Effect of Device Scaling for Low Power Environment

Vijay Kumar Sharma

Department of Electronics & Communication, College of Engineering, Teerthanker Mahaveer University, U.P. (India)

vijay.buland@gmail.com

Abstract: MOS ICs have met the world's growing needs for electronic devices for computing, communication, entertainment, automotive, and other applications with steady improvements in cost, speed, and power consumption. Such steady improvements in turn stimulate and enable new applications and fuel the growth of IC sales. Microelectronics has grown tremendously in the past three decades because of the consistent scaling of CMOS technology. This reduction in size has enabled very dense transistors chips that have improved speed, functionality, and power compared to their predecessors. To achieve an optimal design, tradeoff exists between power and performance at each stage of the design. Therefore the designer must understand the sources of power consumption and make these tradeoffs.

[Vijay Kumar Sharma. **Effect of Device Scaling for Low Power Environment.** Report and Opinion 2011;3(2):24-30]. (ISSN: 1553-9873). http://www.sciencepub.net.

Keywords: Scaling, Leakage, low Power, MOS

1. Introduction

During the early 1970s, the basic MOS transistor structure could be scaled to smaller physical dimensions. One could postulate a "scaling factor" of S, the fractional size reduction from one generation to the next generation, and this scaling factor could then be directly applied to the structure and behaviour of the MOS transistor in a straightforward multiplicative fashion [1-3]. A CMOS technology generation could have a minimum channel length L_{min}, along with technology parameters such as the oxide thickness tox, the substrate doping N_A, the junction depth xj, the power supply voltage V_{DD}, the threshold voltage Vth, etc. Thus, the structure of the next generation process could be known beforehand, and the behaviour of circuits in that next generation could be predicted in a straightforward fashion from the behaviour in the present generation. The scaling theory is solidly grounded in the basic physics and behaviour of the MOS transistor. Scaling theory allows a "photocopy reduction" approach to feature size reduction in CMOS technology, and while the dimensions shrink, scaling theory causes the field strengths in the MOS transistor to remain the same across different process generations. Thus, the "original" form of scaling theory is constant field scaling [4].

In recent years, there has been an increasing trend towards the use of many types of portable electronic equipment. In such portable applications, it is extremely important to minimize current consumption due to the limited availability of battery power. When the whole circuit or segments of it are not in use, they must quickly be switched into a sleep mode in which they almost consume no power. Still leakage current may cause some power consumption even in the sleep mode. If the circuit could be designed such that there is very low leakage current in this mode, then the lifetime of the portable application will increase dramatically. Moreover, the sleeping circuit must be easy to start up. This start up must also be soft, which means that no short circuit current should be allowed to flow through the circuit. In order to manage the active power consumption of high-performance digital circuits, there is a need for active leakage control techniques to gain significant leakage power savings as well as fast time constants for entering and exiting idle mode. Dynamic sleep transistors and body bias used along with clock gating to control active leakage for a 32-bit integer execution core in 130-nm CMOS technology. Their measurements of PMOS sleep transistor showed that there was a substantial reduction in leakage power, while the reactivation of block was achieved in less than two clock cycles [5].

PMOS body bias reduces leakage power with no performance penalty and similar reactivation time. Power measurements at 4 GHz, 1.3V, showed that there was an 8% total power reduction using dynamic body bias and 15% power reduction using a PMOS sleep transistor, for a typical activity profile. An external dual switch leakage controlled flip-flop which effectively reduces the leakage current by 4 times over the other leakage controlled flip flops. The generally unnoticed fact that the sleep transistors for leakage reduction can significantly damp the resonant supply noise due to their series resistance. An optimal sleep transistor sizing method considering the dominant resonant supply noise and showed that a smaller sleep transistor can offer a smaller worst case supply noise due to the increased damping. An adaptive sleep transistor technique which automatically dampens the resonant noise only when it is detected with simulations in the 32nm CMOS and showed that the resonant noise was reduced by 32%.

2. Technology scaling

Since the 1960's the price of one bit of semiconductor memory has dropped 100 million times and the trend continues. The cost of a logic gate has undergone a similarly dramatic drop. This rapid price drop has stimulated new applications and semiconductor devices have improved the ways people carry out just about all human activities. The primary engine the powered the ascent of electronics is "miniaturization". By making the transistors and the interconnects smaller, more circuits can be fabricated on each silicon wafer and therefore each circuit becomes cheaper. Miniaturization has also been instrumental in the improvements in speed and power consumption.

