1X cellular baseband chip with GSM/AMPS/GPS/bluetooth/multimedia capabilities and ZIF RF support, *IEEE International Solid-State Circuits Conference Digest of Technical Papers*, 2004, pp.422–536
6. S.S. Mohan, M. del Mar Hershenson, S.P. Boyd, T.H. Lee, Simple accurate expressions for planar spriral inductors. IEEE J. Solid-State Circ. **34**(10), 1419–1424 (1999)
7. V. Kursun, S.G. Narendra, V.K. De, E.G. Friedman, Efficiency analysis of a high frequency buck converter for on-chip integration with a dual-VDD microprocessor, *Proceedings of the European Solid-State Circuits Conference*, Sep 2002, pp. 743–746
8. V. Kursun, S.G. Narendra, V.K. De, E.G. Friedman, Monolithic DC-DC converter analysis and MOSFET gate voltage optimization, *Proceedings of the Fourth International Symposium on Quality Electronic Design*, Mar 2003
9. V. Kursun, S.G. Narendra, V.K. De, E.G. Friedman, Analysis of buck converters for on-chip integration with a dual supply voltage microprocessor, *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 11, no. 3, Jun 2003
10. V. Kursun, S.G. Narendra, V.K. De, E.G. Friedman, Low-voltage-swing monolithic dc-dc conversion, *IEEE Transactions on Circuit Systems*, vol. 51, no 5, May 2004
11. G. Schrom, P. Hazucha, J. Hahn, D.S. Gardner, B.A. Bloechel Greg Dermer, S.G. Narendra, T. Karnik, Vivek De, A 480-MHz, multi-phase interleaved buck DC-DC converter with hysteretic control, *2004 35th Annual IEEE Power Electronics Specialists Conference*, Jun 2004, pp. 4702–4707
12. T. Ezaki, K. Kondo, H. Ozaki, N. Sasaki, H. Yonernura, M. Kitano, S. Tanaka, T. Hirayarna, A 160Gb/s interface design configuration for multichip LSI, *IEEE International Solid-State Circuits Conference Digest of Technical Papers*, 2004, pp. 140–141
13. P. Hazucha, G. Schrom, J.-H. Hahn, B. Bloechel, P. Hack, G. Dermer, S. Narendra, D. Gardner, T. Karnik, De Vivek, S. Borker, A 233-MHz, 80%–87% efficient four-phase DC-DC converter utilizing air-core inductors on package. IEEE J. Solid-State Circ. **40**(4), 838–845 (2005)
14. G. Schrom, P. Hazucha, J.-H. Hahn, V. Kursun, D. Gardner, S. Narendra, T. Karnik, V. De, Feasibility of monolithic and 3D-stacked DC-DC converters for mircoprocessors in 90nm technology generation, *Proceedings of the 2004 International Symposium on Low Power Electronics and Design*, Aug 2004, pp. 263–268
15. K. Onizuka, H. Kawaguchi, M. Takamiya, T. Sakurai, VDD-hopping accelerators for on-chip power supply circuit to achieve nanosecond-order transient time, *JSSC* **41**(11), Nov 2006
16. A. Kapum, M. Milanovic, J. Korelic, Voltage ripple cancellation in buck converter based on hybrid structured connection, *IEEE Proc. EPE-PEMC*, Aug 2006, pp. 112–117
17. T.J. Barber Jr., S. Ho, P. Ferguson, Jr., Multi-mode CMOS low dropout voltage regulator for GSM handsets, *Symp. VLSI Circuits*, June 2002, pp. 284–287
18. G. Thiele, E. Bayer, Current-mode LDO with active dropout optimization, *IEEE Power Electronics Specialists Conference*, 2005, pp.1203–1208
19. J.T. Stauth, S.R. Sanders, Optimum biasing for parallel hybrid switcing-linear regulators. *IEEE Transactions on Power Electronics*, vol. 22, no. 5, Sept 2007

Sampled Analog Signal Processing: From Software-Defined to Software Radio

François Rivet, André Mariano, Yann Deval, Dominique Dallet, Jean-Baptiste Begueret, and Didier Belot

1 The Software Radio Concept

1.1 Introduction

Wireless communication systems are faced with the emergence of various standards dedicated to voice transmission, data transfer and localization. The past decade has seen a fast evolution regarding the standards. Data rates have been increased. Carrier frequencies are higher. Modulations are more complex. Considering these changes, conventional architectures cannot challenge the multimedia convergence in the case of mobile terminals. Thus, new architectures are to be studied in order to respond to mobile terminal constraints. This chapter presents an overview of new solutions to overcome technological matters.

1.2 Definition and History

The wireless industry is looking for new RF architectures. The concept of Software Radio (SR) is part of the solution. It aims at designing a reconfigurable radio architecture accepting all cellular and non-cellular standards working in a 0–5 GHz frequency range.

Our researches are focused on commercial handsets. Software Radio brings flexibility and adaptability. Many gains are expected by telecommunication industry:

F. Rivet (✉)
François RIVET IC Design team - IMS Lab - France,
e-mail: francois.rivet@ims-bordeaux.fr

A. Mariano, Y. Deval, D. Dallet and J. B. Begueret,
University of Bordeaux, France

D. Belot
STMicroelectronics Crolles, France

A. Amara et al. (eds.), *Emerging Technologies and Circuits*,
Lecture Notes in Electrical Engineering 66, DOI 10.1007/978-90-481-9379-0_17,

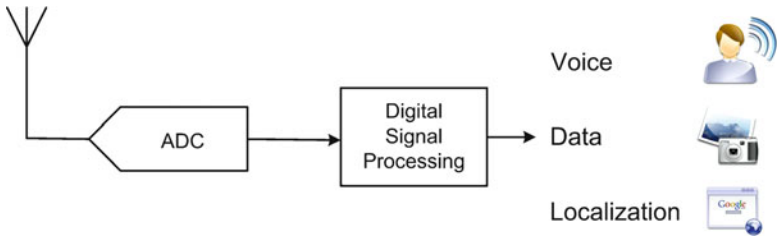


Fig. 1 Software radio receiving architecture

- Gain of compatibility: a common system can address any kind of standards and thus can be used wherever in the world. A mass production leads to a cost reduction.
- Gain of production time: research and development time is optimized between the apparition of a new standard and its use. Only updates (design, software) are required to accept new standards.
- Gain of performance: a SR system is able to reconfigure itself depending on the context (place, data rate, etc). It can adapt the data rate, the bandwidth using the most effective standard.

