Challenges and Design Choices in Nanoscale CMOS

SIVA G. NARENDRA
Portland, OR

The driving force for the semiconductor industry growth has been the elegant scaling nature of CMOS technology. In this article, we will first review the history of technology scaling that follows Moore's law from the prespective of microprocessor designs. Challenges to continue the historical scaling trends will be highlighted and design choices to address two specific challenges, process variation and leakage power, will be discussed. In nanoscale CMOS technology generations, supply and threshold voltages will have to continually scale to sustain performance increase, limit energy consumption, control power dissipation, and maintain reliability. These continual scaling requirements on supply and threshold voltages pose several technology and circuit design challenges. One such challenge is the expected increase in process variation and the resulting increase in design margins. Concept of adaptive circuit schemes to deal with increasing design margins will be explained. Next, with threshold voltage scaling, subthreshold leakage power has become a significant portion of total power in nanoscale CMOS systems. Therefore, it has become imperative to accurately predict and minimize leakage power of such systems, especially with increasing within-die threshold voltage variation. A model that predicts system leakage based on first principles will be presented and circuit techniques to reduce system leakage will be discussed. It is essential to point out that this article does not cover all challenges that nanoscale CMOS systems face. Challenges that are not detailed in the main sections of the article and speculation on what future nanoscale silicon based CMOS systems might resemble are summarized.

Categories and Subject Descriptors: B.7.0 [**Integrated Circuits**]: General

General Terms: Design

Additional Key Words and Phrases: CMOS, nanoscale, leakage power, process variation

1. INTRODUCTION

MOS transistor-based integrated circuits have transformed the world we live in. It is estimated that there are more than 15 billion silicon semiconductor chips currently in use with an additional 500,000 sold each day [Smolan and Erwin 1998]. The ever-shrinking size of the MOS transistors that result in faster, smaller, and cheaper systems have enabled ubiquitous use of these chips. Among these semiconductor chips, a prevalent component is the

Author's address: 7180 SW 84th Avenue, Portland, OR 97223; email: siva@tyfone.com.
Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or direct commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 1515 Broadway, New York, NY 10036 USA, fax: +1 (212) 869-0481, or permissions@acm.org.
© 2005 ACM 1550-4832/05/0400-0007 $5.00

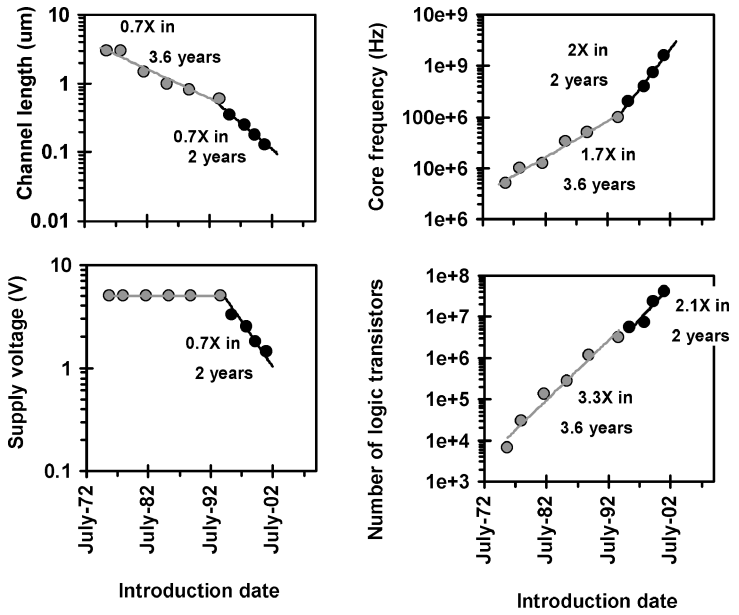


Fig. 1. Timeline on technology scaling and new microprocessor architecture introduction.

high-performance general-purpose microprocessor. Figure 1 illustrates the timeline on technology scaling and new high-performance microprocessor architecture introductions in the past three decades [Intel]. This trend holds in general for other segments of the semiconductor industry as predicted by Moore's law [Moore 1965]. In 1965, Gordon Moore showed that, for any MOS transistor technology, there exists a minimum cost that maximizes the number of components per integrated circuit. He also showed as transistor dimensions are shrunk (or scaled) from one technology generation to the next, the minimal cost point allows significant increase of the number of components per integrated circuit as shown in Figure 2.

Historically, technology scaling resulted in scaling of vertical and lateral dimensions by 0.7X each generation resulting in delay of the logic gates to be scaled by 0.7X and the integration density of logic gates to be increased by 2X. From the timeline shown in Figure 1 it is clear that there were two distinct eras in technology scaling—constant voltage scaling and constant electric field scaling.

Constant Voltage Scaling Era (First Two Decades). Technology scaling and new architectural introduction in this era happened every 3.6 years. Technology scaling should scale delay by 0.7X translating to 1.4X higher frequency. However, frequency scaled by 1.7X with the additional increase primarily brought about by increase in the number of logic transistors. As it can be seen from Figure 1 the number of logic transistors increased by 3.3X in each of the new introductions. Technology scaling itself would have provided only 2X—the additional increase was enabled by increase in die area of about 1.5X every generation [De and Borkar 1999].

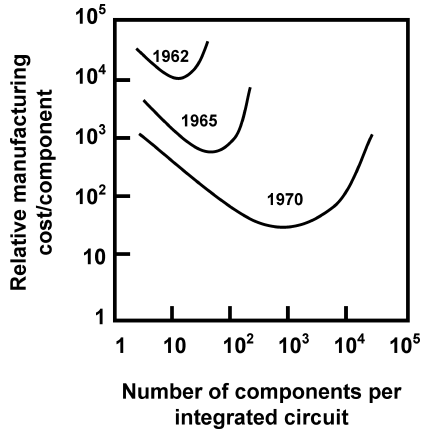


Fig. 2. Basic form of Moore's law.

Constant Electric Field Scaling Era (Past Decade). Technology scaling and new architectural introduction in this era happened every 2 years along with voltage scaling of 0.7X. As always technology scaling should scale delay by 0.7X translating to 1.4X higher frequency, but frequency increased by 2X in each new introduction. The additional increase in frequency was primarily brought by decrease in logic depth through architectural and circuit design advancements. The number of logic transistors grew only by about 2.1X every generation, which could be achieved without significant increase in die area. Since switching power is proportional to Area $\times$ ε/distance $\times$ V_{dd} $\times$ V_{dd} $\times$ F, it increased by ($1 \times 1/0.7 \times 0.7 \times 0.7 \times 2 =$) 1.4X every generation. Although the die size growth is not required for logic transistor integration, it is important to note that the total die area did continue to grow at the rate of 1.5X per generation [De and Borkar 1999] due to increase in the capacity of integrated memory.

In the past decade, technology and new architecture product cycles reduced from 3.6 years to 2 years. From the product development perspective, this requires concurrent engineering in product design, process design, and building of manufacturing supply lines [Kempf 1998]. The past decade also required supply voltage scaling imposed by oxide reliability and the need to slow down the switching power growth rate. From the process design stand point supply voltage scaling requires threshold voltage scaling [Thompson etl al. 1998; Taur and Ning 1998] so that the technology scaling can continue to provide 1.4X frequency increase. To prolong the tremendous growth the industry has experienced in the past three decades threshold voltage scaling and concurrent engineering has to continue. These requirements pose several challenges in the coming years including increase in process variation, worsening interconnects RC delay, and increase in subthreshold, gate, and tunneling leakage components [Taur and Ning 1998; Antoniadis and Chung 1991]. This article will focus in detail on two of the challenges—the increasing importance of process variation and leakage power, how it impacts digital CMOS circuits used in microprocessors and other high-performance integrated circuits.

1.1 Organization of the Article

In the subsequent sections, the effects of MOS parameter variation and leakage power on high-performance digital CMOS circuits, and potential circuit solutions that alleviate these effects will be presented.

Section 2 provides a brief background on the reasons for the increasing importance of process variation and leakage power. Section 3 focuses on different aspects of die-to-die threshold voltage variation and its impact on delay and power of the integrated circuit. Scalable adaptive circuit solutions that minimize impact of process variation will be discussed. Section 4 introduces the importance of taking into account the influence process variation on system's leakage power especially as technology scales and present several techniques that help reduce system leakage power. Finally, Section 5 enlists other challenges to continue the historical scaling trends and potential solutions. Summary and opinion on what future nanoscale MOS designs might encompass are also discussed.

2. BACKGROUND ON PROCESS VARIATION AND LEAKAGE POWER

Conventionally, CMOS technology has been scaled from one generation to the next so as to provide 30% smaller gate delay with 30% smaller dimensions, resulting in CMOS systems operating at about 40% higher frequency in half the area with reduced energy consumption. Scaled CMOS systems, such as new generation microprocessors, achieve at least an additional 60% frequency increase with augmented architecture and circuit techniques. This complexity increase results in higher energy consumption, peak power dissipation and power delivery requirements [De and Borkar 1999].

To limit the energy and power increase in future CMOS technology generations, supply voltage will have to continually scale. The amount of energy reduction depends on the magnitude of supply voltage scaling [Chandrakasan et al. 1992]. Along with supply voltage scaling, MOS device threshold voltage will have to scale to sustain the traditional 30% gate delay reduction. This supply and threshold voltage scaling requirements pose several technology and circuit design challenges [De and Borkar 1999; Antoniadis and Chung 1991; Chen et al. 1994]. One such challenge is the increase in the variation of threshold voltage, a key process parameter, due to worsening short channel effects. This is explained in the following section.

2.1 Technology Scaling, Threshold Voltage Variation, and Leakage Power

With technology scaling, the MOS device channel length is reduced. As the channel length approaches the source-body and drain-body depletion widths, the charge in the channel due to these parasitic diodes become comparable to the depletion charge due to the MOS gate-body voltage [Poon et al. 1973], rendering the gate and body terminals to be less effective. As the band diagram illustrates in Figure 3, the finite depletion width of the parasitic diodes do not influence the energy barrier height to be overcome for inversion formation in a long channel device. Figure 3 shows cross-sectional schematic of long channel and

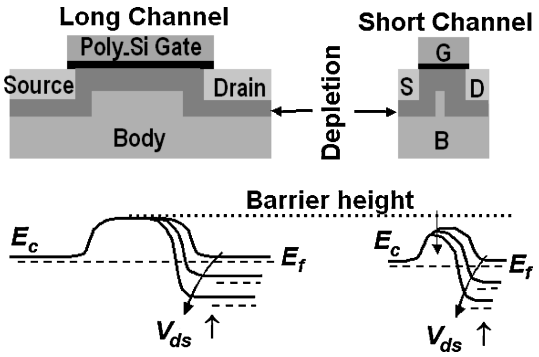


Fig. 3. Barrier height lowering due to channel length reduction and drain voltage increase in an NMOS field-effect transistor.

short channel transistors and their corresponding band conduction bands. The band diagram indicates the barrier that majority carriers in the source terminal have to overcome to enter the channel. In a given technology generation, since the source-body and drain-body depletion widths are predefined based on the dopings, the rate at which the barrier height increases as a function of distance from the source into the channel is constant. Therefore, when the channel length is reduced the barrier for the majority carriers to enter the channel also is reduced as indicated in the figure. This results in reduced threshold voltage. In other words, anytime the depletion charge between the source-body and drain-body terminals become a larger fraction of the channel length the threshold voltage reduces. For the same reason, in short channel transistor, the barrier height and therefore the threshold voltage are a strong function of the drain voltage. As the figure indicates, the barrier reduces as the drain voltage is increased. Therefore, in short channel devices, threshold voltage depends on the drain voltage.

To reiterate, as the channel length becomes shorter, both channel length and drain voltage reduce this barrier height. This two-dimensional effect makes the barrier height to be modulated by channel length variation resulting in threshold voltage variation as shown in Figure 4. The amount of barrier height lowering, threshold voltage variation, and gate and body terminal's channel control loss will directly depend on the charge contribution percentage of the parasitic diodes to the total channel charge.

Figure 5 shows measurements of threshold voltage variations for three device lengths in an 180 nm technology confirming this behavior. It is essential to mention that in nanoscale technologies variation in several physical and process parameters lead to variation in the electrical behavior of the MOS device. The discussions in this article will address variation in the electrical behavior manifested as threshold voltage variation because of process parameter variations. In addition, the threshold voltage variations addressed here are due to short channel effect in scaled MOS devices and not on threshold voltage variation due to random dopant fluctuation effect. Random dopant fluctuation effect is expected to be one of the significant sources of threshold voltage variation in devices of small area [Asenov et al. 2001].

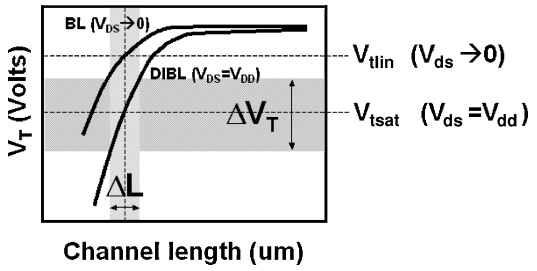


Fig. 4. Barrier lowering (BL) resulting in threshold voltage roll-off with channel length reduction. Drain induced barrier lowering (DIBL) reduces threshold voltage for short channel devices and increases threshold voltage roll-off. For short channel devices channel length variation (ΔL) translates to threshold voltage variation (ΔV_t).

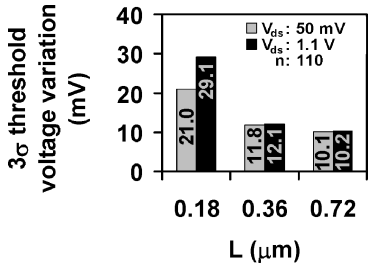


Fig. 5. Dependence of threshold voltage variation on channel length and drain voltage; n is the number of MOS device samples measured.

It was mentioned in Section 1 that in order to maintain the performance increase trend with technology scaling threshold voltage would have to be scaled along with supply voltage. However, reduction in threshold voltage increases the subthreshold leakage current significantly. Relationship between threshold voltage and subthreshold leakage is illustrated in Figure 6. Typically, reduction in threshold voltage of about 85 mV, as shown in Figure 6, will increase the subthreshold leakage current by 10X.