Gordon Moore made an empirical observation in the 1960's that the number of devices on a chip doubles every 18 months or so. The "Moore's Law" is a succinct description of the persistent periodic increase in the level of miniaturization. Each time the minimum line width is reduced, we say that a new technology generation or technology node is introduced. Examples of technology generations are 180nm, 130nm, 90nm, 65nm, 45nm…generations. The numbers refer to the minimum metal line width. Poly-Si gate length may be smaller. At each new node, the various feature sizes of circuit layout, such as the size of contact holes, are 70% of the previous node. This practice of periodic size reduction is called scaling. Historically, a new technology node is introduced every three years or so.

The main reward for introducing a new technology node is the reduction of circuit size by 2. (70% of previous line width means ≈50% reduction in area, i.e. 0.7 x 0.7= 0.49.) Since nearly twice as many circuits can be fabricated on each wafer with each new technology node, the cost per circuit is reduced significantly. That is the engine that drives down the cost of ICs.

Besides line width, some other parameters are also reduced with scaling such as the MOSFET gate oxide thickness and the power supply voltage. The reductions are chosen such that the transistor current density increases with each new node. Also, the smaller transistors and shorter interconnects lead to smaller capacitances. Together, these changes cause the circuit delays to drop. Historically, integrated circuit speed has increased roughly 30% at each new technology node [7]. Scaling does another good thing that, it reduce power supply voltage so lowering the power consumption.

With each new process generation, the entire lateral and some of the vertical dimensions of the transistors are scaled down to allow a higher level of integration. Figure 1 reflects the reduction of the key dimensions of a typical MOSFET with the corresponding increase of the doping densities.

Scaled dimensions and doping densities have an immediate impact on reducing the power dissipation, as well as increasing the circuit speed. The primary effect of process scaling is the reduction of all the capacitance, which provides a proportional decrease in the power consumption and circuit delays.

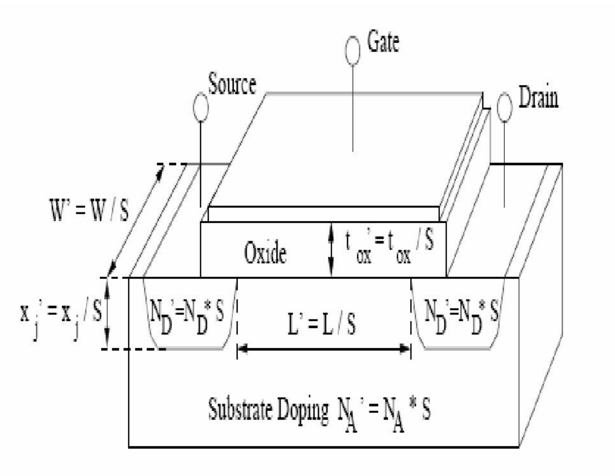


Figure 1 Scaling of a MOSFET by a factor of S

As today's technology scales below 90nm, the transistor density will continue to grow. The transistor delay will also continue to improve, at least modestly, to a 30% reduction per generation. The continued scaling of the technology has meant that designs that were limited by the amount of functionality on a chip are now limited by the amount of constrained power. In practice, there are two types of scaling strategies for MOSFET devices: full scaling or constant field scaling and constant voltage scaling.

2.1 Constant Field (CF) scaling, all the horizontal and vertical dimension of the transistor, as well as the power supply, are scaled down by a factor of S. In order to preserve the magnitude of the internal electric field, the doping densities need to be increased by the same factor S.