Performances of SR are not only technological. Industrial ones and easiness are thus proven. A Software-Radio receiving architecture is depicted in Fig. 1. The concept is to bring as close as possible the antenna to the analog to digital conversion. Thus, the ideal system is composed by an antenna, an ADC and a DSP. The DSP is reconfigurable by software and can address any standard, known or unknown. This architecture can adapt itself to any kind of radio context and treat any RF signal.

But, nowadays technological bottlenecks prevent from realizing such a utopian system at the lowest cost. Intermediate solutions are studied to achieve the Software Radio concept.

The US Army is the first to study SDR projects. The aim was to secure radio communications between operational units on a hostile battlefield. Radio-communications had to be reconfigured rapidly in order to escape to listening and jammers enemies. Defense Advanced Research Projects Agency (DARPA) financed researches. Project "Speakeasy" gave the first result at the beginning of the 1990s. The evolution was first to handle several standards on a 2 MHz to 2 GHz frequency band and reconfigure "on the fly" the communication device with one known standard. It went on the implementation of new standards through known standards in order to switch rapidly on a new way of communication. This necessity found obviously all its meaning in the military market.

1.3 Technological Bottlenecks Constrained by Mobile Context

SR ADC requirements are given by:

- At least a 10 GHz sampling frequency is required to convert from analog to digital any RF signals. This is imposed by Shannon theorem.

- A 16 Effective Number Of Bit (ENOB) is required to accept any dynamic range among all defined RF standards.
- The power consumption is directly a function of the sampling frequency. The higher the frequency is, the higher the power consumption is. In a context of mobility, the battery life is the major parameter to take into account.
- Silicon area is important as it determines the cost of the component. CMOS technology is preferred because of its low cost.

Given the Fig. 2a, such an ADC is not feasible nowadays at low power (Fig. 2b) [1], at high frequencies with an acceptable accuracy. Extrapolating current A/D converter characteristics the A/D converter for SR would consume about 1 kW (Fig. 2b). This is too much for handsets. The progress in A/D converters at the same power level (at the same sample frequency) is about 1.5 bit in 8 years [2]. As the power consumption issue depends on frequency and resolution, the next paragraph gives clues to understand the technological bottleneck [3].

Two mains limiting factors are thus exhibited: frequency and resolution. Power consumption is considered as depending on those factors. A/D convertors performances are mainly limited by three physical phenomena: thermal noise, jitter and quantification (minimal resolution). Currently, ADCs found on the market have specifications from 1 MHz with a 24-bit resolution to 2 GHz with an 8-bit resolution. These specifications are under the requirements of an ideal Software Radio system. It can be determined that at least 15 years of works is required to target a low power ADC answering to SR constraints, if one day it is feasible!

Technical information given show how critical is the A/D conversion. ADC suiting strong SR requirements is not expected to be achieved soon. Designers must find new architectures which relax ADC requirements to achieve a SR system, structures in rupture with traditional ones. In a first time, a Digital Software Defined Radio receiver will be presented. Then in a next part, as the digital domain is on obstacle to a SR system realization, an analog signal processor will be presented as a key for a true SR system achievement.

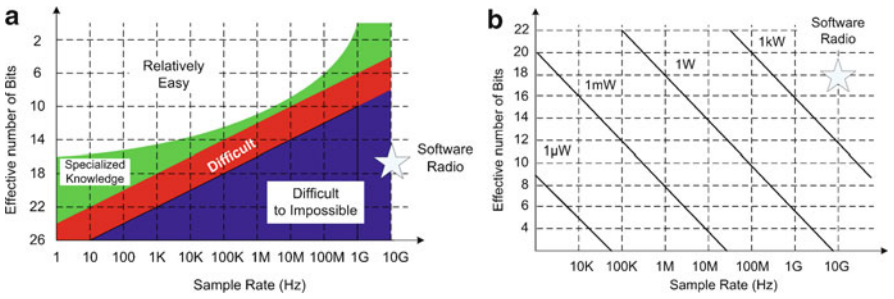


Fig. 2 (**a**) ADC feasibility – (**b**) ADC power consumption

2 Digital Software Defined Radio Receiver

Wireless front-end receivers of last generation mobile devices operate at least two frequency translations before I/Q demodulation. Frequency translation increases the system complexity, introducing several problems associated with the mixers (dynamic range limitation, noise injection from the local oscillator, etc.). Herein, the position of the analog-to-digital interface in the receiver chain can play an important role. Moving the analog-to-digital converter (ADC) as near as possible to the antenna, permits to simplify the overall system design and to alleviate requirements associated with analog functions (filters, mixers). The analog part has a substantial impact on the size and the cost of a component. Shifting the ADC closer to the antenna, allow eliminating some analog functions such as amplifiers, filters and mixers (integrated circuits and/or external components). These functions can be easily implemented in the digital domain, reducing the receiver complexity and therefore the power consumption. Nevertheless, the bottlenecks are the ADC required specifications, which become more severe with the antenna closeness.