As indicated in Section 1, switching power increases by 1.4X per generation. With scaling of threshold voltage subthreshold leakage power will increase at a very rapid rate due to its strong dependence on the threshold voltage. Figure 7 illustrates the comparison between the increase in the switching power and subthreshold leakage power with technology scaling. As is evident from the figure, subthreshold leakage power will be comparable to the switching power in the 60–65 nm node. This "inefficient" leakage power manifests itself as active leakage that influences the total power budget during operation and as standby leakage that influences the battery life of hand-held systems. It therefore becomes important to not only reduce subthreshold leakage power but also accurately estimate it.

With supply and threshold voltage scaling, control of threshold voltage variation becomes essential for achieving high yields and limiting worst-case leakage [Sun and Tsui 1994]. Maintaining good device aspect ratio, by scaling gate

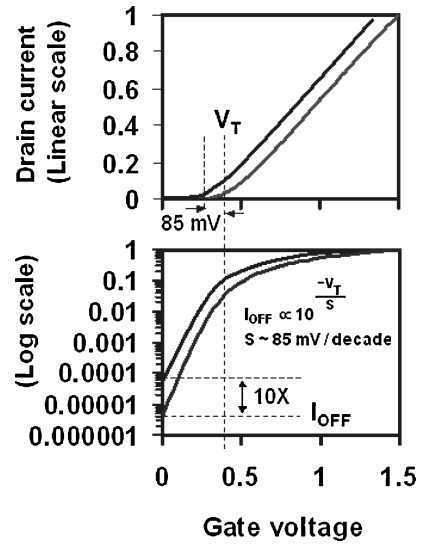


Fig. 6. Relationship between threshold voltage (V_t) and sub-threshold leakage current (l_{off}).

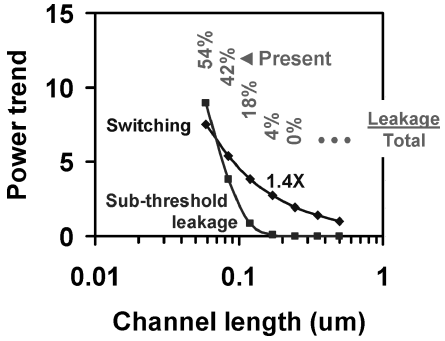


Fig. 7. Trend in subthreshold leakage and switching power with technology scaling.

oxide thickness is important for controlling threshold voltage tolerances [Taur and Ning 1998]. With the silicon dioxide gate dielectric thickness approaching scaling limits due rapid increase in gate tunneling leakage current [Muller et al. 1999; Schulz 1999] researchers have been exploring several alternatives, including the use of high-permittivity gate dielectric, metal gate, novel device structures and circuit-based techniques [Lee, C. H. et al. 2000; Mohapatra et al. 2002; Lee, J. et al. 1999; Chau et al. 2003; Kohno et al. 2000; Kuroda et al. 1996]. The use of high-permittivity gate dielectric will result in thicker and easier to fabricate dielectric for iso-gate oxide capacitance with potential for significant reduction in gate leakage. Identification of a proper high-permittivity dielectric material that has good interface states with silicon along with limited gate leakage is in progress [Lee et al. 2000]. However, it has also been shown that use of high-permittivity gate dielectric has limited return [Mohapatra et al. 2002].

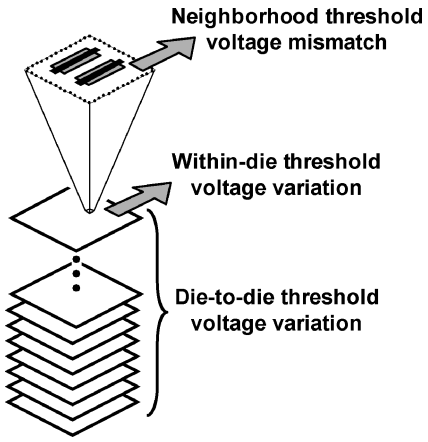


Fig. 8. Threshold voltage variation categories.

Use of metal gate prevents poly-depletion resulting in a thinner effective gate dielectric. However, identification of dual metal gates to replace the $n+$ and $p+$ doped polysilicon is essential to maintain threshold voltage scaling. In addition, novel device structures such as self-aligned double gate, FinFET, and tri-gate MOS devices provide better device aspect ratio [Lee et al. 1999; Huang et al. 1999; Chau et al. 2003]. Other than material and device based solutions, circuit design solutions such as threshold canceling logic [Kohno et al. 2000] and adaptive body bias [Kuroda et al. 1996; Miyajaki et al. 1998] enable supply and threshold voltage scaling. Threshold-canceling logic mimics threshold voltage scaling by defining the MOS off state with $|V_{gs}| > 0$, instead of $|V_{gs}| = 0$. Although threshold-canceling logic enables threshold voltage scaling, it requires larger area due to increase in logic complexity and number of power grids.

2.2 Threshold Voltage Variation Categories

The three threshold voltage variation categories illustrated in Figure 8, which impact high-performance circuit design. In general, neighborhood threshold voltage mismatch affects primarily analog circuits, within-die threshold voltage variation determines the maximum frequency (F_{max}) of operation of that particular die sample, and die-to-die threshold voltage variation determines the distribution of F_{max} for a population of dies. This article will focus on within-die and die-to-die threshold voltage variation. In Section 3 of this article we show that traditional adaptive reverse body bias circuit technique to reduce die-to-die threshold voltage variation is not scalable for future generations and additionally this technique results in increased within-die threshold voltage variation [Narendra et al. 1999]. Use of bidirectional adaptive forward and reverse body bias to limit threshold voltage variation is more promising [Miyazaki et al. 2000]. Forward body bias can be used not only to reduce threshold voltage [De 2000; Wann et al. 2000], but also to reduce die-to-die and within-die threshold voltage variations as will be shown in

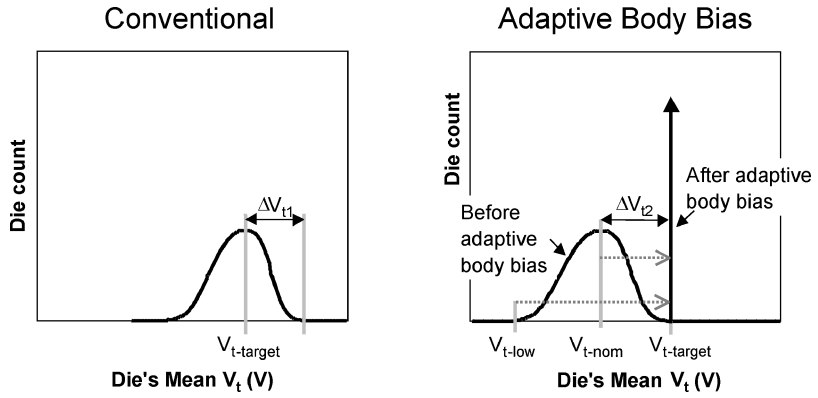


Fig. 9. Die-to-die threshold voltage distributions (a) Conventional approach without adaptive body bias (b) Adaptive body bias approach.

Section 3. Combining adaptive supply voltage with adaptive body bias is also presented.

It is important to note that threshold voltage variation not only affects supply voltage scaling but also the accuracy of leakage power estimation. Accurate leakage power estimation is very critical for future CMOS systems since the leakage power is expected to be a significant portion of the total power due to threshold voltage scaling [De and Borkar 1999]. In Section 4, leakage power estimation that takes into account within-die threshold voltage variation will be presented. In a leakage-dominant CMOS system, it also becomes inevitable to identify techniques to reduce this variation and leakage power. In Section 4, the use of different techniques to reduce system leakage will be discussed.

3. ADAPTIVE SCHEMES FOR PROCESS VARIATION

3.1 Adaptive Body Bias

Supply voltage (V_{dd}) and threshold voltage (V_t) scaling is the most effective approach to keep active power dissipation under control while maintaining performance improvement [Chandrakasan et al. 1992]. One of the limits to V_{dd} scaling is the expected increase in V_t variation [Antoniadis and Chung 1991; Sun and Tsui 1994]. Increase in die-to-die V_t variation will result in slow dies that do not meet the frequency target and fast dies that exceed the allowed power limits due to excessive leakage. The resulting reduction in yield will lead to increases in manufacturing cost and time to market, neither of which is acceptable especially with the technology life cycle shrinking from 3.6 to 2 years (Figure 1). Adaptive body bias schemes have been proposed in the past to reduce this expected increase in die-to-die V_t variation [Kuroda et al. 1996; Miyazaki et al. 1998].

Figure 9(a) illustrates that in a conventional approach without adaptive body bias the mean V_t of all the die samples do not match the target V_t. By using adaptive body bias, a sharper distribution in die-to-die V_t variation can be achieved,

as shown in Figure 9(b). Adaptive body bias first requires modification of the process so that mean V_t of all the dies are lower than the target V_t, as depicted in Figure 9(b). This lowering of V_t for a given technology is accomplished by reducing the channel doping which increases the depletion width of the MOSFET parasitic junction diodes. It was shown in Section 2.1 that this would result in increased V_t variation due to worsened short channel effect (SCE)! Therefore, $\Delta V_{t2} > \Delta V_{t1}$ in Figure 9. After this process modification, depending on the mean V_t of a die sample an adaptive amount of reverse body bias is applied to the entire die so that its mean V_t will be increased to match the target V_t, as illustrated in Figure 9(b).

Reverse body bias increases the depletion width of the MOSFET parasitic junction diodes [Tsividis 1987]. It was shown in Section 2.1 that this would result in increased V_t variation due to worsened short channel effect (SCE)! To study the effectiveness of adaptive body bias in controlling die-to-die V_t variation as technology is scaled and (2) to determine impact of adaptive body bias on within-die V_t variation. It will shown that as MOSFET technology is scaled, the body bias required for compensating die-to-die V_t variation increases, which in turn further increases SCE, and, because of this increase in SCE, within-die V_t variation becomes worse. It will also be shown that the die that requires larger body bias to match its mean V_t to the target V_t will end up with a higher within-die V_t variation. The resulting increase in within-die V_t variation due to adaptive body bias can impact clock skew, worst-case gate delay, worst-case device leakage current, total chip leakage power, and analog circuit performance. More importantly, increase in within-die V_t can also reduce the frequency of operation in high performance designs that have increasingly lesser logic stages between flip-flops [Bowman et al. 2001; Tschanz et al. 2002]. In the rest of this section, the effectiveness of adaptive body bias and within-die V_t variation due to adaptive body bias will be presented. To reiterate the point from Section 2.1, the focus of V_t variation in this atricle is due to worsening SCE with technology scaling and channel length variation.

3.1.1 *Adaptive Body Bias and Short Channel Effect (SCE).* For adaptive body bias, the V_t of the process technology has to be retargeted to be lower as shown in Figure 9. In a given technology, this is achieved by lower channel doping that will result in lower body effect to begin with. Since adaptive body bias depends on body effect to modulate V_t with reverse body bias, lowering V_t will render adaptive body bias less effective. The body effect is further reduced in short channel devices because lower V_t with reduced channel doping will increase diode depletion charge and SCE. Figure 10 illustrates the reduction in body effect due to V_t lowering in a 0.25 μm technology. For an MOS device with V_t of 0.4 V, reverse body bias of 0.6 V increased the V_t by 25%. V_t modulation for the same amount of reverse body bias reduces to less than 8% for an MOS device with V_t of 0.25 V.

Furthermore, since V_t reduction degrades short channel effect, V_t-roll-off with channel length reduction should be more for the lower-V_t device. In addition, reverse body bias will further increase the V_t-roll-off as shown in Figure 11.

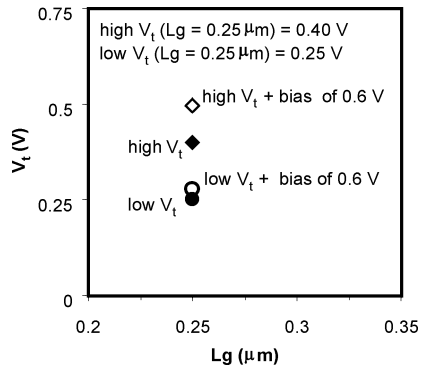


Fig. 10. Reduction in V_t modulation with reverse body bias with reduction in V_t.

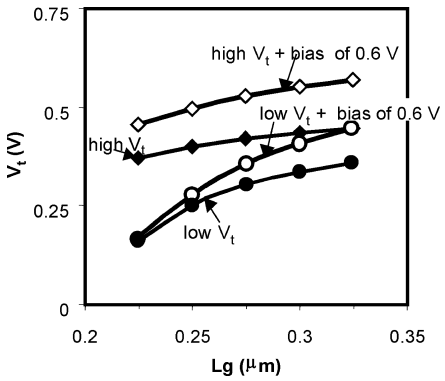


Fig. 11. Increase in V_t-roll-off with V_t reduction and reverse body bias increase.

It is known that increase in reverse body bias worsens MOSFET's short channel effect. Figure 12 shows subthreshold characteristics of a 0.25 μm NMOS device. Using Drain-Induced Barrier Lowering (DIBL), which is ΔV_t observed for a given ΔV_{ds}, as another figure of merit to indicate short channel effect, we see that increasing reverse body bias (V_{sb}) from 0 V to 2 V increases ΔV_t and hence DIBL, by 88%.

In summary, although adaptive body bias reduces die-to-die V_t variation, it increases within-die V_t variation, due to increase in short channel effects. The analysis in Narendra et al. [1999] showed that the increase in within-die V_t variation due to adaptive bias worsens with technology scaling and is more pronounced for aggressive V_t scaling. Consequently, to make effective use of the traditional adaptive body bias scheme one should consider (a) the maximum acceptable within-die V_t variation increase that can be tolerated for a given design and (b) the use of multiple adaptive bias generators within-die on a triple well process. Even if these techniques are employed to minimize impact of adaptive body bias on within-die V_t variation, adaptive body bias is still destined to become less effective with scaling due to increased SCE and weakening body effect. In addition, circuits that cannot tolerate increase in short channel effect due to reverse body bias should be isolated not to receive

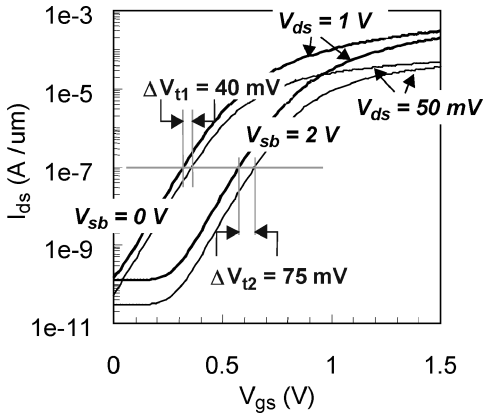


Fig. 12. Increase in DIBL due to increase in reverse body bias.

body bias. This will require triple-well process if adaptive body bias needs to applied for both PMOS and NMOS devices.