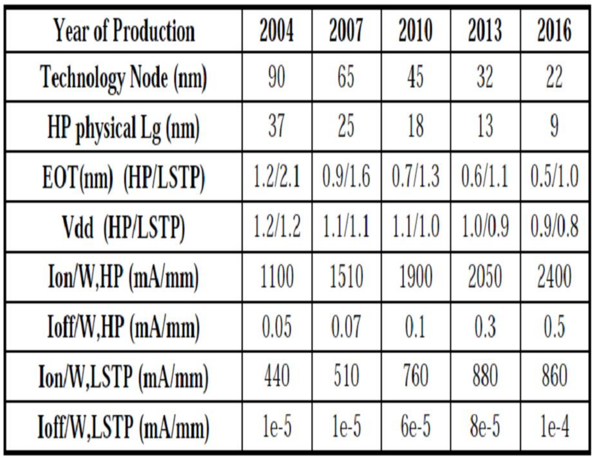
2.2 Constant Voltage (CV) scaling, all the dimensions of the MOSFET are reduced by a factor of S, as in full scaling, but the power supply voltage and the terminal voltage remain unchanged. The doping densities are increased by a factor of S^2 in order to preserve the charge-field relation. Table 1 summarizes the scaling factors for all the significant dimensions, power supply, doping densities of the MOS transistors, and changes in the key device characteristics for these two scaling strategies.

Table 1 Influence of scaling on MOS device characteristics

Parameter	**Constant Field (CF)**	**Constant Voltage (CV)**
Channel Length (L)	1/ S	1/ S
Channel Width (W)	1/ S	1/ S
Gate Oxide thickness (t_{ox})	1/ S	1/ S
Junction depth(x_j)	1/ S	1/ S
Power Supply Voltage (V_{DD})	1/ S	1
Threshold Voltage (V_{th})	1/ S	1
Doping Densities (N_A, N_D)	S	S^2
Oxide Capacitance (C_{ox})	S	S
Drain Current (I_D)	1/ S	S
Delay (τ)	1/ S	1/ S^2
Power Dissipation (P_{diss})	1/ S^2	S
Leakage Power ($P_{leakage}$)	Exp	1
Power Density (P/Area)	1	S^3
Power Delay Product (PDP)	1/ S^3	1/ S

It is evident that CF scaling reduces both the drain and the supply voltage by a factor of S. Therefore, the power dissipation of the transistor decreases by a factor of S^2, and increases by the factor S in CV scaling. This significant reduction of the power dissipation is one of the most attractive features of CF scaling. However, Intel has used CV scaling in their microprocessors until the appearance of 0.8um technology, where a 5V supply voltage has been used to maintain the compatibility with the supply voltage of conventional systems, and also to obtain a higher operation speed. CF scaling has been used since 0.5µm technology has evolved. The main reason for the supply voltage scaling that began in the 0.5µm generation is that CV scaling increases the drain current densities and the power density by a factor of $S^{3.}$ This large increase in the current and power densities can eventually cause serious reliability problems such as electro migration, hot carrier degradation, oxide breakdown, and electrical over-stress, for the scaled transistor. Another reason for reducing the power supply voltage is to decrease the power consumption of the chip. However, the CF scaling causes the sub-threshold leakage currents to grow exponentially and become an increasingly larger component of the total power dissipation. Therefore, effective leakage minimization techniques need to be designed.

Table 2 Technology scaling from 90nm to 22nm



Year of Production	2004	2007	2010	2013	2016
Technology Node (nm)	90	65	45	32	22
HP physical Lg (nm)	37	25	18	13	9
EOT(nm) (HP/LSTP)	1.2/2.1	0.9/1.6	0.7/1.3	0.6/1.1	0.5/1.0
Vdd (HP/LSTP)	1.2/1.2	1.1/1.1	1.1/1.0	1.0/0.9	0.9/0.8
Ion/W,HP (mA/mm)	1100	1510	1900	2050	2400
Ioff/W,HP (mA/mm)	0.05	0.07	0.1	0.3	0.5
Ion/W,LSTP (mA/mm)	440	510	760	880	860
Ioff/W,LSTP (mA/mm)	1e-5	1e-5	6e-5	8e-5	1e-4

3. Subthreshold Current

Circuit speed improves with increasing I_{on}, therefore it would be desirable to use a small V_{th}. At $V_{gs}<V_{th}$, an N-channel MOSFET is in the off-state. However, an undesirable leakage current can flow between the drain and the source. The MOSFET current observed at $V_{gs}<V_{th}$ is called the subthreshold current. This is the main contributor to the MOSFET off-state current, I_{off}. I_{off} is the I_{ds} measured at $V_{gs}=0$ and $V_{ds}=V_{dd}$. It is important to keep I_{off} very small in order to minimize the static power that a circuit consumes even when it is in the standby mode. For example, if I_{off} is a modest 100nA per transistor, a cell-phone chip containing one hundred million transistors would consume so much standby current (10A) that the battery would be drained in minutes without receiving or transmitting any calls. A desk-top PC chip may be able to tolerate this static power but not much more before facing expensive problems with cooling the chip and the system.