2.1 Heterodyne Front-End Receiver Architecture with IF A/D Conversion – "Digital Receiver"

The principle of the so-called digital receiver architecture (Fig. 3) is to convert the analog signal as close as possible to the antenna [4]. The A/D conversion requires prior an extremely selective filtering of the desired channel. The antenna RF filters (BAW or SAW) can not perform this operation because of their wide bandwidth. Therefore, this conversion is generally not carried out directly on RF signal, but rather on IF signal. Moreover, the channel selection at RF frequency would require a filter array bank or a variable-frequency filter.

The frequency of the local oscillator is variable in order to translate the desired signal to the defined IF (placed in the center of the channel filter bandwidth).

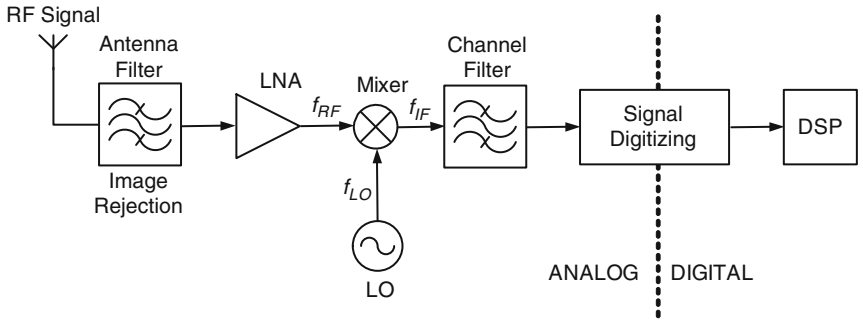


Fig. 3 Digital receiver architecture

The usual I/Q demodulation was performed after the signal digitization. This allows simplifying the overall receiver architecture which, in digital receivers, uses no more than a single mixer. This aspect makes digital heterodyne architecture competitive in terms of power consumption and design complexity, compared to the conventional analog receiver architectures. Nevertheless, the use of an IF passive filter, which integration is still off-chip technology (SAW), limit the scale integration of the receiver. An active filter can also be used in place of the passive one. However, the power consumption of the active filter increases when the IF increases.

In this type of architecture, the requirements for the A/D converter become more severe. Even though the channel filter in front of the A/D converter conserves the dynamic range, bandwidth and linearity requirements, the sampling frequency should be at least twice the IF frequency. Moreover, dynamic range and linearity requirements are more difficult to meet at higher frequencies due to circuit non-idealities and parasitic effects. All these restrictions make the A/D converter for high-IF digitizing much less efficient in terms of power consumption compared to baseband A/D converters.

These currently requirements have led to a great effort in designing high-resolution and high-speed ADCs. Delta-Sigma ($\Delta\Sigma$) modulators are popular nowadays for A/D conversion applications. There is an increasing interest in designing $\Delta\Sigma$ modulators using continuous-time circuitry for the loop filter. This is because continuous-time $\Delta\Sigma$ modulators allow dealing with higher clock frequency compared to discrete-time $\Delta\Sigma$ modulators. Consequently, for a given oversampling ratio (OSR), the conversion bandwidth is greatly increased. However, it is known that the continuous-time $\Delta\Sigma$ converters suffer from performance degradation due to non-idealities such as excess loop delay and clock jitter in the $\Delta\Sigma$ modulator loop. To perform the conversion of analog signal into digital one near to antenna, Band Pass (BP) $\Delta\Sigma$ modulator architectures using high-frequency resonators need to be developed. Band pass $\Delta\Sigma$ modulators sampling at high-IF allows reducing analog hardware and further realization of fully-integrated software-programmable receivers.

2.2 *Digital Receiver Using Band Pass Delta-Sigma A/D Converter*

The band pass Delta-Sigma converter architectures incorporate the idea of digitizing the analog signal directly into intermediate frequency as proposed in [5–7]. The BP $\Delta\Sigma$ converter is based on the principle of oversampling the analog signal. Indeed, the sampling frequency f_s is chosen according to the desired OSR:

$$\mathrm{OSR} = \mathrm{fs}/2\mathrm{B}$$

where B is the signal bandwidth.

The ratio between sampling frequency f_s and the center frequency of the resonator is chosen generally equal to four ($f_s = 4.f_{in}$) in order to simplify the $\Delta\Sigma$

modulator design and I/Q demodulation [8]. However, this ratio may be different. For example, to reduce the sampling frequency and therefore power consumption, this ratio could be equal to 5/4 [9] or 4/3 [10]. On the other hand, a reduction of this ratio implies a reduction of the OSR and thus a reduction of the SNR.

Figure 4 shows the implementation of a band pass ΔΣ modulator within a radio receiver. An IF filter is placed between the output of the mixer and the ΔΣ modulator to perform the channel selection. While DT modulators necessarily require a channel filter to avoid spectral aliasing [11], CT modulators can be implemented without this filter thanks to its intrinsic anti-aliasing filtering property inferred in its signal transfer function (STF) [12].

The CT band pass ΔΣ converter architecture is presented in Fig. 5. The analog signal is converted by the band pass ΔΣ modulator in a low-resolution and high frequency digital signal. Thus, a decimation filter is used to convert the modulator output in a high-resolution and low frequency digital signal. This signal is equal to twice the bandwidth (or the Nyquist frequency).