In the next section, a scheme called bidirectional adaptive body bias is introduced. This scheme does not require process modification for V_t retargeting, minimizes die-to-die V_t variation without impacting V_t within-die variation, and more importantly, its effectiveness scales better with technology compared to the traditional adaptive body bias. The bidirectional adaptive body bias scheme discussed in the next section is designed to minimize the variation in microprocessor operating frequency due to within-die and die-to-die V_t variations.

3.2 Bidirectional Adaptive Body Bias

Both die-to-die and within-die V_t variations, which are becoming worse with technology scaling, impact clock frequency and leakage power distributions of microprocessors in volume manufacturing [Bowman et al. 2001]. In particular, they limit the percentage of processors that satisfy both minimum frequency requirement and maximum active switching and leakage power constraints. Their impacts are more pronounced at the low supply voltages used in processors for mobile systems where the active power budget is limited by constraints imposed by heat removal, power delivery and battery life considerations.

In bidirectional adaptive body bias, the mean V_t of all die samples are matched to the target V_t by applying both forward and reverse body bias. Reverse body bias is a well known concept that is dealt in all introductory MOS text books, where the body-source diode is reverse biased that results in the increase in threshold voltage of the MOS device. Forward body bias is defined as a condition where the body-source diode is forward biased, which reduces the threshold voltage of the MOS device.

In bidirectional adaptive body bias, forward body bias is applied to die samples that are slower than the target and reverse body bias is applied to die samples that are faster than the target, as shown in Figure 13. It is important to note that, while forward bias reduces V_t, it also increases the junction

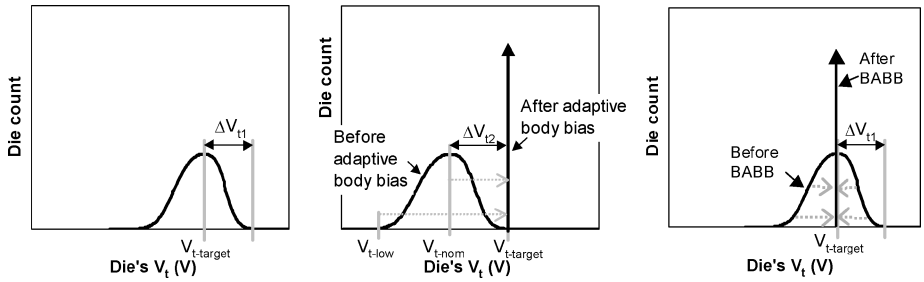


Fig. 13. Die-to-die threshold voltage distributions (a) Conventional approach without adaptive body bias (b) traditional adaptive body bias approach—die sample that requires maximum reverse body bias is $2\Delta V_{t2}$ away from $V_{t-target}$ (c) bidirectional adaptive body bias approach—die sample that requires maximum reverse body bias is ΔV_{t1} away from $V_{t-target}$. Note: $\Delta V_{t2} > \Delta V_{t1}$ since SCE of devices with lower V_t will be more.

current and junction capacitance. Hence, there is a maximum forward bias beyond which the junction current increase will inhibit proper operation of CMOS circuits. It has been determined that, at a temperature of 110°C, the maximum amount of forward bias that can be applied is 450 mV. This increases to 750 mV at an operating temperature of 30°C [Narendra et al. 2003]. Since both V_t reduction and increase are possible, process retargeting to reduce V_t is not required. By avoiding process retargeting increase in within-die V_t variation, due increase in SCE for lower V_t transistors is prevented. In addition, the die samples that forward body bias improves SCE. Since it reduces the diode depletion and hence reduces within-die V_t variation, the maximum reverse body bias required under bidirectional adaptive body bias clearly would be smaller. So, this scheme will always scale better than the traditional adaptive body bias. This technique was first reported in Miyazaki et al. [2000] as a follow-up to Miyazaki et al. [1998] and Narendra et al. [1999]. In rest of this section, further improvements over Miyazaki et al. [2000] will be presented.

A testchip (Figure 14) has been implemented in a 150-nm CMOS technology to evaluate effectiveness of the bidirectional adaptive body bias technique for minimizing impacts of both die-to-die and within-die V_t variations on processor frequency and active leakage power [Tschanz et al. 2002]. The testchip contains 21 "subsites" distributed over a 4.5×6.7 mm^2 area in two orthogonal orientations. Each subsite has (i) a circuit block (CUT) containing key circuit elements of a microprocessor critical path, (ii) a replica of the critical path whose delay is compared against an externally applied target clock frequency (ϕ) by a phase detector, (iii) a counter that updates a 5-bit digital code based on the phase detector output, and (iv) a "resistor-ladder D/A converter + op-amp driver" which, based on the digital code, provides one of 32 different body bias values to PMOS transistors in both the CUT and the critical path delay element. The circuit block diagram of each subsite is shown in Figure 15. N-well resistors are used for the D/A converter implementation. For a specific externally applied NMOS body bias, this on-chip circuitry automatically generates the PMOS body bias that minimizes leakage power of the CUT while meeting a

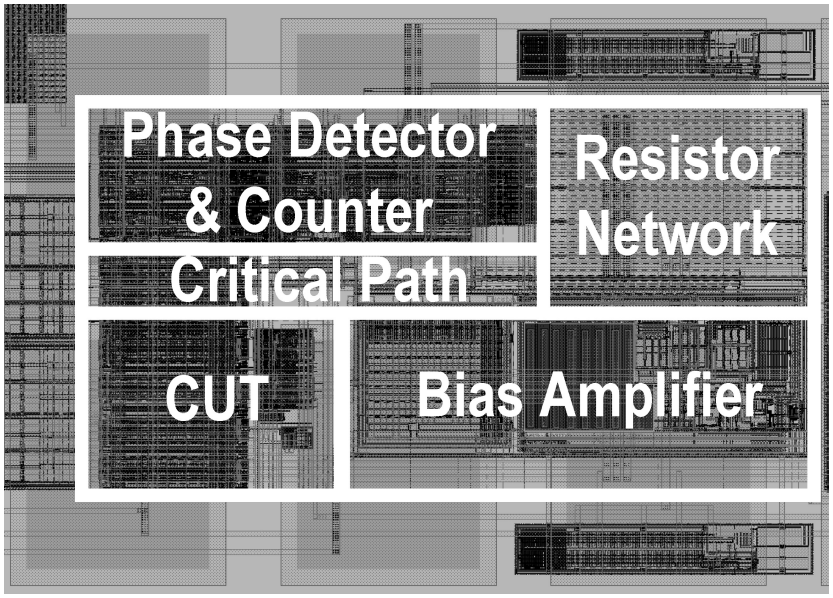


Fig. 14. Chip micrograph of a subsite.

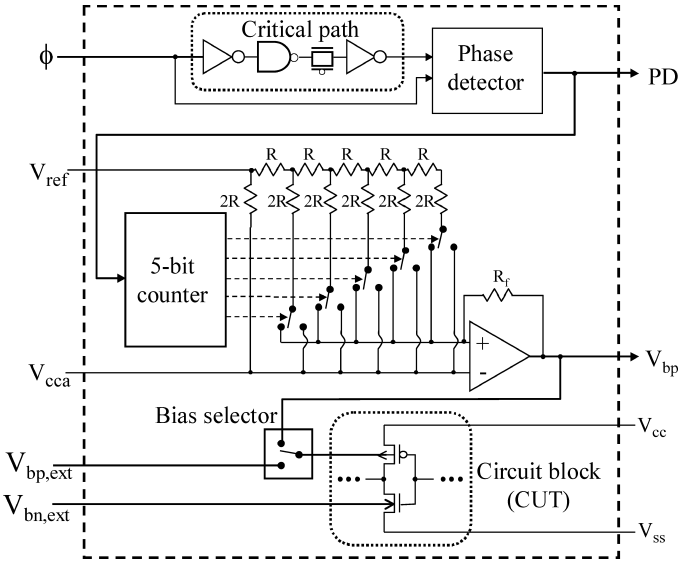
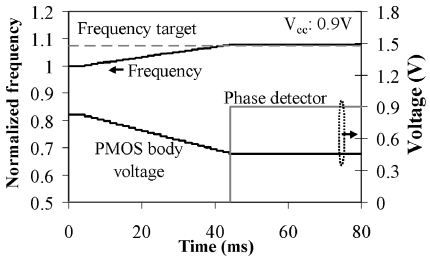


Fig. 15. Circuit block diagram of each subsite.

target clock frequency, as demonstrated by measurements in. Different ranges of unidirectional—forward (FBB) or reverse (RBB)—or bidirectional body bias values (Figure 16) can be selected by using appropriate values of V_{REF} and V_{CCA}, and by setting a counter control bit. Adaptive body biasing can also be accomplished by using the phase detector output (PD) to continually adjust



Bias Mode	Condition	Range
NBB → FBB	$V_{cca} = V_{cc}$ $V_{ref} > V_{cca}$	FBB: $0 \rightarrow V_{ref} - V_{cca}$
NBB → RBB	$V_{cca} = V_{cc}$ $V_{ref} < V_{cca}$	RBB: $0 \rightarrow V_{cca} - V_{ref}$
FBB → RBB	$V_{cca} < V_{cc}$ $V_{ref} < V_{cca}$	FBB: $V_{cc} - V_{cca} \rightarrow$ RBB: $2V_{cca} - V_{ref} - V_{cc}$

Fig. 16. Demonstration of frequency adapting to meet target and list of possible on-chip bias modes.

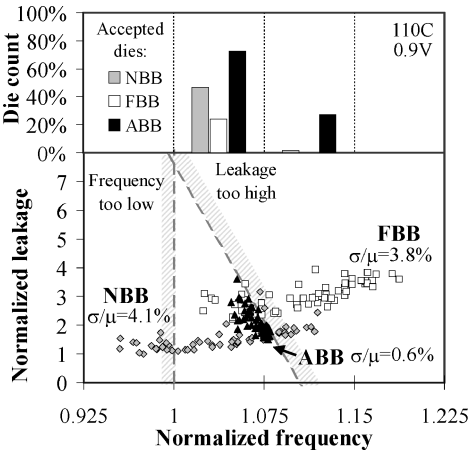


Fig. 17. Die-to-die variation in frequency and leakage for no body bias (NBB), 0.2 V static forward body bias (FBB), and adaptive body bias applied to compensate die-to-die variation (ABB).

off-chip bias generators through software control, instead of using the on-chip circuitry, until the frequency target is met.

Clock frequency, switching power and active leakage power of the 21 CUTs per die are measured independently at 0.9 V V_{dd} and 110C, for 62 dies on a wafer. Die clock frequency is the minimum of the CUT frequencies, and active leakage power is sum of the CUT leakages. When no body bias (NBB) is used, 50% of the dies meet both the minimum frequency requirement and the maximum active leakage constraint set by a total power density limit of 20 W/cm^2 (Figure 17). Using 0.2 V forward body bias (FBB) allows all of the dies to meet the minimum frequency requirement, but most of them fail to satisfy the leakage constraint. As a result, only 20% of the dies are acceptable although variations are reduced slightly by FBB due to improved short-channel effects [Miyazaki et al. 2000].

Bidirectional ABB is used for both NMOS and PMOS devices to increase the percentage of dies that meet both frequency requirement and leakage constraint. For each die, we use a single combination of NMOS and PMOS body bias values that maximize clock frequency without violating the active leakage power limit. As a result, die-to-die frequency variations (σ/μ) reduce

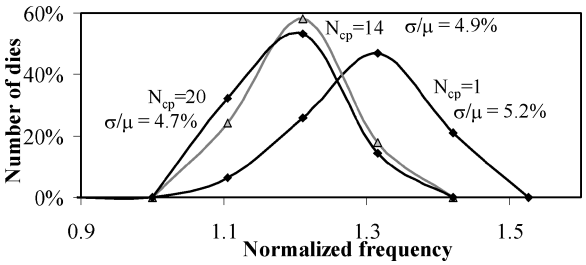


Fig. 18. Frequency vs. number of critical paths that determine the frequency.

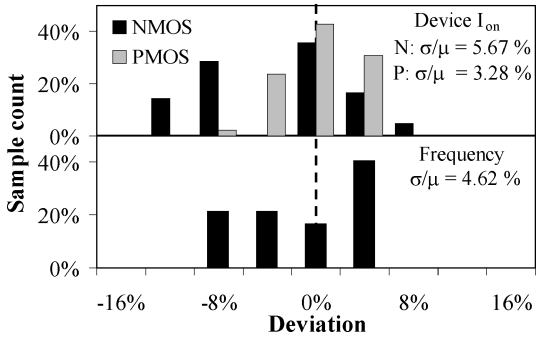


Fig. 19. Comparison of variations in within-die device current and frequency.

by an order of magnitude, and 100% of the dies become acceptable (Figure 17). In addition, 30% of the dies are now in the highest frequency bin allowed by the power density limit when leakage is negligible.

In a simpler ABB scheme, within-die variations can be neglected [Miyazaki et al. 2000] and the required body bias for a die can be determined from measurements on a single CUT per die. However, testchip measurements in Figure 18 show that as the number of critical paths (N_{cp}) on a die increases, WID delay variations among critical paths cause both μ and σ of the die frequency distribution to become smaller. This is consistent with statistical simulation results [Bowman et al. 2001] indicating that the impact of WID parameter variations on die frequency distribution is significant. As N_{cp} exceeds 14, there is no change in the frequency distribution with N_{cp}. Therefore, using measurements of 21 critical paths on the testchip to determine die frequency is sufficiently accurate for obtaining frequency distributions of microprocessors, which contain 100s of critical paths.