Figure 2 shows a typical subthreshold current plot. It is almost always plotted in a semilog I_{ds} versus V_{gs} graph. When V_{gs} is below V_{th}, I_{ds} is an exponential function of V_{gs}. Figure 2 explains the subthreshold current. At V_{gs} below V_{th}, the inversion electron concentration is small but nonetheless can allow a small leakage current to flow between the source and the drain.

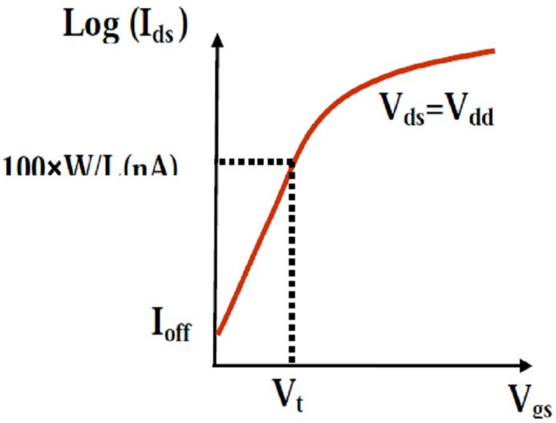


Figure 2 subthreshold current

For given W and L, there are two ways to minimize I_{off}. The first is to choose a large V_{th}. This is not desirable because a large V_{th} reduces I_{on} and therefore increases the gate delays. The preferable way is to reduce the subthreshold swing. S can be reduced by increasing C_{ox} i.e. using a thinner t_{ox}, and by decreasing C_{dep}, i.e. increasing W_{dep}. Second an additional way to reduce S, and therefore to reduce I_{off}, is to operate the transistors at a lower temperature.

SiO_2 has been the preferred gate insulator for silicon MOSFET since its very beginning in the 1960’s and the oxide thickness has been reduced over the years from 300nm for 10mm technology to 1.2nm for 65nm technology.

There are two reasons for the relentless drive to reduce the oxide thickness. First, a thinner oxide, i.e. a larger C_{ox} raises Ion. A large I_{on} is desirable for maximizing the circuit speed. The second reason is to control V_{th} roll-off (and therefore the subthreshold leakage) in the presence of falling L. Figure 3 shows that the oxide thickness has been scaled roughly in proportion to the line width.

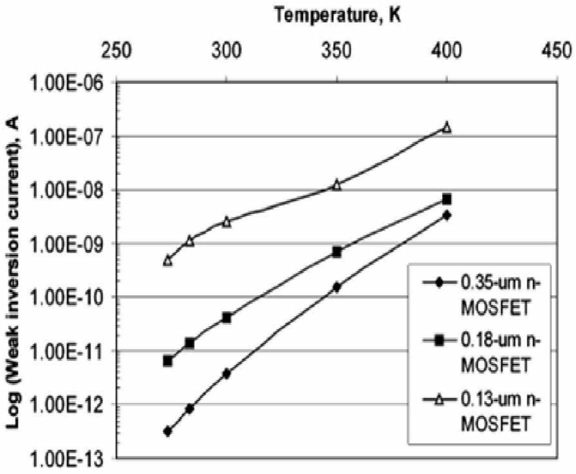


Figure 3 subthreshold current vs. Temperature curve

So, thinner oxide is desirable. Manufacturing thin oxide is not easy, but it is possible to grow very thin and uniform gate oxide films with high yield. Oxide breakdown is another limiting factor. If the oxide is too thin, the electric field in the oxide can be so high as to cause destructive breakdown.