The ΔΣ converters use two fundamental principles to improve the SNR. First, the analog signal is oversampled in order to spread the quantization noise on the whole spectrum [13]. Then, the modulator feedback loop imposes quantization noise-shaping: the quantization noise is rejected out of the desired bandwidth. These two features allow ΔΣ converters to achieve high resolutions without

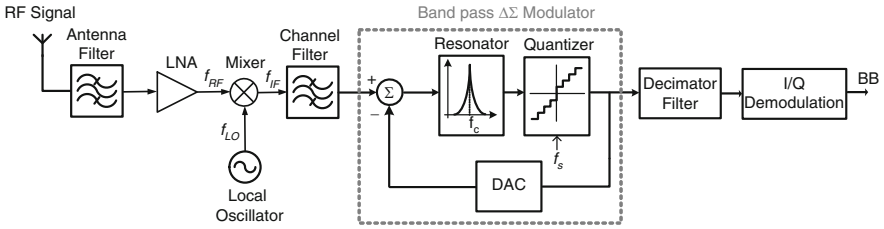


Fig. 4 Digital receiver using a band pass ΔΣ modulator

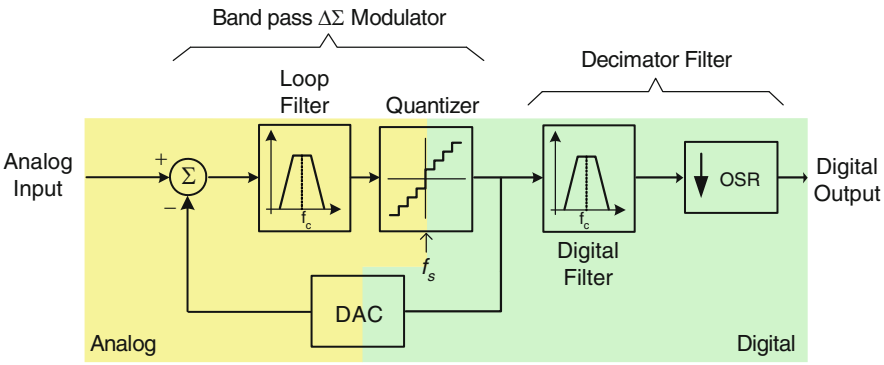


Fig. 5 Continuous-time band pass Delta-Sigma converter

requiring severe constraints in terms of analog blocks, such the anti-aliasing filter. This makes ΔΣ converters very interesting in comparison with Nyquist A/D converters, because these last ones require a very selective anti-aliasing filter.

2.3 *Proposed Continuous-Time Band Pass Delta-Sigma Modulator*

The architecture proposed is based on an association of resonators [$H(s) = As/(s^2 + \omega_o^2)$] with two different types of feedback Digital-to-Analog Converter (DAC). It offers a full control over the noise-shaping behavior, leading to the so-called multi-feedback architecture. This architecture allows the achievement of a higher order noise-shaping which maintains the modulator stability. The DAC used are return-to-zero (RZ) and half-return-to-zero (HRZ). The quantizer implements a 3-bit flash architecture. We choose a 3-bit quantizer instead of a usual 1-bit quantizer in order to reduce the quantization noise and to improve stability, at the cost of introducing mismatch errors [5]. The quantizer employs an input adapter amplifier, seven comparators (one per comparison level) with associated latches, an encoder matrix and three output buffers. Figure 6 shows the Continuous-Time Delta-Sigma modulator architecture.

A behavioral model of the proposed architecture was developed using VHDL-AMS. The input frequency is set to $f_c = 1$ GHz and the sampling frequency to $f_s = 4$ GHz. The results were analyzed within a bandwidth of BW = 20 MHz [14]. Figure 7 depicts the output spectrum of the proposed multi-bit CT BP ΔΣ modulator. Note that the noise is large everywhere, except in the desired band

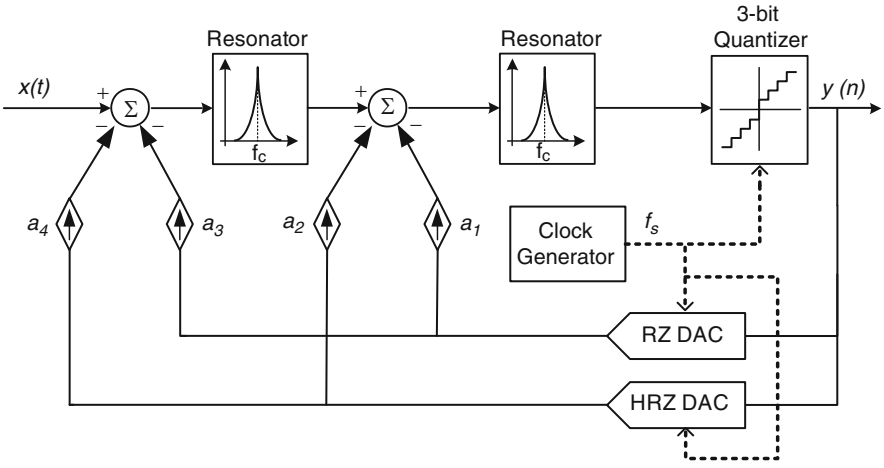


Fig. 6 CT BP ΔΣ modulator architecture

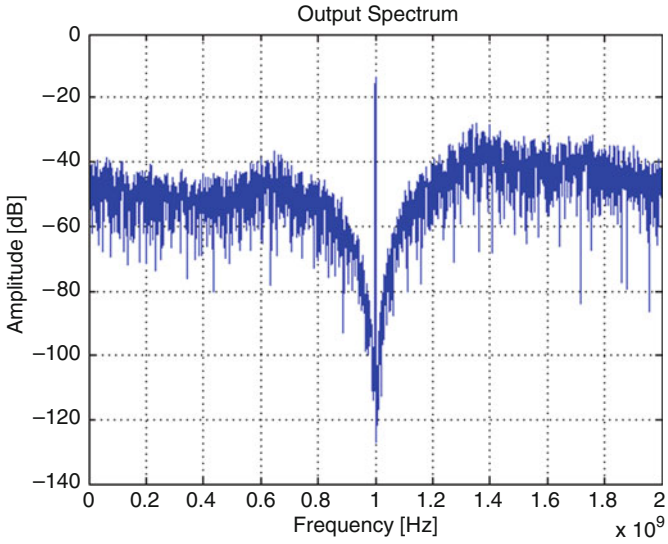


Fig. 7 Output spectrum plot of the proposed multi-bit CT BP $\Delta\Sigma$ modulator

around the carrier. This architecture achieves a SNR of 87 dB in a 20 MHz bandwidth. The SNR in a narrow 1 MHz bandwidth extrapolates 100 dB.