Previous measurements [Miyazaki et al. 2000] on 49-stage ring oscillators showed that σ of the WID frequency distribution is 4X smaller than σ of the device saturation current (ION) distribution. However, measurements on the testchip containing 16-stage critical paths (Figure 19) show that σ's of WID critical path delay distributions and NMOS/PMOS ION distributions are comparable. Since typical microprocessor critical paths contain 10–15 stages, and this number is reducing by 25% per generation [De and Borkar 1999] impact of within-die variations on frequency is becoming more pronounced. This is further evidenced by the fact that the number of acceptable dies reduces from

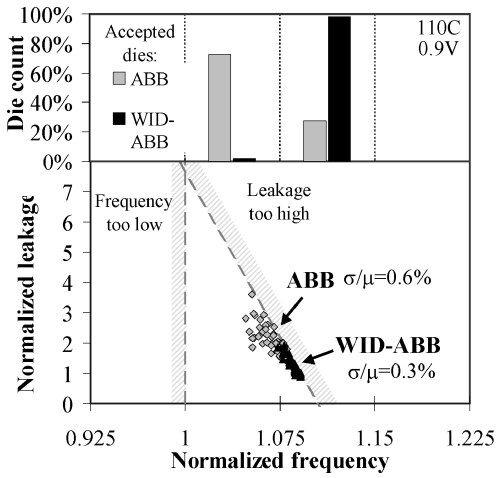
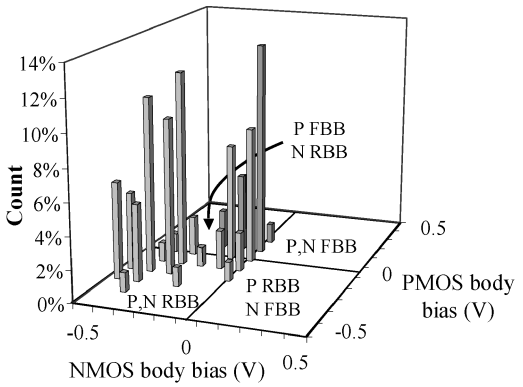


Fig. 20. Die-to-die variation in frequency and leakage for adaptive body bias applied to (i) compensate die-to-die variation (ABB) and (ii) compensate within-die variation (WID-ABB).

100% to 50% in the simpler ABB scheme which neglects within-die variations, although die count in the highest frequency bin increases from 0% to 11% when compared with NBB.

The ABB scheme, which compensates primarily for die-to-die parameter variations by using a single NMOS/PMOS bias combination per die, can be further improved to compensate for WID variations as well. In this WID-ABB scheme, different NMOS/PMOS body bias combinations are used for different circuit blocks on the die. A triple-well process is needed for NMOS implementation. For each CUT, the NMOS body bias is varied over a wide range using an off-chip bias generator. For each NMOS bias, the on-chip circuitry determines the PMOS bias that minimizes leakage power of the CUT while meeting a particular target frequency. The optimal NMOS/PMOS bias for the CUT at a specific clock frequency is then selected from these different bias combinations as the one that minimizes CUT leakage. This produces a distribution of optimal NMOS/PMOS body bias combinations for the CUTs on a die at a specific clock frequency. If the die leakage power exceeds the limit at that frequency, the target frequency is reduced and the process is repeated until we find the maximum frequency where the leakage constraint is also met.

WID-ABB reduces σ of the die frequency distribution by 50%, compared to ABB (Figure 20). In addition, virtually 100% of the dies are accepted in the highest possible frequency bin, compared to 30% for ABB. Distribution of optimal NMOS/PMOS body bias combinations for a sample die in the WID-ABB scheme reveals that while RBB is needed for both PMOS and NMOS devices, FBB is used mainly for the PMOS devices. In addition, body bias values in the range of 0.5 V RBB to 0.5 V FBB are adequate. Finally, measurements (Figure 21) show that ABB and WID-ABB schemes need at least 300 mV and 100 mV body bias resolutions, respectively, to be effective. The 32 mV bias resolution provided by the on-chip circuitry in the testchip is, therefore, sufficient for both ABB and WID-ABB.



Bias resolution	Die-to-die ABB		Within-die ABB	
	dies, F > 1	σ/μ	dies, F > 1.075	σ/μ
0.5	79 %	2.87 %	2 %	1.89 %
0.3	100 %	1.47 %	66 %	0.50 %
0.1	100 %	0.58 %	97 %	0.25 %

Fig. 21. Histogram of bias voltages within a die sample and effect of bias resolution on frequency distribution.

3.3 Adaptive Supply Voltage

Bidirectional-forward (FBB) and -reverse (RBB) adaptive body bias (V_{bs}) to reduce impacts of die-to-die and within-die (WID) parameter variations on clock frequency and active leakage power of microprocessors in volume manufacturing was discussed [Miyazaki et al. 2000; Tschanz et al. 2002] and was summarized in the previous section. In this section, we investigate the effectiveness of adaptive supply voltage (V_{dd}) and frequency binning, used individually and in conjunction with adaptive V_{bs}, for improving distributions of die frequency and power in low power and high performance microprocessors. We compare usefulness of these schemes for maximizing the percentage of dies accepted in the highest frequency bin, subject to constraints of total active power, burn-in leakage power and standby leakage power.

3.3.1 *Implementation and Measurements.* Testchip, containing 21 "subsites" or circuit blocks (CUT) distributed over a 4.5 × 6.7 mm^2 area in two orthogonal directions, has been designed and fabricated in a 1 V–1.6 V, 150 nm CMOS technology (Figures 22 and 23). Figure 23 shows two orientations of device layout in 150 nm CMOS. It should be pointed out that orientation restrictions are becoming a requirement by design rules to control process variation.

A CUT contains key circuit elements of microprocessor critical paths. Separate pads are available for V_{dd}, ground and body bias of NMOS and PMOS devices. Thus, a range of V_{dd} and V_{bs} values can be applied externally to each CUT, and their switching and leakage powers measured accurately. Frequency,

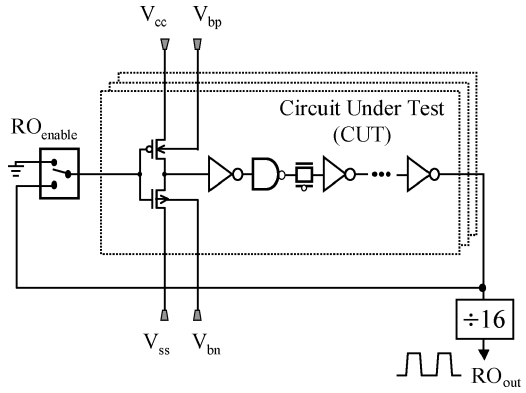


Fig. 22. Block diagram of testchip circuits for measuring critical path frequency and power.

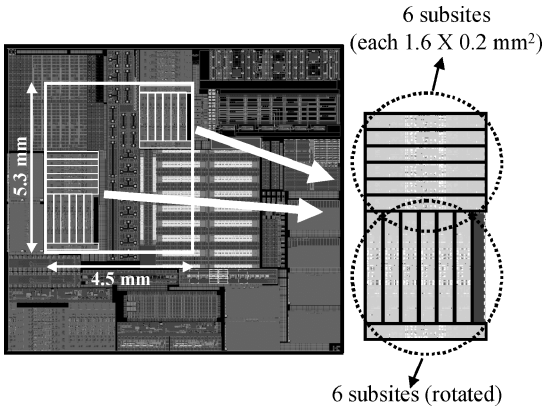


Fig. 23. Chip micrograph showing testchip subsites distributed in two regions of the die in two orientations in 150 nm technology.

power, and switched capacitance of the 21 CUTs per die are measured independently at 0.8 V–1.6 V V_{dd} and 40°C–110°C, with V_{bs} values ranging from 500 mV RBB to 500 mV FBB, for 62 dies on a wafer. Die clock frequency is the minimum of the CUT frequencies and die power is the sum of CUT powers.

3.3.2 *Effectiveness of Adaptive Supply Voltage.* Distributions of frequency, active power, switched capacitance and standby leakage power are shown in Figure 24 for a fixed 1.05 V V_{dd} and 0 V V_{bs}, chosen to maximize frequency of the median die for a power density limit of 10 W/cm^2, typical of low power microprocessors in mobile systems. We see that dies with higher frequencies have larger leakage and smaller switched capacitance. Excessive leakage causes many dies to violate constraints of active power and standby leakage power (0.5 W/cm^2). Many other dies satisfy the power constraints, but are significantly slower than the median die. Since lowering frequency reduces switching power, active power limit can be satisfied for all dies by simply moving them to a frequency bin, equal

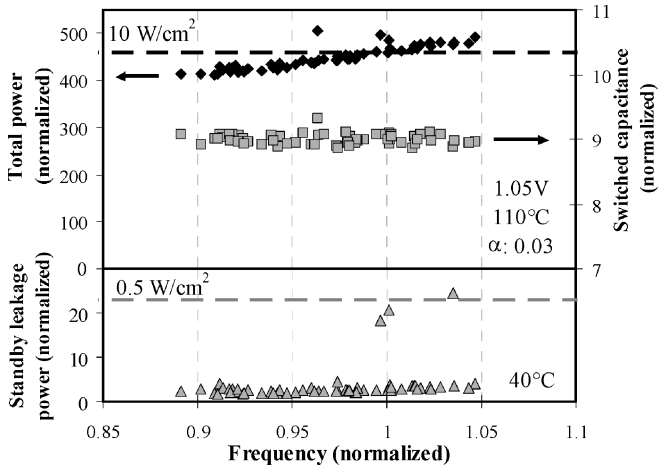


Fig. 24. Total power, switched capacitance, and standby leakage power vs. frequency. $V_{dd} = 1.05$ V for maximum frequency of median die at 10 W/cm^2 total power limit.

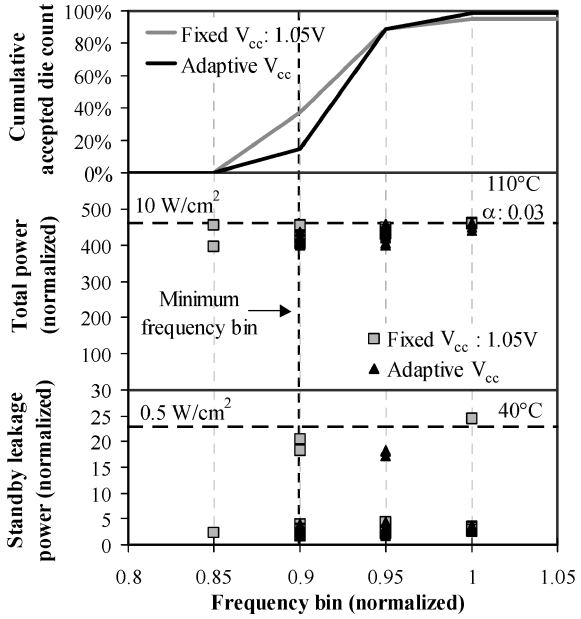


Fig. 25. Cumulative accepted die count, total power, and standby leakage power vs. frequency bin for fixed V_{dd} and adaptive V_{dd}. $V_{dd} = 1.05$ V for maximum frequency of median die at 10 W/cm^2 total power limit.

to or below their natural operating frequencies, where the total power is less than the maximum allowed.

However, some dies then fail to meet the minimum frequency requirement and still violate the standby leakage constraint. Thus, 95% of dies are accepted. In addition, 37% of dies are in the lowest frequency bin (Figure 25). Adaptive V_{dd} can be used to improve the percentage of dies accepted in higher frequency

bins. Larger V_{dd} values are used for slow dies to increase their natural operating frequency and move them to the highest frequency bin allowed by the active power limit. Gate oxide reliability considerations limit the maximum allowed V_{dd}. For dies above the power limit, V_{dd} is reduced in tandem with their natural operating frequencies so that they meet the active power constraint at a frequency bin higher than that achievable by frequency reduction alone. In contrast to simple frequency reduction, lowering V_{dd} reduces standby leakage power as well. As a result, the accepted die count improves to 98%, with only 15% in the minimum frequency bin.

For higher power density limit of 40 W/cm^2, typical of nonportable high-performance microprocessors, nominal V_{dd} of 1.5 V and 100 mV FBB are chosen to maximize frequency of the median die. Although 97% of dies are accepted when simple frequency binning is used, only 19% are in the highest frequency bin [Tschanz et al. 2003]. Imposing an additional burn-in leakage power limit of 5 W/cm^2 reduces the percentage of accepted dies to 89% [Tschanz et al. 2003]. Adaptive V_{dd} improves the accepted die count to 97%, with 44% of dies in the highest frequency bin. However, effectiveness of adaptive V_{dd} depends critically on the V_{dd} resolution. Using 50 mV V_{dd} resolution, instead of 20 mV, renders this technique ineffective for compensating across-wafer variations, but may be adequate when larger variations (across multiple wafers and lots) are considered [Tschanz et al. 2003].

3.3.3 *Combining Adaptive Supply Voltage and Adaptive Body Bias.* Using adaptive V_{dd} in conjunction with adaptive V_{bs} (adaptive $V_{dd} + V_{bs}$) is more effective than using either of them individually. In this combined scheme, a single V_{dd} and NMOS/PMOS V_{bs} combination is used per die to move it to the highest frequency bin subject to the active power limit. Adaptive V_{bs} uses FBB to speed up dies that are too slow, and RBB to reduce frequency and leakage power of dies that are too fast and leaky. Adaptive $V_{dd} + V_{bs}$, on the other hand, recovers these dies above the active power limit by (1) first lowering V_{dd} and natural operating frequency together to bring the sum total of their switching and leakage powers well below the active power limit, and (2) then applying FBB to speed them up, and move them to the highest frequency bin allowed by the active power limit.