Yet another limiting factor is that long term operation at high field, especially at elevated chip operating temperatures, breaks the weaker atomic bonds at the Si/SiO_2 interface thus creating oxide charge and V_{th} shift. V_{th} shifts cause circuit behaviors to change and raise reliability concerns. For SiO_2 films thinner than 1.5nm, tunneling leakage current becomes the most serious limiting factor. This large leakage would drain the battery of a cell phone in minutes. Researchers are developing high-k dielectrics to replace SiO_2.

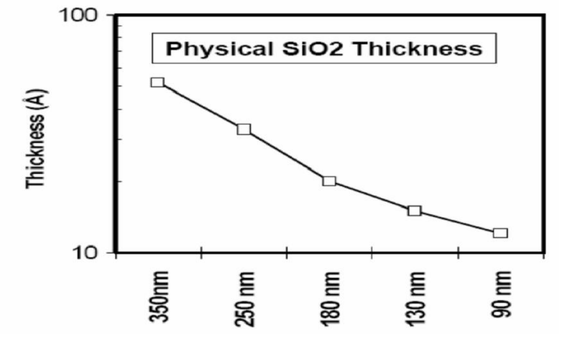


Figure 4 Oxide thickness at different technologies

The consequence is a leakage current that is several orders of magnitude smaller than that in SiO_2. A metal gate is used to reduce the poly-Si gate depletion. These problems can be minimized by inserting a thin SiO_2 interfacial layer between the silicon substrate and the high-k dielectric and using a metal gate instead of a poly-Si gate.

4. Low Power Applications

There is a steadily growing market for low-power applications of CMOS technology and it is the battery-powered nature of most of these applications that particularly creates the low-power constraint. To achieve good battery life, these circuits simply cannot dissipate very much power. Roughly speaking, these are circuits that consume less than 1 W/cm with a subgroup of ultralow power circuits in the range below 1 mW/cm . Low-power constraints fall into two broad categories: those that relate to active mode power dissipation and those that relate to dissipation in the quiescent state. Some types of applications are primarily sensitive to active power considerations, since they are switched off when not in use. Other applications may be turned on almost all the time, but rarely ever actually compute anything and so are more concerned with the quiescent power dissipation. Since MOSFET design limits are different for these two cases, they need to be considered separately.

Redefining the problem, the architecture, the algorithms, and/or the protocols can often save several orders of magnitude in power dissipation. At the device design level, the important low-power variables are the threshold voltage, the gate leakage current, and the device size, which largely determines the body-to-drain tunneling dissipation. For current generations of technology, the latter effect is not usually significant, but at the limits of scaling, it should become quite important. For active mode dissipation, these parameters offer strong tradeoffs between speed and low power.

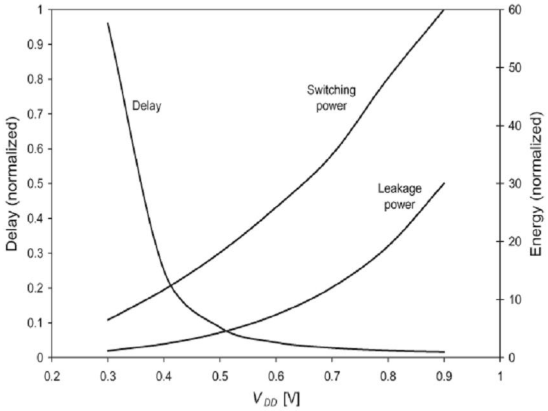


Figure 5 Delay, Power vs. V_{DD} Curve

This tradeoff occurs because all three variables tend to simultaneously increase the circuit's speed and its dissipation during the time it is not switching.

The optimum value for increases for slow circuits to reduce static dissipation and increases by 20–100 mV when the tolerances are doubled from their nominal values.

5. Power limited scaling

The fundamental limit to constant field scaling regime is related to the non-scaling of the sub-threshold slope and increase of gate leakage as most of the other limiting factors are under designers control (voltage, frequency, die size, and architecture).

Reducing the supply voltage significantly reduces the switching power, but lowers the device switching speeds because of lower saturation currents. It is necessary to scale the threshold voltages according to the constant field model to maintain the performance. Threshold voltage reduction results in an exponential increase in transistor drain leakage currents, which represent a significant portion of the overall power budget today. With scaling of both the supply and threshold voltages, a minimum power is achieved when a balance is struck between the active and leakage power components. This optimum is at the point where leakage contributes to about 30%–40% of the total power during active operation of the circuit [9].