3 A Disruptive Software Radio Receiver Architecture

3.1 Sampled Analog Signal Processor

3.1.1 Principle

A Software-Radio (SR) architecture was proposed to challenge frequency and resolution constraints. It is composed by a LNA, an analog processor to perform low power analog calculations at RF frequencies, a low frequency ADC and a DSP. A component located between the LNA and the ADC pre-conditions analogically the RF signal (Fig. 8). Technological bottleneck of the ADC is consequently avoided. This component was called Sampled Analog Signal Processor (SASP) [15].

The SASP carries out basic analog operations on analog discrete time voltage samples. Its purpose is to reduce the RF signal data rate before digital conversion. The processor is fully analog and allows working directly at RF frequencies at acceptable power consumption. It is reconfigurable by software through analog parameters defined by a DSP. The SR concept dedicated to mobile terminals is consequently addressed by a hardware component fully controlled by a DSP

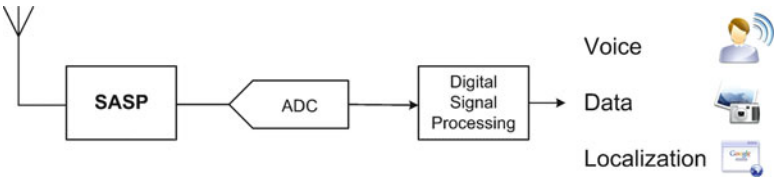


Fig. 8 SR receiving chain

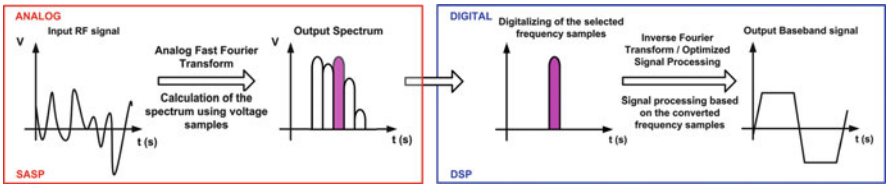


Fig. 9 SASP principle

which is able to adapt all the receiving chain by itself at RF frequency and low power consumption. Whereas the digital conversion technological bottleneck is avoided, new challenges appear in the analog domain.

3.1.2 A Fast Fourier Transform

The SASP selected a spectral envelope of a RF signal among any RF signals. The SASP processed analogically the RF input signal spectrum thanks to an analog Discrete Fourier Transform (DFT) to reach that target, the. Once the spectrum processed, voltage samples representing the signal envelope to be treated are converted into digital. The selection of few analog voltage samples among thousands replaces the classical mixing and filtering operations. These both operations are done into the frequency domain at the same time by the only conversion of the desired signal envelope. It reduces dramatically the A/D conversion frequency from GHz frequencies to MHz frequencies (Fig. 9).

3.2 Architecture

The SASP implements basic analog blocks to provide an analog FFT calculation at RF frequencies. It integrates three main parts to carry out the FFT (Fig. 10) [16].

- A continuous-time signal pre-conditioning (Filtering)
- A sampler and the analog FFT
- A RF signal envelope selection unit and the Analog to Digital Conversion

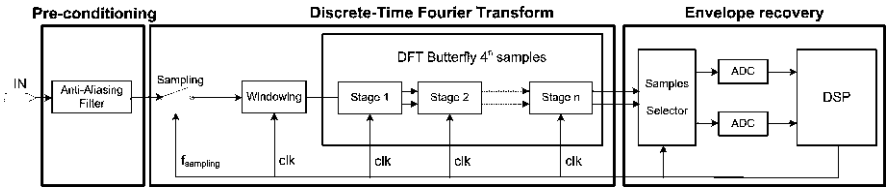


Fig. 10 SASP architecture

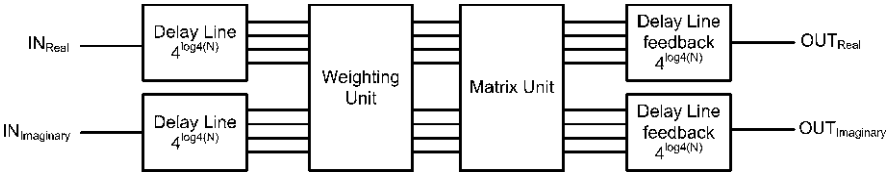


Fig. 11 Stage architecture

3.2.1 A Pipelined Analog FFT

The SASP processed the FFT algorithm of Cooley-Tukey. A radix-4 pipeline FFT was chosen to improve the speed efficiency. The FFT uses $\log_4(N)$ stages to process analogically the RF signal. Each stage implements one basic module which runs with two processing phases [17, 18]:

- Summation/subtraction and weighting factor [19].
- Feedback storage.

3.2.2 Design Matters

The signal is processed using analog voltage samples. As their values represent complex numbers and are stored analogically, real and imaginary parts of the FFT have to be calculated separately. Thus, a basic module implements the basic operations through the three main parts (Fig. 11):

- A delay line
- A Processing Unit [20] composed by a Weigthing Unit and a Matrix Unit to process the basic analog operations on voltage samples
- A feedback delay line

Every stage described in the DFT part of the SASP architecture is designed thanks to this architecture. Differences lie on the depth of the delay line and the coefficients applied in the weighting unit.