As a result, more dies use lower V_{dd} values than adaptive V_{dd}. In addition, more dies use FBB, instead of RBB, compared to adaptive V_{bs}. Since effectiveness of RBB for leakage power reduction diminishes with technology scaling [Keshavarzi et al. 2001], adaptive V_{dd}+V_{bs} will be more effective in future technology generations. Combining adaptive V_{dd} with adaptive WID-V_{bs} can compensate for within-die variations as well. In this adaptive V_{dd}+WID-V_{bs} scheme, different NMOS/PMOS body bias combinations are used for different circuit blocks on a die, while a single V_{dd} is used for all circuit blocks. A triple-well process is needed for within-die NMOS body bias implementation, since the regular *n*-well process has all NMOS devices in a common substrate. Summary of dies in the highest frequency bins in 150 nm CMOS with 10 W/cm^2 active and 0.5 W/cm^2 standby power density limits is shown in Figure 26. Increase in number parts in the highest frequency bins were achieved by

Number of dies in F=1 and F=1.05 frequency bins	
6%	Fixed V_{dd}
10%	Adaptive V_{dd}
16%	Adaptive V_{bs}
26%	Adaptive V_{dd} + V_{bs}
80%	Adaptive V_{dd} + Within-die V_{bs}

Fig. 26. Summary of adaptive techniques and their benefits from a 150 nm testchip.

sharpening the distribution, as shown earlier, thus minimizing the impact of process variation present in nanoscale CMOS. Resolutions of supply voltage and body bias used were 20 mV and 100 mV, respectively.

4. LEAKAGE POWER

It has been established that to limit the energy and power increase in future CMOS technology generations, the supply voltage (V_{dd}) will have to continually scale. The amount of energy reduction depends on the magnitude of V_{dd} scaling. Along with V_{dd} scaling, the threshold voltage (V_t) of MOS devices will have to scale to sustain the traditional 30% gate delay reduction. These V_{dd} and V_t scaling requirements pose several technology and circuit design challenges. One such challenge is the rapid increase in subthreshold leakage power due to V_t scaling. The present scaling trend have lead to sub-threshold leakage power being as much as 40% of the total power in the 90-nm generation [Grove 2002]. Under this scenario, it is not only important to be able to reduce subthreshold leakage power, but also to be able to predict subthreshold leakage power more accurately.

It should be noted that there are other sources of leakage power—namely gate oxide, junction tunneling, and direct source-to-drain tunneling. Increase in gate oxide leakage can be addressed by switching to high-k dielectric material for the gate-to-channel insulator [Chau et al. 2004]. While junction leakage is expected to increase with scaling, reducing junction area by changing the device structure from bulk to tri-gate will help [Chau et al. 2003]. Gate-oxide and source-to-drain tunneling were expected to be the key limiters of scaling [Hoeneisen and Mead 1972]. This article we primarily focus on subthreshold leakage power, since this component is the most dominant source of leakage power in nanoscale CMOS logic design. There are two types of leakage power that is of importance (i) active leakage power and (ii) standby leakage power. Active leakage power is defined leakage power consumed by a nanoscale CMOS system when it doing useful work and standby leakage power is consumed when the system is idle. In the rest of this section, we will discuss statistical method for estimating subthreshold leakage power and techniques that help reduce active and standby leakage power.

4.1 Leakage Power Estimation

Under the present scaling scenario, it is important to be able to predict sub-threshold leakage power more accurately. Present leakage current estimation techniques do not take into account the variation in within-die threshold

voltage. It will be shown that this assumption leads to significant inaccuracies. A mathematical model for chip leakage current that considers within-die threshold voltage variation will be derived. Microprocessor measurements that verify the improvement in leakage estimation with the new model are also presented. In rest of this article, the term "leakage," unless specified, refers to subthreshold leakage.

4.1.1 *Present Leakage Current Estimation Techniques.* Due to the wide variation expected threshold voltage of MOS devices from die-to-die and within-die during the lifetime of a process, present leakage current estimation techniques provide lower and upper bounds on the leakage current. The upper and lower bounds are at least an order of magnitude apart and leakage power of most chips lies between the two bounds as shown in Keshavarzi et al. [2001]. In older technology generations, basing system design on the two leakage current bounds was acceptable since leakage power was a negligible component of the total power. In most systems, the worst-case bound is assumed for the design. In technology generations where as much as half of the system power during active mode can be due to leakage, using the worse-case bound estimation technique will lead to extremely pessimistic and expensive design solutions. One cannot base the system design on the lower bound since it will lead to overly optimistic and unreliable design solutions. Therefore, it will be crucial to estimate leakage current as accurately as possible. The upper-and lower-bound estimate equations and measurements are provided in the next part of this section. The lower-bound leakage current estimation of a chip is given as follows,

$$I_{leak-l} = \frac{w_p}{k_p} I_p^o \quad + \quad \frac{w_n}{k_n} I_n^o,$$

where w_p and w_n are the total PMOS and NMOS device widths in the chip; k_p and k_n are factors that determine percentage of PMOS and NMOS device widths that are in off state; I_p^o and I_n^o are the expected mean leakage currents per unit width of PMOS and NMOS devices in a particular chip. The mean leakage current is obtained for devices with mean threshold voltage or channel length. The upper-bound leakage current estimation of a chip is related to the device leakage as follows:

$$I_{leak-u} = \frac{w_p}{k_p} I_{off-p}^{3\sigma} \quad + \quad \frac{w_n}{k_n} I_{off-n}^{3\sigma},$$

where $I_{off-p}^{3\sigma}$ and $I_{off-n}^{3\sigma}$ are the worst-case leakage current per unit width of PMOS and NMOS devices. The worst-case leakage current is obtained for devices with threshold voltage or channel length 3σ lower than the mean leakage currents per unit width of PMOS and NMOS devices in a particular chip.

4.1.2 *Leakage Current Estimation Including Within-Die Variation* [Narendra et al. 2004]. To include the impact of within-die threshold voltage or channel-length variation, it is necessary to consider the entire range of leakage currents, not just the mean leakage or the worst-case leakage. Let us assume that the within-die threshold voltage or channel length variation, follows a normal distribution with respect to transistor width, with μ being

the mean and σ being the sigma of the distribution. Let I^o be the leakage of the device with the mean threshold voltage or channel length. Then, by performing the weighted sum of devices of different leakage, we can estimate the total leakage of the chip. This is achieved by integrating the threshold voltage or channel length distribution multiplied by the leakage, as shown below.

$$I_{leak} = \frac{I^o w}{k} \frac{1}{\sigma\sqrt{2\pi}} \int_{x_{min}}^{x_{max}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \exp\left(\frac{(\mu - x)}{a}\right) dx.$$

In the above equation, the first exponent estimates the fraction of the total width for the device leakage estimated by the second exponent. If the distribution considered within-die is threshold voltage variation, then x in the above equation represents threshold voltage and a will be equal to $n\phi_t$. If the distribution considered is channel length then x in the above equation will represent channel length and a will be equal to λ. λ can be estimated for a technology by measuring the relationship between channel length and device leakage. In the rest of this section, we will assume that the distribution of interest is the channel length, since this parameter is used to characterize a technology. The derivation of the chip leakage is then given as follows:

$$\begin{aligned} I_{leak} &= \frac{I^o w}{k} \frac{1}{\sigma\sqrt{2\pi}} \int_{l_{min}}^{l_{max}} \exp\left(\frac{-(l-\mu)^2}{2\sigma^2}\right) \exp\left(\frac{(\mu - l)}{\lambda}\right) dl \\ &= \frac{I^o w}{k} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{\sigma^2}{2\lambda^2}\right) \int_{l_{min}}^{l_{max}} \exp\left(\frac{-(l-\mu)^2}{2\sigma^2}\right) \exp\left(\frac{(\mu - l)}{\lambda}\right) \exp\left(\frac{-\sigma^2}{2\lambda^2}\right) dl \\ &= \frac{I^o w}{k} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{\sigma^2}{2\lambda^2}\right) \int_{l_{min}}^{l_{max}} \exp\left(-\left[\frac{l-\mu}{\sqrt{2}\sigma} + \frac{\sigma}{\sqrt{2}\lambda}\right]^2\right) dl. \end{aligned}$$

Let

$$t = \left[\frac{l-\mu}{\sqrt{2}\sigma} + \frac{\sigma}{\sqrt{2}\lambda}\right] \Rightarrow dl = \sqrt{2}\sigma\, dt$$

$$\therefore I_{leak} = \frac{I^o w}{k} \frac{1}{\sqrt{\pi}} \exp\left(\frac{\sigma^2}{2\lambda^2}\right) \int_{\frac{l_{min}-\mu}{\sqrt{2}\sigma} + \frac{\sigma}{\sqrt{2}\lambda}}^{\frac{l_{max}-\mu}{\sqrt{2}\sigma} + \frac{\sigma}{\sqrt{2}\lambda}} \exp -([t]^2)\, dt\ .$$

The integral can be rewritten as

$$\begin{aligned} I_{leak} &= \frac{I^o w}{2k} \exp\left(\frac{\sigma^2}{2\lambda^2}\right) \left[\frac{2}{\sqrt{\pi}} \int_o^{\frac{l_{max}-\mu}{\sqrt{2}\sigma} + \frac{\sigma}{\sqrt{2}\lambda}} \exp(-t^2) dt - \frac{2}{\sqrt{\pi}} \int_o^{\frac{l_{min}-\mu}{\sqrt{2}\sigma} + \frac{\sigma}{\sqrt{2}\lambda}} \exp(-t^2) dt\right] \\ &= \frac{I^o w}{2k} \exp\left(\frac{\sigma^2}{2\lambda^2}\right) \left[\mathrm{erf}\left(\frac{l_{max}-\mu}{\sqrt{2}\sigma} + \frac{\sigma}{\sqrt{2}\lambda}\right) - \mathrm{erf}\left(\frac{l_{min}-\mu}{\sqrt{2}\sigma} + \frac{\sigma}{\sqrt{2}\lambda}\right)\right] \\ &\quad \therefore \mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_o^z \exp(-t^2) dt \\ &= \frac{I^o w}{2k} \exp\left(\frac{\sigma^2}{2\lambda^2}\right) \left[\mathrm{erf}\left(\frac{l_{max}-\mu}{\sqrt{2}\sigma} + \frac{\sigma}{\sqrt{2}\lambda}\right) + \mathrm{erf}\left(\frac{\mu - l_{min}}{\sqrt{2}\sigma} - \frac{\sigma}{\sqrt{2}\lambda}\right)\right] \\ &\quad \therefore \mathrm{erf}(-z) = -\mathrm{erf}(z). \end{aligned}$$

Since

$$\text{erf}(z) \to 1 \quad \text{if} \quad z > 1 \quad \text{and} \quad \frac{l_{max} - \mu}{\sqrt{2}\sigma} + \frac{\sigma}{\sqrt{2}\lambda}, \quad \frac{\mu - l_{min}}{\sqrt{2}\sigma} - \frac{\sigma}{\sqrt{2}\lambda} \gg 1$$

$$\Rightarrow I_{leak} = \frac{I^o w}{k} \exp\left(\frac{\sigma^2}{2\lambda^2}\right).$$

Using the above result, we can now estimate the leakage of a chip that has both PMOS and NMOS devices including within-die variation as follows:

$$I_{leak-w} = \frac{I_p^o w_p}{k_p} \exp\left(\frac{\sigma_p^2}{2\lambda_p^2}\right) + \frac{I_n^o w_n}{k_n} \exp\left(\frac{\sigma_n^2}{2\lambda_n^2}\right),$$

where w_p and w_n are the total PMOS and NMOS device widths in the chip; k_p and k_n are factors that determine percentage of PMOS and NMOS device widths that are in off state; I_p^o and I_n^o are the expected mean leakage currents per unit width of PMOS and NMOS devices in a particular chip; σ_p and σ_n are the standard deviation of channel length variation within a particular chip; λ_p and λ_n are constants that relate channel length of PMOS and NMOS devices to their corresponding subthreshold leakages. It is also worth pointing out that from the formula for I_{leak}, if I_{leak} can be measured for a chip, a macroscopic standard deviation (σ) representing parameter variation in that chip can be determined as,

$$\sigma = \lambda \sqrt{2 \ln\left(\frac{k}{w} \frac{I_{leak}}{I^o}\right)}.$$

4.1.3 *Measurement Results.* Leakage power measurements on several samples of an 180-nm 32-bit microprocessor were carried out. The current and effective channel length measurements on test devices that accompany each microprocessor were measured to determine I_p^o, I_n^o, λ_p, and λ_n. σ_p and σ_n were assumed as a constant percentage of the measured channel length in the test device of each sample. Using these individual device measurements, with w_p and w_n obtained from the design, the leakage power was calculated using the I_{leak-l}, I_{leak-u}, and I_{leak-w} formulas. In addition, we assumed that on an average half of the devices will be in off state, that is, $k_p = k_n = 2$. The three calculated leakages are then compared with the measured leakage.

Figure 27(a) clearly illustrates that the upper-bound technique overestimates the leakage current of the chips while the lower bound techniques underestimates the leakage current. However, the estimation technique explained in this section that includes within-die variation matches the measurement better, as illustrated in Figure 27(b). Data shown in Figure 27 is summarized in Figure 28. As the figure indicates, the leakage power for most of the samples are underestimated by 6.5X if the lower-bound technique is used and overestimated by 1.5X if the upper-bound technique is used. The measured-to-calculated leakage ratio for majority of the device samples is 1.04 for the technique described in this section. The calculated leakage is within ±20% of the measured leakage for more than 50% of the samples, if the new I_{leak-w} technique is used. Only

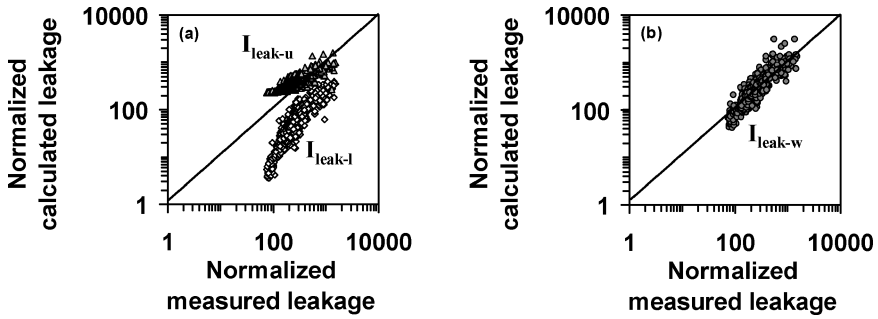


Fig. 27. Comparison of calculated leakage versus measured leakage for (a) existing leakage current estimation techniques and (b) leakage current estimation technique explained in this section.