As a result, continued scaling in the 90, 65, and 45nm nodes and beyond departs from the constant field model and enters the power limited scaling regime. The continued scaling of technology outlined by ITRS still introduces new devices with lower thresholds. The power limited scaling regime is characterized by the use of multiple devices in the design optimized for different performance/power targets, together with slowed-down supply and threshold voltage scaling, and dramatic changes in chip architectures.

The power dissipation of high performance applications such as microprocessors, digital signal processors, and random access memories has increased along with the progress in CMOS technologies, where the design emphasis has been on maximizing the operational frequency. The increased power consumption raises a chip temperature which leads to electro-migration reliability problems, and degradation in the device performance. Thus, lowering the power dissipation is crucial for high performance VLSI designs. Also, applications are emerging for which the energy consumption is the key metric, and the speed of operation becomes less relevant. Generally, energy constrained VLSI applications such as micro-sensor networks and nodes, radio frequency identification, and biomedical devices have low activity rates and low speed, but the concern is to lengthen battery life. Ideally, the power consumption of these systems should decrease to the extent that they can harvest energy from environmental resources such as solar power, thermal gradients, radio-frequency, and mechanical vibration, and theoretically have unlimited lifetimes. Such ultra-low power applications have established a significant niche for sub-threshold circuits.

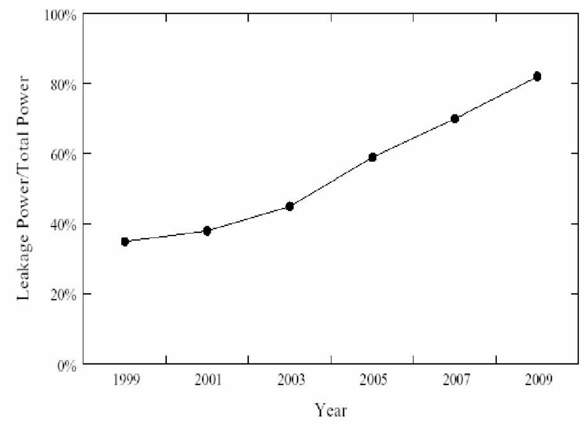


Figure 6 Projected leakage power as a fraction of the total consumption to ITRS

Figure 7 shows the trends of power supply voltage, threshold voltage, and gate oxide thickness versus channel length for high performance CMOS logic technologies. Sub-threshold non-scaling and standby power limitations bound the threshold voltage to a minimum of 0.2 V at the operating temperature.

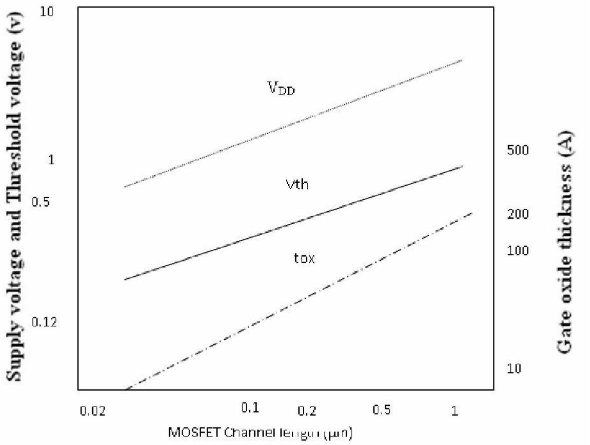


Figure 7 supply voltage V_{DD}, threshold voltage V_{th}, and gate oxide thickness t_{ox}, vs. channel length

Thus, a significant reduction in performance gains is predicted below 1.5 V due to the fact that the threshold voltage decreases more slowly than the historical trend, leading to more aggressive device designs at higher electric fields.

6. Scaling effect on circuit design

With continuing aggressive technology scaling, it is increasingly difficult to sustain supply and threshold voltage scaling to provide the required performance increase, limit energy consumption, control power dissipation, and maintain reliability. These requirements pose several difficulties across a range of disciplines. On the technology front, the question arises whether we can continue along the traditional CMOS scaling path reducing effective oxide thickness, improving channel mobility, and minimizing parasitic. On the design front, researchers are exploring various circuit design techniques to deal with process variation, leakage and soft errors [9-11].