The delay line is composed by capacitors and switches. A differential structure was chosen to improve accuracy during the charge transfers. A digital signal Clk_{IN} governs voltage sample load into the capacitor (Fig. 12a). The charge is stored

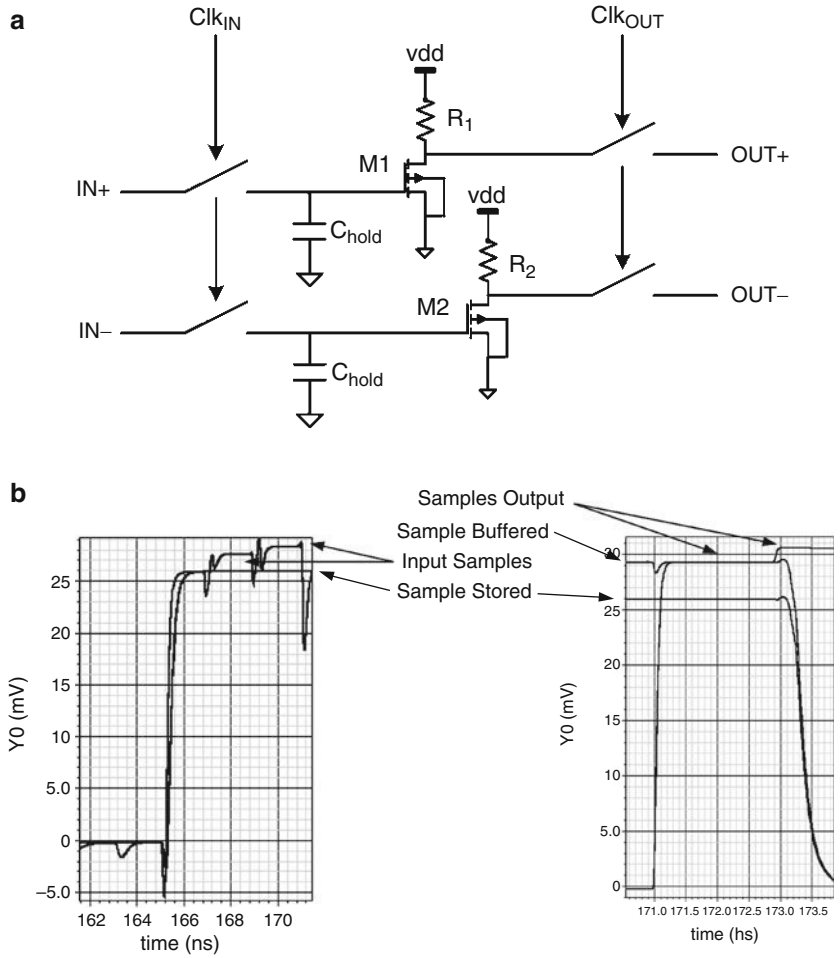


Fig. 12 (**a**) Delay schematics – (**b**) Simulation of a charge transfer

during a given delay time and output on the digital signal Clk_{OUT} command (Fig. 12a). The charge is buffered to avoid any loss during storing and transfer. Figure 12b exhibits a charge transfer simulation.

3.2.3 A 64-Sample SASP

The circuit considered here is a 64-point radix-4 FFT processor (Fig. 13b). It is a demonstrator designed with 65 nm CMOS technology of STMicroelectronics. A Post Layout Simulation is proposed (Fig. 13a). The sampling frequency is 500 MHz. A sine wave with a frequency of 56.64 MHz is sent. Real and Imaginary part of the spectrum are output by the way of 64 samples for a given FFT. Only the

desired sample among the 64 is stored and displayed during a FFT time. As the sine wave frequency is not an entire number of the sampling frequency, the spectrum is expected to change every processed FFT. As 56.64 MHz is exactly 7.25 times the frequency resolution of a voltage sample, four different kinds of spectrum are noticed (Fig. 13a). This observation leads to think to SASP applications such as frequency demodulation and concurrent reception.

The SASP demonstrator has a die area of 1.44 mm^2 (Fig. 13b). It maximal working frequency is 1 GHz with a dynamic range of 200 mV under 1.2 V supply voltage. It consumes 360 mW.

3.3 A Software Radio System

3.3.1 SASP Configuration

The SASP is able to receive any RF signal. But, its configuration ruled by the sampling frequency and the number of samples taken into account has to be adapted for a given standard. The maximal number of samples is 65536. That is why a 65536-sample SASP is the chip to be designed to be integrated in a mobile phone [21]. Configurations are exhibited in Fig. 14 for known standards. The number of

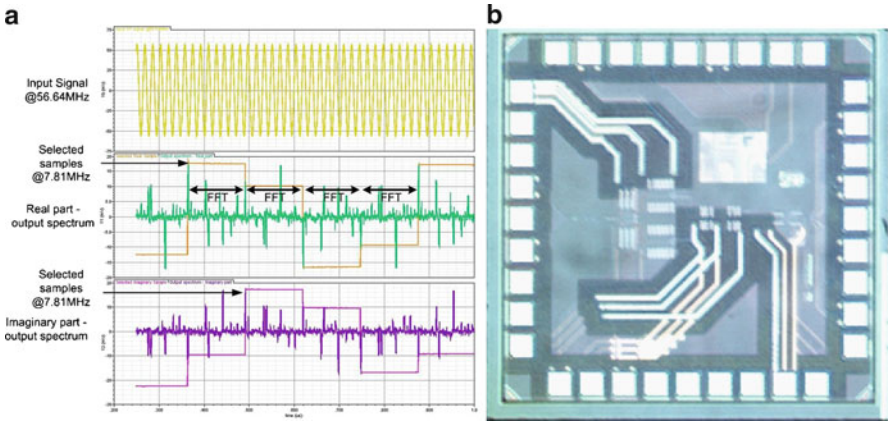


Fig. 13 (**a**) 64-sample SASP Post Layout Simulation – (**b**) 64 sample-SASP die photography