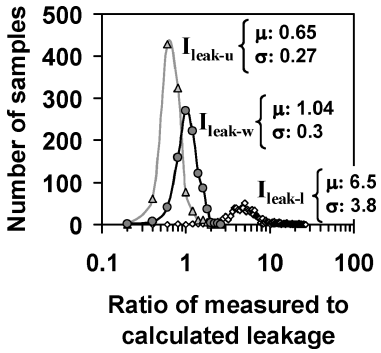


Fig. 28. Ratio of measured to calculated leakage current ratio distribution for I_{leak-u}, I_{leak-l}, and I_{leak-w} techniques (Sample size: 960).

11% and 0.2% of the samples fall into this range for the I_{leak-u} and I_{leak-l} techniques respectively. I_{leak-w} technique can be used to predict chip level leakage with better accuracy once device level leakage, parameter variation, and total transistor widths are known.

4.2 Leakage Power Reduction

Developing circuit techniques to reduce subthreshold leakage currents in both the active and standby periods to minimize total power consumption is imperative for taking advantage of nanoscale CMOS. Standby leakage currents are especially wasteful in burst mode systems, such as mobile devices, where circuits spend a significant portion of the time in an idle mode where no computation takes place. A large number of circuit techniques have been developed to turn off these leakage currents when performance is not needed. As technology has continued to scale, subthreshold leakage currents have become so large that they must be controlled during the active state as well.

In general, there are three main approaches to control subthreshold leakage currents: source biasing including stack effect, direct V_t manipulation, and power gating. In source biasing, the main idea is to bias the source terminal of an "off" transistor in order to exponentially reduce the leakage currents of that device. The second way to lower subthreshold leakage currents is to

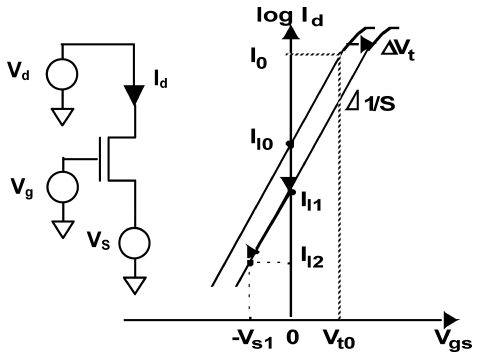


Fig. 29. Source biasing effect concept.

directly adjust the V_t's of transistors within a circuit. This can be accomplished by using a multiple threshold voltage process where a combination of low and high V_t devices achieved by process design can be used to select between high performance and low leakage requirements. This is accomplished by using low V_t in devices that are timing critical and high V_t in devices that are not timing critical. Since not all paths are timing critical the expectation is that one can achieve low V_t performance with about 2-3X lower leakage that using all low V_t [Wei et al. 1999]. Alternatively, a variable threshold voltage technology, such as body biasing, could also be used to explicitly alter the threshold voltage. Finally power gating is a technique where the power supply to the core is withdrawn by turning off a series switch to either V_{dd} or GND.

Using multiple V_t by process to combat leakage without performance degradation reduces both active and standby leakage powers. Modern process technologies allow V_t assignment at the transistor level by using different dopings and/or channel lengths. However, the main challenge in effectively using this technique is the ability to determine accurately which paths are critical and which paths are not, in design phase. This has proven to be challenging in the presence of process variation. In general all the other techniques are inherently standby leakage reduction techniques. They can be extended to active leakage reduction by enabling or disabling the scheme. Associated with enabling and disabling are (i) multiclock cycle time constants that require co-designing at the circuit and architecture levels and (ii) energy overheads that require minimum time to stay in leakage power saving state below which no net energy is saved.

It is worth noting that power gating is a technique that reduces all sources of leakage including gate and junction tunneling currents, since the power is turned off.

4.2.1 *Source Biasing.* The source biasing concept is illustrated in Figure 29, where a positive bias is applied during the standby state to the source terminal of an off device. The subthreshold leakage current is significantly reduced because the body effect tends to raise the threshold voltage of the device, and also the V_{gs} voltage becomes negative, which more strongly turns off the transistor. The net effect is that the off device is turned off more strongly and leakage currents can be reduced during standby modes. This source biasing

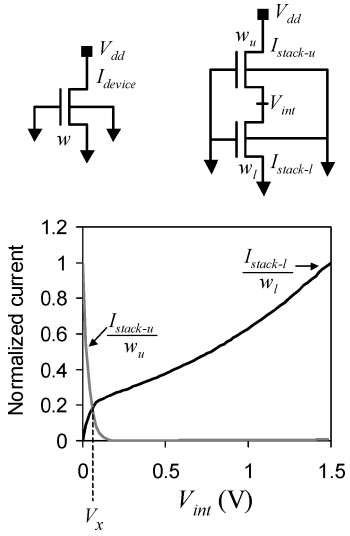


Fig. 30. Leakage reduction due to stack effect for series off devices.

principle is the underlining mechanism for several different standby leakage reduction schemes. The switched source impedance concept [Horiguchi et al. 1993] is a special case where a degenerating resistor is used to generate the biased source voltage. For high performance, the degenerating resistor is bypassed to ground, but during the standby state, the resistor is used to bias the source terminal of the off device. Another variation known as self-reverse biasing [Kawahara et al. 1993; Sakata et al. 1994] replaces the switched source impedance with another off transistor so that the equilibrium value is set through a series of "off devices." This technique was first applied to decoded word line driver circuits where the large drivers can have large leakage currents.

4.2.1.1 *Stack Effect.* A final example of the source biasing principle is illustrated by efficiently using transistor stacks within the logic gates themselves to control leakage [Chen et al. 1998; Ye et al. 1998]. In stack effect, two series-connected off transistors will have much lower leakage currents compared to a single off device due to self-reverse biasing effects. For example, Figure 30 shows leakage difference between one off devices versus two series off devices.

For the case where the two series devices are both turned off, the leakage currents are much smaller since the internal series node causes V_{gs} for the upper device to become reverse biased. Another way to interpret the leakage is that V_{ds} of lower device will be V_x, which is much smaller than that of the single off device whose V_{ds} will be V_{dd}. Modeling of leakage reduction factor, X, due to a stack of two series off devices can be expressed based on fundamental technology parameters [Narendra et al. 2001] as,

$$X = 10^{\frac{\lambda_d V_{dd}}{S}\left(\frac{1+\lambda_d}{1+2\lambda_d}\right)} = 10^U,$$

where U is the universal two-stack exponent which depends only on the process parameters, DIBL (λ_d) and subthreshold swing (S), and the design parameter,

V_{dd}. Once these parameters are known, the reduction in leakage due to a two-stack can be determined from the above model. It is essential to point out that the model assumes the intermediate node voltage to be greater than 3kt/q. In the above equation we assumed $w = w_u = w_l$. A more generic equation can be found in Narendra et al. [2001].

If an appropriate vector can be clocked into a logic block during the standby state, then leakage currents can be reduced by maximizing the number of series connected off paths. Associated with using this concept for active leakage reduction is the time constant required in leakage convergence when the vector is clocked in and the time constant associated with connecting the inputs back to the regular data path. Fortunately, the time constant for leakage convergence is proportional to the leakage current and therefore has the desired trend.

In Narendra et al. [2001], this idea was extended to actually inserting extra series off devices into single stack paths. This provides moderate leakage reduction while using a standard single threshold voltage technology. The difficulties with this approach is in developing the proper CAD tools to identify single stack paths with enough slack such that inserting extra series devices will not impact performance and the slack uncertainty due to process variation.

4.2.2 *Direct V_t Manipulation*

4.2.2.1 *Multiple V_t.* Subthreshold leakage currents are exponentially dependent on device V_t, and can be changed by several orders of magnitude by switching between high and low threshold voltages. In many modern processes, multiple threshold voltage devices are readily available to the circuit designer. A dual V_t process, for example, requires an extra mask layer to select between high- and low-threshold voltages, which provide the designer with transistors that are either fast (but high leakage) or slow (but low leakage).

A straightforward way to take advantage of these modern technologies is through a dual V_t partitioning algorithm. A circuit can be partitioned into high and low threshold voltage gates or transistors, which will tradeoff between performance and increased leakage currents. For instance, critical paths within a circuit should be implemented with low V_t to maximize performance, while noncritical paths should be implemented with high V_t devices to minimize leakage currents. By using fast leaky devices only when necessary, leakage currents can be significantly reduced in both the standby and active modes compared to an all low V_t implementation.

Dual V_t partitioning is a popular leakage reduction technique because the circuit operation remains the same as for a single V_t implementation, yet critical parts of a circuit can use scaled V_t devices to maintain performance at low supply voltages. Lee et al. [1997] and Yamashita et al. [2000] discuss some examples that illustrate the effectiveness of this technique.

Practically, there are limitations to the use of dual V_t partitioning to reduce leakage currents. In many optimized designs, there are many critical delay paths. Therefore, a large fraction of all paths in the circuit must be implemented with low V_t devices, which reduces the effectiveness of this technique.

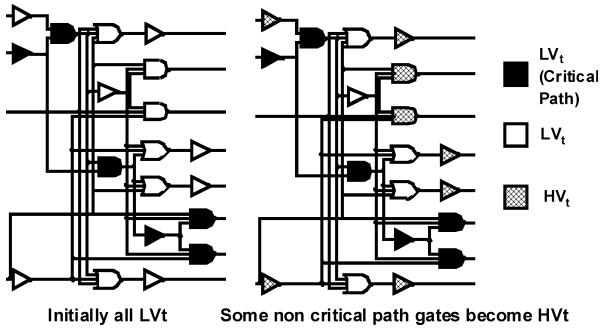


Fig. 31. Dual V_t partioning. Only some noncritical path gates can be made high V_t.

Another limitation is that CAD tools must be developed and integrated into the design flow to help optimize the partitioning process. It is not straightforward to identify which gates can be made high and low V_t without changing the delay profiles of the circuit. For example, one partitioning scheme that can be applied to random combinational logic is to first implement the circuit with all low V_t devices to ensure the highest possible performance, and then to selectively implant non critical gates to be high V_t. However, noncritical gates which are converted to be high V_t devices can become critical gates as illustrated in Figure 31. Additionally, due to process variation, that is not well understood prior to manufacturing, assumption on critical and noncritical paths may end up being incorrect post manufacturing.

Significant research is still required to improve multiple V_t algorithms, especially in the presence of process variation. There exists a natural limit where dual V_t partitioning may not reduce standby leakage currents enough for ultra low power, high performance applications. As a result, other leakage reduction techniques are important.

4.2.2.2 *Variable V_t.* Variable V_t CMOS, or VTCMOS, is another technique that has been developed to reduce standby leakage currents. Rather than employ multiple threshold voltage process options, VTCMOS relies on biasing the body terminal of the device to dynamically adjust the V_t. By applying maximum reverse biasing during the standby mode, the threshold voltage is shifted higher and subthreshold leakage current reduced. The VTCMOS principle was first proposed in Kuroda et al. [1996].

Reverse body bias, as explained in Section 3.1, does not scale well in nanoscale CMOS [Keshavarzi et al. 2001]. Under the same principle, forward body bias controls the channel better and therefore modulates the threshold voltage better. Figure 32 summarizes the different body bias in nanoscale PMOS. As is evident from Figure 32, due to degradation in barrier lowering with reverse body bias, the ability to control the threshold voltage reduces. This effect is opposite with forward body bias.

It is necessary to point out that there exists an optimal forward bias beyond which the increase in current and capacitance due to the parasitic diodes become detrimental. Figure 33 illustrates that the optimal bias is 500 mV at

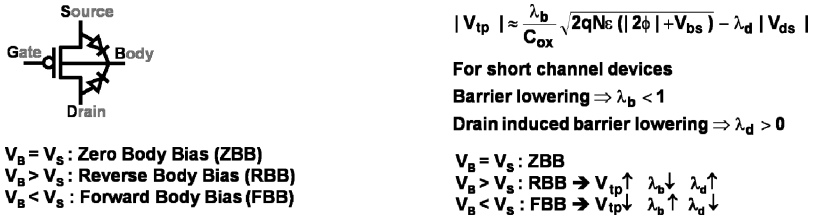


Fig. 32(a). Variable V_t CMOS using forward and reverse body bias.

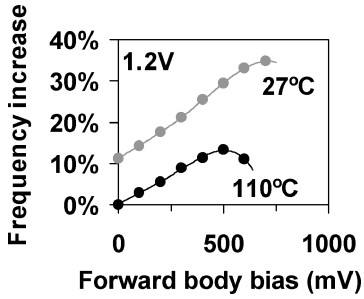


Fig. 32(b). Optimal forward body bias values.

110°C and 750 mV at 27°C. The optimal bias values are fairly independent of technology generation and supply voltage. On the other hand, optimal reverse body bias values are dependent on technology generation [Keshavarzi et al. 2001].

Due to these reasons, in nanoscale CMOS, applying forward body bias in operation mode to achieve the target frequency of operation and withdrawing forward body bias in standby mode reduces leakage power is more effective than the original VTCMOS concept. Using reverse body bias in standby in conjunction with forward body bias in active mode provides the maximum modulation of leakage power. Scaling challenges with reverse body bias and benefits of forward body bias are summarized in Keshavarzi et al. [2001] and Narendra et al. [2003], respectively.

4.2.3 *Power Gating.* Power gating technique in its original form was called Multi-Threshold CMOS (MTCMOS). MTCMOS is a dual V_t technique that is extremely effective at reducing standby leakage currents at the expense of performance penalty [Mutoh et al. 1995]. The basic principle is illustrated in Figure 33 where a low V_t computation block is gated with high V_t power gating switches. Performance penalty of this technique arises from the fact that there is voltage drop across the power gating transistors, resulting in smaller supply voltage for the logic block.

The high V_t switches are used to disconnect the power supplies during the standby state, resulting in very low leakage currents set by the high-threshold voltage of the series transistor. During the active state, the high V_t switches are turned on, and the internal logic transition through fast low V_t devices. Although both PMOS and NMOS power gating transistors are shown in Figure 33, only one polarity sleep device is actually required to reduce leakage current.