7. Manage leakage power

For CMOS technologies beyond 90nm, leakage power is one of the most crucial design components which must be efficiently controlled in order to utilize the performance advantages of these technologies. It is important to analyze and control all components of leakage power, placing particular emphasis on sub-threshold and gate leakage power. A number of issues must be addressed, including low voltage circuit design under high intrinsic leakage, leakage monitoring and control, effective transistor stacking, multi-threshold CMOS, dynamic threshold CMOS, well biasing techniques, and design of low leakage data-paths and caches. While supply voltage scaling becomes less effective in providing power savings as leakage power becomes larger due to scaling, it is suggested that the goal is to no longer have simply the highest performance, but instead have the highest performance within a particular power budget by considering the physical aspects of the design. In some cases, it may be possible to balance the benefit of using high threshold devices from a low leakage process running at the higher possible frequency at a full V_{DD}, as opposed to using faster but leakier devices which require more voltage scaling in order to reach the desired power budget.

8. Conclusions

We have described most of the important physical phenomena that stand in the way of continued scaling of Si CMOS technology and have shown how these effects determine different limits for different circuit applications. Most of the application limits are set by limitations on the amount of power that can be dissipated in the three primary leakages: subthreshold channel current, gate-to-channel tunneling through the insulator, and body-to-drain junction tunneling currents for very short channels and at low temperature.

Correspondence to:

Vijay Kumar Sharma

Assistant Professor, College of Engineering,

Teerthanker Mahaveer University, Moradabad UP, India

Mobile phone: +918979620726

Email: vijay.buland@gmail.com

References

1. S. Borkar, “Design Challenges of Technology Scaling,” IEEE Micro, pp. 23–29, 1999.

2. G. G. Shahidi, “Challenges of CMOS scaling at be low 0.1µm,” The 12th International Conferenceon Microelectronics, October 31–November 2, 2000.

3. S. Kang and Y. Leblebici, CMOS Digital Integrated Circuits, McGraw-Hill, New York, 2003.

4. R. Puri, T. Karnik, R. Joshi, "Technology Impacts on sub-90 nm CMOS Circuit Design & Design methodologies," Proceedings of the 19th International Conference on VLSI Design, 2006.

5. Y. Taur and T. H. Ning, Fundamentals of Modern VLSI Devices. New York: Cambridge Univ. Press, 1998, ch. 3, pp. 120–128.

6. K. Roy and S. C. Prasad, Low-Power CMOS VLSI Circuit Design. New York: Wiley, 2000, chap. 2, pp. 28–29.

7. V. De, Y. Ye, A. Keshavarzi, S. Narendra, J. Kao, D. Somasekhar, R. Nair, and S. Borkar "Techniques for leakage power reduction," in Design of High- Performance Microprocessor Circuits, A. Chandrakasan, W. Bowhill, and F. Fox, Eds. Piscataway, NJ: IEEE, 2001, chap. 3, pp. 48–52.

8. K. Nose, M. Hirabayashi, H. Kawaguchi, S. Lee, a nd T. Sakurai, "Vth Hopping scheme to reduce subthreshold leakage for low-power processors," IEEE J. Solid-State Circuits, vol. 37, pp. 413–419, Mar. 2002.

9. Z. Chen, L.Wei, A. Keshavarzi, and K. Roy, "IDDQ testing for deep submicron ICs: Challenges and Solutions," IEEE Des. Test Comput, pp. 24–33, Mar.– Apr. 2002.

10. D. Duarte, Y. F. Tsai, N. Vijaykrishnan, and M. J. Irwin, "Evaluating runtime techniques for leakage power reduction," in Proc. 7th Asia and South Pacific and 15th Int. Conf. VLSI Design, 2002, pp. 31–38.

11. N. Sirisantana, L. Wei, and K. Roy, "High performance low-power CMOS circuits using multiple channel length and multiple oxide thickness," in Proc. Int. Conf. Computer Design, 2000, pp. 227–232.

12. Y. Oowaki, "A sub-0.1 µm circuit design with substrate-overbiasing," in Dig. Tech. Papers IEEE Int. Solid-State Circuits Conf., 1998, pp. 88–89.

1/1/2011