System	Carrier Frequency	Channel Bandwith	Modulation	Number of samples	$f_{sampling}$
GSM	925–960MHz	200kHz	GMSK	6	2.184GHz
DCS	1805–1880MHz	200kHz	GMSK	3	4.368GHz
UMTS	2110–2170MHz	5MHz	QPSK, HPSK	65	5.041GHz
Bluetooth	2402–2480MHz	1MHz	GFSK	12	5.461GHz
802.11g	2412–2472MHz	20MHz	OFDM	250	5.243GHz

Fig. 14 65536-sample SASP configuration

samples represents the number of sample of a signal envelope to convert into digital.

System	Carrier frequency (MHz)	Channel bandwidth	Modulation	Number of samples	$f_{sampling}$ (GHz)
GSM	925–960	200 kHz	GMSK	6	2.184
DCS	1,805–1,880	200 kHz	GMSK	3	4.368
UMTS	2,110–2,170	5 MHz	QPSK, HPSK	65	5.041
Bluetooth	2,402–2,480	1 MHz	GFSK	12	5.461
802.11g	2,412–2,472	20 MHz	OFDM	250	5.243

3.3.2 Frequency Demodulation

Figure 15 depicts the example of a BPSK modulation. The input bits are encoded through a binary phase shifting. The RF signal amplitude remained the same, but as

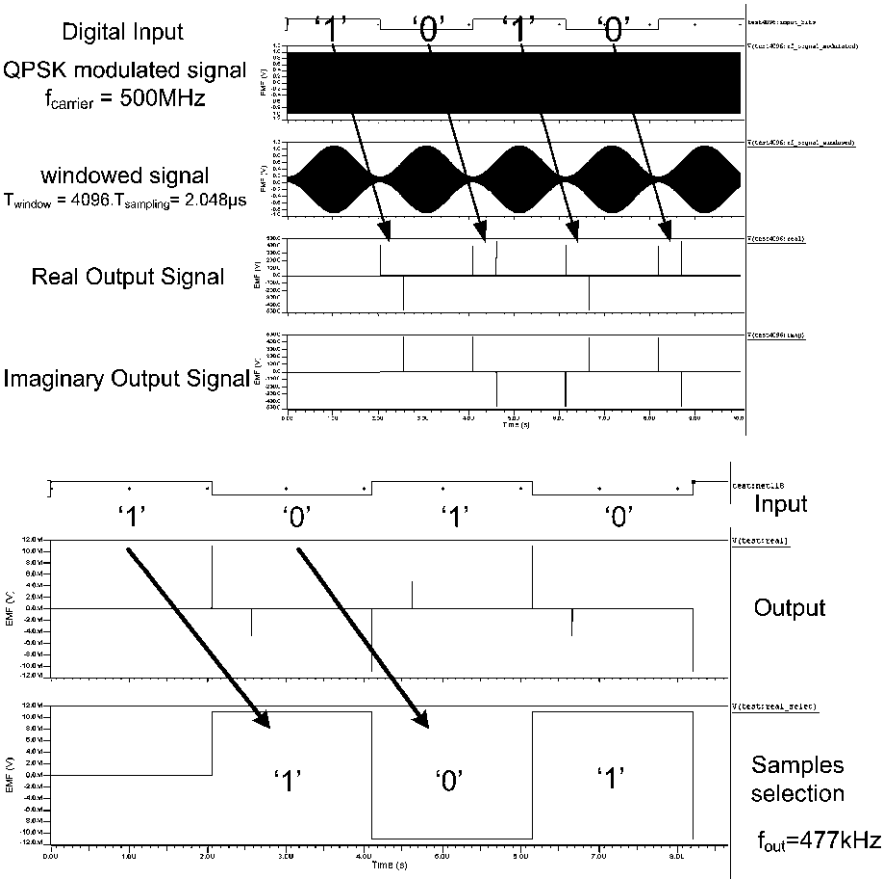


Fig. 15 Frequency demodulation example

the phase changes, the real and imaginary output spectrum were not the same depending on whether a '0' or a '1' is encoded. A BPSK demodulation could be optimized with the SASP by a relevant interpretation of the output spectrum (Fig. 15). A VHDL-AMS simulation of a 4096-sample SASP was done with a carrier frequency of 500 MHz for a matter of simplicity. The BPSK modulated signal received was first windowed. Its length was sized to be the timing of a modulated bit (here 2.048 μs). The sampling frequency was 2 GHz. The spectral accuracy was thus 477 kHz. Only the voltage sample among the 4096 is selected. The output data rate is consequently 477 kHz. The direct lecture of the spectrum enables a direct demodulation of the modulated BPSK signal. It is easy to imagine for any kind of modulation an optimized algorithm into the frequency domain to perform the signal recovering. A costly Inverse Fast Fourier Transform is no more required. In the example (Fig. 15), the working frequency was divided by more than 1,000. The ADC and the DSP constraints are relaxed.

3.3.3 Concurrent Reception

The envelope selection carried out by the selection of only few samples at the output of the SASP was not limited to the selection of only one RF signal envelope. The output samples representing several envelopes could be buffered to be converted at a lower rate. This is the concept of concurrent reception. Figure 16 depicts the capture of samples representing two signal envelopes among N samples output by the SASP. It is just a matter of selecting the samples of both envelopes processed from the same received RF signal. This principle enables to receive different standards at the same time, like GSM and Bluetooth. A common configuration is just to be defined.