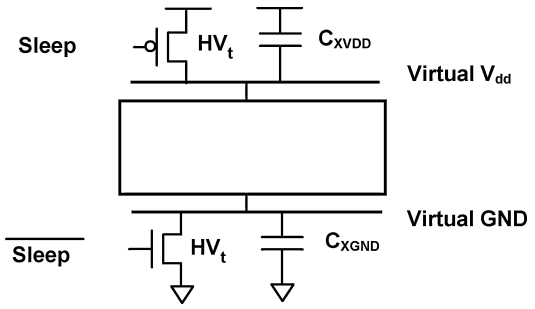


Fig. 33. MTCMOS—The first power gating concept.

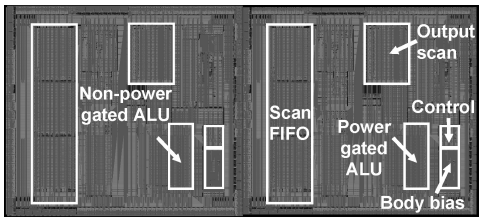
It is important in power-gated circuits to determine the proper sizing of the gating transistor. This device needs to be large enough to maintain performance, yet still be area efficient and incur minimum energy overhead when switching between modes. Because these transistors can be large, it is important to develop CAD tools to efficiently size these devices to ensure functionality and to minimize costs [Kao and Chandrakasan 2000].

The optimal way to size power-gating transistors though is to identify the worst case critical path to size the power-gating transistor. Unfortunately, identifying the worst case path is not straightforward because the delay not only depends on a signal propagating through a datapath, but also on how all other noncritical gates switch and contribute to the virtual ground or power bounce.

It is important to develop CAD tools to help efficiently size power-gating transistors and to trade off between minimizing silicon area overhead and increasing design complexity. Kao et al. [1998] describes a preliminary sizing algorithm that ensures that the power-gated circuit will always meet performance constraints without performing an exhaustive search of all possible input vectors. This technique relies on dividing a power-gated block into smaller mutual exclusive pieces that can be simulated more easily, and then merged together to determine the required power-gating transistor width.

In most cases, the ratio of device width in the power dating device to the power gated block is small enough to justify using low-V_t for the power gating-devices. In this case, the majority of leakage power savings comes from the smaller leaking width and stack effect. Using low V_t device for the power-gating transistor has the potential to make it smaller for given performance penalty, thereby reducing the power overhead in enabling and disabling of these transistors [Tschanz et al. 2003]. Tschanz et al. [2003] refers to power gating transistors as sleep transistors.

4.2.4 *Case Study: Active Leakage Power Reduction* [Tschanz et al. 2003]. Clock gating is used in high-performance microprocessors to reduce average active power and energy consumptions [Kurd et al. 2001]. Disabling the clock to idle functional units saves power by preventing wasteful switching power dissipation in the local clock distribution network and sequentials during the idle period. However, with technology scaling, leakage power of idle units is



Technology	130nm dual-V_T CMOS
Die size	1.61 x 1.44 mm^2
Number of transistors	160k

Fig. 34. Dynamic power gating (sleep transistors) and body bias case study.

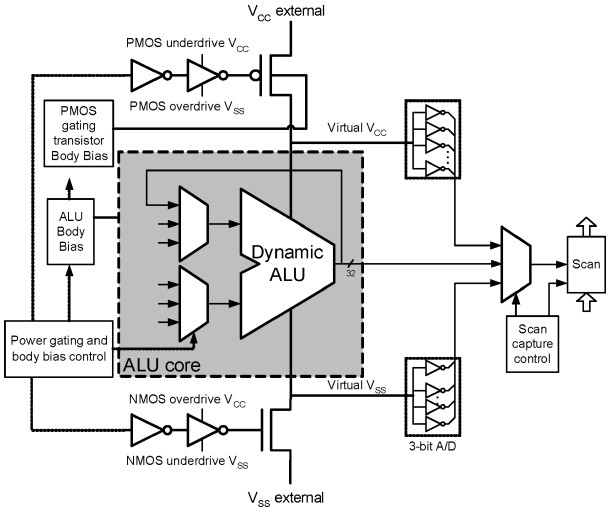


Fig. 35. Block diagram of ALU with power gating (sleep) transistors, body bias, and control circuitry.

becoming a large fraction of the total chip power. As a result, the overall power savings achievable by clock gating alone is diminishing. Using dynamic power-gating and body bias techniques, used in conjunction with clock gating can help to control the active leakage power of an ALU in a 32-bit integer execution core [Vangal et al. 2002]. Performance impacts, area overheads, active leakage and total power reductions achievable by these techniques are measured on the prototype chip (Figure 34) implemented in a 130 nm dual-V_t CMOS technology.

The 32-bit, 2-phase domino, Han–Carlson adder design contains NMOS and PMOS power gating transistors, inserted between the virtual and real supply grids on chip (Figure 35). The power gating transistors are ON during active mode and are turned OFF during the idle phase along with the local clock. To further improve on-resistance and leakage reduction capabilities of the power gating transistors we use combinations of (1) gate overdrive during "active" and underdrive during "idle", and (2) forward body bias (FBB) during "active" and zero body bias (ZBB) during "idle". Low-V_t devices are used for the power-gating transistors to minimize performance and area impacts. The

Reference case: adder without power gating transistors, 450mV FBB (1.28V, 75ºC, 4GHz)	Frequency change	Leakage reduction	Leakage reduction including decap oxide leakage	Area overhead
PMOS power gating transistor	-2.3%	37X	13.3X	11%
PMOS power gating transistor with 200mV overdrive/underdrive	-1.8%	44X	13.8X	12%
PMOS power gating transistor with 450mV FBB/ 500mV RBB	-1.8%	64X	15.5X	12%
Adder with PMOS FBB during active mode & ZBB during idle	0%	1.9X	1.8X	2%

Fig. 36. Adder frequency and leakage with PMOS power gating (sleep) transistor, compared to no power gating transistor.

power-gating transistors are distributed uniformly across the adder layout to prevent undesirable current crowding in the power grids. M4 and M5 metal levels in the traditional power grid are reconnected through the power-gating transistors instead of M4/M5 vias. The virtual supply transition waveform, following power gating transistor turn-off, can be measured directly using an on-chip 8-level (3-bit) A/D converter that uses inverters of varying precharacterized trip points. Dynamic body bias for the adder core, to apply FBB during "active" and ZBB during "idle", is also implemented.

High-frequency testing is accomplished using a scan methodology for the ALU input vectors and at-speed output capture. Measurement of the virtual supply transition waveform, following power-gating transistor turn-off, is done by capturing outputs of the two 3-bit A/D converters after a preset number of clock cycles. Separate supply pins are provided for the buffers driving the power-gating transistors so that switching and leakage power of the adder can be measured independently of the energy required to switch the power-gating transistors between active and idle.

The adder operation frequency, without power gating transistors, ranges from 3.3 GHz (1.1 V) to 4.3 GHz (1.4 V) at 75°C with FBB applied to the adder core. Using PMOS power gating transistors degrade frequency by 2.3%, with an associated area overhead of 11%. Idle leakage power is 13X smaller at 75°C (Figure 36). Gate oxide leakage through the MOS decoupling capacitors (decap) on the real supply grid cannot be reduced by power-gating transistor. The adder leakage power reduction is 37X when the decap leakage is excluded. When 200 mV gate overdrive and underdrive are used, the frequency loss is 1.8%, but leakage power savings remain largely unchanged because the decap leakage is significant. Using 450 mV FBB for the PMOS power-gating transistors reduces frequency impact from 2.3% to 1.8%. In contrast to the power-gating transistor technique, using dynamic body biasing for PMOS devices in the adder core reduces leakage power by 1.8X. The area overhead for body bias generators and bias grid routing is 2%. While body biasing is useful for controlling subthreshold leakage only, power-gating transistors help reduce gate oxide, junction and subthreshold leakage components of the leakage power. However, logic state of the adder is lost during the "idle" period when power-gating transistors are turned off, but is preserved when dynamic body biasing is used.

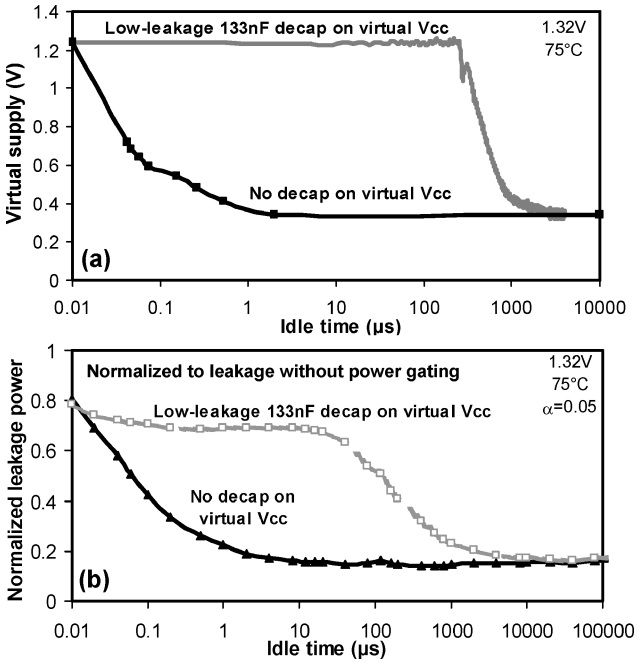


Fig. 37. (a) Convergence of virtual VCC as PMOS power gating (sleep) transistor is turned off. (b) Effect of virtual supply capacitance on time constant and leakage savings.

Overall performance impact of these dynamic leakage control techniques is dictated by the time required to switch the power-gating transistor gate or the body node voltage during "idle" to "active" transition. Measurements and simulations show that this time is 1 clock cycle for power-gating transistors and 3–4 cycles for body bias, compared to less than a cycle for clock gating. For any of these dynamic leakage control techniques to achieve reduction in overall power, the leakage energy saved during the "idle" period must be larger than any energy overhead incurred during transitions between "idle" and "active" modes. For power-gating transistors, energy is consumed during active-idle transitions due to discharging and charging of the capacitances at the virtual supply and internal circuit nodes by adder leakage currents as they converge to steady-state values. This "leakage convergence" time is measured to be 1 μs at 75°C (Figure 37(a)) and increases to 10 ms when an external 133 nF decap is added to the virtual supply. While adding low-leakage decap to the virtual supply alleviates the performance impact of power gating transistors, the transition energy overhead becomes larger. As a result, the overall leakage power savings for 10-μs idle time is only 30%, instead of 84% (Figure 37(b)). However, if the leaky MOS decap, which accounts for 10% of the adder area, is moved from the real supply to the virtual supply, decap leakage is virtually eliminated during idle period since the virtual supply collapses to 0.4 V. Then, overall leakage power savings by power-gating transistor improves from 80% to 90%, in spite of the additional transition energy overhead arising from extra capacitance at the virtual supply (Figure 38(a)). The idle leakage power also depends on

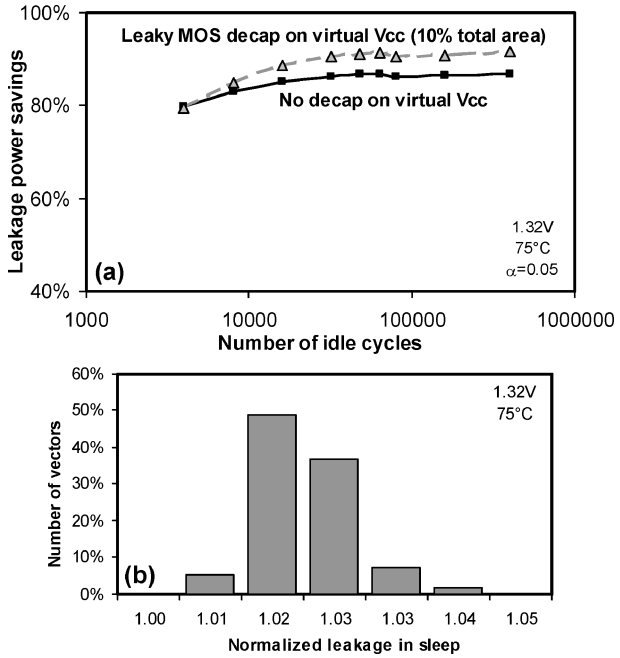


Fig. 38. (a) Leakage savings as a function of decoupling capacitor placement. (b) Dependence of power gated leakage on adder input vector. Leakage measured after 100 ns in sleep mode.

the input vector and clock mode of the domino adder (Figure 38(b)). Therefore, effectiveness of these leakage control techniques can be improved by loading, during active to idle transition, an input vector or clock mode that provides the smallest leakage power.

The minimum "idle" time required to achieve overall power saving is also dictated by the energy spent in switching the power-gating transistor gates and body nodes. We compare total active power of these techniques at 4-GHz clock frequency where higher supply voltage (1.32 V) is used for the power-gating transistor to meet the frequency target. Power measurements for different adder activity profiles, including switching energy overheads, show that the minimum "idle" time is ∼100 clock cycles for both power-gating transistors and body bias when the activity factor (α) is 0.05 (Figure 39(a)). For 400 consecutive active cycles, the maximum activity factor, beyond which no power reduction is achieved, is 0.1 to 1 (Figure 39(b)). With wider execution pipelines and instructions spending larger amounts of time waiting for memory access, activity factors of execution units in high-performance microprocessors are becoming smaller. Therefore, the minimum number of idle cycles or maximum activity factor limits for power-gating transistor and body bias techniques to be effective (Figure 40(a)) are within the range of values encountered in microprocessors. Measurements for a typical activity profile demonstrate 10% overall active power reduction by power-gating transistor and body bias (Figure 40(b)), in addition to that achieved by clock gating.

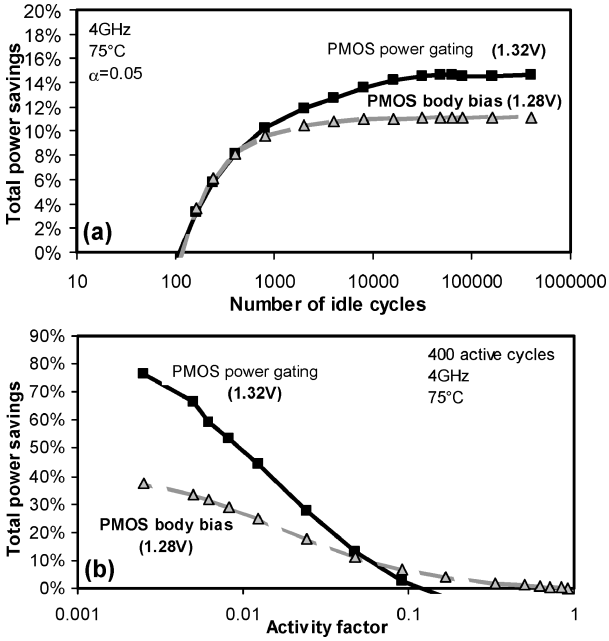


Fig. 39. (a) Total power savings for power gating (sleep) transistor and body bias as compared to clock gating only for a fixed activity factor as a function of the number of consecutive idle cycles. (b) Total power savings as a function of activity factor for 400 consecutive active cycles.