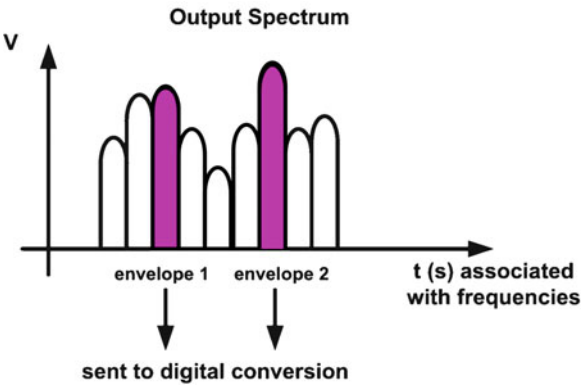


Fig. 16 Frequency demodulation principle

4 Conclusion

This chapter has presented an overview of Software-Defined to Software Radio systems de dedicated to handsets. Technological bottlenecks were exhibited and two solutions were proposed to answer it. The first one is an intermediate frequency digitalization solution, but it is a narrow band system that locks the architecture to the Software-Defined Radio concept. A degree of reconfigurability is given by a second solution. It is a sampled analog signal processing system that allows to access to a full Software Radio architecture. This disruptive system opens researches to a new kind of radio architectures.

References

1. R. Schiphorst, F.W. Hoeksema, C.H. Slump, The front end of software-defined radio: Possibilities and challenges, in *Proceedings of the Annual CTIT Workshop on Mobile Communications*, 2001
2. R. Walden, Performance trends for analog to digital converters. Commun. Mag. IEEE **37**(2), 96–101 (Feb 1999)
3. P. Seen, Radio logicielle dans les terminaux: quels impacts technologiques ? 2007
4. A.E. Cosand, J.F. Jensen, H.C. Choe, C.H. Fields, IF-sampling fourth-order Bandpass ΔΣ modulator for digital receiver applications. IEEE J. Solid-State Circ. **39**, 1633–1639 (2004)
5. J.A. Cherry, W.M. Snelgrove, *Continuous-Time Delta-Sigma Modulators for High-Speed A/D Conversion* (Kluwer, Boston, MA, 2000) ISBN: 0-7923-8625-6
6. A. Jayaraman, P. Asbeck, K. Nary, S. Beccue, K.C. Wang, Bandpass delta-sigma modulator with 800MHz center frequency, in *GaAs IC Symposium*, 1997, pp. 95–98
7. W. Gao, J.A. Cherry, W.M. Snelgrove, A 4GHz fourth-order SiGe HBT band pass ΔΣ modulator, in *VLSI Circuits Symposium*, 1998, pp. 174–175
8. R. Schreier, W.M. Snelgrove, Decimation for bandpass sigma-delta analog-to-digital conversion. IEEE Int. Symp. Circ. Syst. **3**, 1801–1804 (1990)
9. T. Salo, S. Lindfors, K.A.I. Halonen, A 80-MHz bandpass ΔΣ modulator for 100-MHz IF receiver. J. Solid State Circ. **37**, 798–808 (2002)
10. U.V. Koc, J. Lee. Direct RF sampling continuous-time bandpass Delta-Sigma A/D converter design for 3G wireless applications, in *IEEE Internatinal Symposium on Circuits and Systems*, 2004, pp. 409–412
11. T. Salo, S.J. Lindfors, K.A.I. Halonen, 80-MHz bandpass ΔΣ modulators for multimode digital IF receivers. J. Solid State Circ. **38**(3), 464–474 (2003)
12. O. Shoaei, W.M. Snelgrove, A multi-feedback design for LC bandpass Delta-Sigma modulators. Int. Symp. Circ. Syst. **1**, 171–174 (1995)
13. R. Gray, Oversampled Sigma-Delta modulation. IEEE Trans. Commun. **35**, 481–489 (1987)
14. A. Mariano, D. Dallet, Y. Deval, J.B. Begueret, VHDL-AMS behavioral modeling of high-speed continuous-time Delta-Sigma modulator, in *IWADC – International Workshop on ADC Modelling and Testing*, Sept 2007, pp. 118–121
15. F Rivet, Y Deval, J-B Bégueret, D Dallet, D Belot, A disruptive software-defined radio receiver architecture based on sampled analog signal processing, *IEEE Radio Frequency Integrated Circuits Symposium (RFIC'07)*, Honolulu, USA, June 3–5, 2007
16. F. Rivet, Y. Deval, J.-B. Begueret, D. Dallet, P. Cathelin, D. Belot, A disruptive receiver architecture dedicated to software defined radio, *IEEE Transactions on Circuits and Systems (TCAS-II), Software Defined Radio Special Issue*, Apr 2008, pp. 344–348

17. E. Swartzlander, W. Young, S. Joseph, A radix 4 delay commutator for fast fourier transform processor implementation. IEEE J. Solid-State Circ. **19**, 702–709 (Oct. 1984)
18. A. El-Khashab, Modular pipeline fast fourier transform algorithm, Ph.D. dissertation, University of Texas at Austin, 2003
19. E. Monastra, J. Huah, Pipelined fast fourier transform processor, US Patent number 5.038.311
20. S. Sayegh, A pipeline processor for mixed-size FFT's. IEEE Trans. Signal Process. **40**, 1892–1900 (Aug. 1992)
21. F. Rivet, Y. Deval, J.-B. Begueret, D. Dallet, P. Cathelin, D. Belot, From software-defined to software radio: Analog signal processor features, *IEEE Radio and Wireless Symposium (RWS'09)*, San Diego, USA, January 18–22, 2009