(a)	Break-even number of idle cycles with α=0.05	Break-even α with 400 active cycles
Adder with PMOS power gating transistor	115	0.13
Adder with PMOS body bias	121	> 0.5

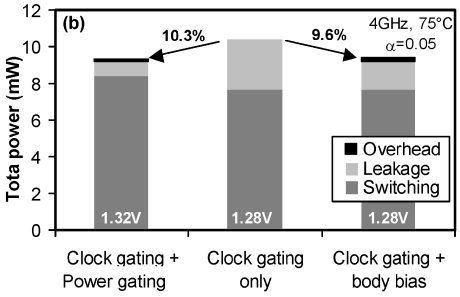


Fig. 40. (a) Break-even point for total power in terms of number of idle cycles and activity factor. (b) Components of total active power for power gating (sleep) transistor and body bias, compared to clock gating only with 400 consecutive active cycles.

5. SUMMARY OF CHALLENGES IN NANOSCALE CMOS

An integrated processing system offering over 200 Giga instructions per second, with 2 billion logic transistors and additionally an order of magnitude more memory transistors, using less than 20-nm physical gate length devices,

operating below 700 mV supply voltage by first half of the next decade—this is the expected roadmap should the scaling trends continue. Can we achieve this—maybe, maybe not! Nevertheless to attempt at implementing the vision of such a processing system, it is essential that its design comprehend variation and leakage power.

We reviewed two challenges, process variation and leakage power, and some techniques that can be used to reduce the impact of these on the behavior of nanoscale CMOS circuits. With increasing variation due to worse short channel effects, it will become inevitable to consider variations explicitly and rigorously in all areas of design. This will require the development of new circuit solutions that are tolerant to process variation and methodologies that combine computational efficiency of simple-minded worst-case methods, with the precision of statistical design methods. Other nanoscale CMOS design challenges includes continuing the scaling trend of on-die memory, dealing with increasing on-die interconnect parasitics, and dealing with increase in soft error rates due to reduction in stored charge.

Challenges at the platform level include the increasing difficulty in power delivery and heat removal. The variation in supply voltage is due to iR and L di/dt drops in the power grid with non-zero parasitic resistance (R) and non-zero loop inductance (L). Ideally, one would like to maintain the historical 10% variation is supply voltage. This is becoming harder due to increase in the current level and the rate of change of current due to faster switching as technology is scaled. In addition, the parasitic resistance and inductance have not been reducing at the same rate as the increase properties of current flow [Larrson 1994]. This problem is compounded by the fact that the supply voltage is expected to scale with technology. Traditionally, passive on-die and off-die decoupling capacitors were used to filter power supply noise. Delivering 100 W at 250-mV supply voltage is a very challenging problem due to high-power supply current and low-power supply voltage. Recently, researchers have shown use of active on-die voltage regulation to be more efficient in controlling supply voltage variation [Ang et al. 2000; Takashima et al. 1998]. To further improve the power distribution efficiency, an on-chip voltage down regulation scheme should be explored. In this scheme, the power distribution is done at higher voltages and converted locally to lower logic voltage level, thereby reducing not only the AC and DC current levels but also the percentage voltage drop in the package and global distribution grids. Such solutions benefit from integrated magnetic inductors [Gardner et al. 2001; Kursun et al. 2002, 2003, 2005].

The challenge of power dissipation or heat removal goes hand-in-hand with that of power delivery. Subsequent to the computation, power delivered to CMOS VLSI circuits gets dissipated as heat. Increase in power delivered with scaling results in increased power dissipation and higher-power density [De and Borkar 1994]. In order to maintain junction temperature constant with increased power dissipation, it maybe necessary to use more exotic cooling and enhanced heat-spreading solutions such as carbon nanotubes [Hone et al. 2000] and electrokinetic microchannel cooling [Goodson]. Since subthreshold leakage power will become more dominant with scaling, total power dissipated

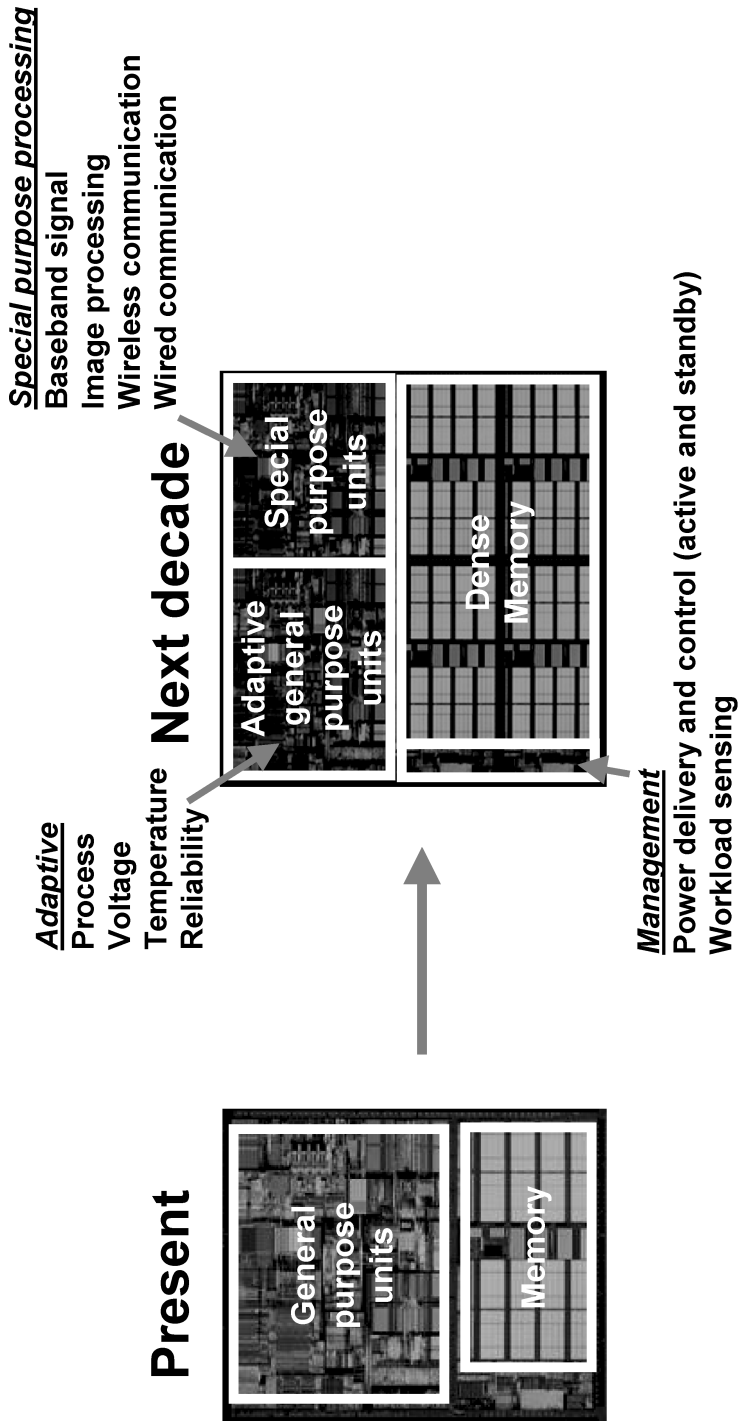


Fig. 41. Speculated evolution for future nanoscale silicon based CMOS systems.

will have strong junction temperature dependence. Therefore, instead of keeping junction temperature constant with scaling, it might be beneficial to decrease the temperature. This temperature scaling will not only reduce leakage power but also improve drive current, interconnect resistance and reliability [Aller et al. 2000; Ghoshal and Schmidt 2000]. Further optimization of device and circuits for low-temperature operation can provide additional scaling benefits [Huang et al. 1999; Aller et al. 2000; Jackson 2001; Banerjee et al. 2003].

As a final note—the expected evolution of nanoscale transistor technology is discussed in Grove [2002] and Moore [2003]. The expected evolution of present day CMOS VLSI computational units to nanoscale CMOS VLSI computational units is speculated in Figure 41. Essential features include adaptive techniques to reduce design margins, special purpose computation units to improve computational energy efficiency, dense memory choices that enable continued scaling of integrated random access memory, and effective power management schemes that while occupying silicon area enable integration of additional transistors for computation. All of this will be possible if and only if there is cohesive interaction between device, circuit, architecture, and platform designers!

ACKNOWLEDGMENTS

The author would like to thank Vivek De, Ali Keshavarzi, James Tschanz, Bradley Bloechel, Yibin Ye, Dinesh Somasekhar, Greg Dermer, Nitin Borkar, Matthew Haycock, and Shekhar Borkar from Intel Corporation; and James Kao (currently at Silicon Labs), Prof. Anantha Chandrakasan, and Prof. Dimitri Antoniadis from MIT for their support, technical discussions, and contributions.

REFERENCES

ALLER, I., BERNSTEIN, K., GHOSHAL, U., SCHETTLER, H., SCHUSTER, S., TAUR, Y., AND TERREITER, O. 2000. CMOS circuit technology for sub-ambient temperature operation. In *Proceedings of the International Solid-State Circuits Conference*. 214–215.

ANG, M., SALEM, R., AND TAYLOR, A. 2000. An on-chip voltage regulator using switched decoupling capacitors. In *Proceedings of the International Solid-State Circuits Conference*. 438–439.

ANTONIADIS, D. AND CHUNG, J. E. 1991. Physics and technology of ultra short channel MOSFET devices. In *Proceedings of the International Electron Devices Meeting.* 21–24.

ASENOV, A., SLAVCHEVA, G., BROWN, A. R., DAVIES, J. H., AND SAINI, S. 2001. Increase in the random dopant induced threshold fluctuations and lowering in sub-100 nm MOSFETs due to quantum effects: A 3-D density-gradient simulation study. *IEEE Trans. Elect. Dev. 48*, 4 (Apr.), 722–729.

BANERJEE, K., LIN, S.-C., KESHAVARZI, A., NARENDRA, S., AND DE, V. 2003. A self-consistent junction temperature estimation methodology for nanometer scale ICs with implications for performance and thermal management. In *Proceedings of the International Electron Devices Meeting* (Dec.). 36.7.1–36.7.4.

BOWMAN, K., DUVALL, S., AND MEINDL, J. 2001. Impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution. In *Proceedings of the International Solid-State Circuits Conference*. 278–279.

CHANDRAKASAN, A., SHENG, S., AND BRODERSEN, R. W. 1992. Low-power CMOS digital design. *IEEE J. Solid-State Circ. 27* (Apr.), 473–484.

CHAU, R., BOYANOV, B., DOYLE, B., DOCZY, M., DATTA, S., HARELAND, S., JIN, B., KAVALIEROS, J., AND METZ, M. 2003. Silicon nano-transistors for logic applications. *Physica E. Low-Dimen. Syst. Nanostruct. 19*, 1–2 (July), 1–5.

CHAU, R., DATTA, S., DOCZY, M., DOYLE, B., KAVALIEROS, J., AND METZ, M. 2004. High-k/metal-gate stack and its MOSFET characteristics. *IEEE Electron Dev. Lett. 25* (June), 408–410.

CHEN, Z., SHOTT, J., BURR, J., AND PLUMMER, J. D. 1994. CMOS technology scaling for low voltage low power applications. In *Proceedings of the IEEE Symposium on Low Power Electrons.* IEEE Computer Society Press, Los Alamitos, Calif. 56–57.

CHEN, Z., WEI, L., AND ROY, K. 1998. Estimation of standby leakage power in CMOS circuits considering accurate modeling of transistor stacks. In *Proceedings of the International Symposium on Low Power Electronics and Design*. 239–244.

DE, V. 2000. Forward biased MOS circuits. United States Patent, Patent number: 6,166,584. Filed: June 1997, Issued: Dec. 2000.

DE, V. AND BORKAR, S. 1999. Technology and design challenges for low power & high performance. In *Proceedings of the International Symposium on Low Power Electronics and Design*. (Aug.). 163–168.

GARDNER, D., CRAWFORD, A. M., AND WANG, S. 2001. High frequency (GHz) and low resistance integrated inductors using magnetic materials. In *Proceedings of the International Technology Conference*. 101–103.

GHOSHAL, U. AND SCHMIDT, R. 2000. Refrigeration technologies for sub-ambient temperature operation of computing systems. In *Proceedings of the International Solid-State Circuits Conference*. 216–217.

GOODSON, K. Research available at www.stanford.edu/group/microheat/hex.html.

GROVE, A. 2002. *IEDM 2002 Keynote Luncheon Speech*. http://www.intel.com/pressroom/archive/speeches/grove_20021210.pdf.

HOENEISEN, B. AND MEAD, C. A. 1972. Fundamental limitations in microelectronics I: MOS technology. *Solid-State Electron. 15* (July), 819–829.

HONE, J., BATLOGG, B., BENES, Z., JOHNSON, A. T., AND FISCHER, J. E. 2000. Quantized phonon spectrum of single-wall carbon nanotubes. *Science 289* (Sept.), 1730–1733.

HORIGUCHI, M., SAKATA, T., AND ITOH, K. 1993. Switched-source-impedance CMOS circuit for low standby sub-threshold current giga-scale LSI's. *IEEE J. Solid-State Circ. 28* (Nov.), 1131–1135.

HUANG, X., LEE, W.-C., KUO, C., HISAMOTO, D., CHANG, L., KEDZIERSKI, J., ANDERSON, E., TAKEUCHI, H., CHOI, Y.-K., ASANO, K., SUBRAMANIAN, V., KING, T.-J., BOKOR, J., AND HU, C. 1999. Sub 50-nm FinFET: PMOS. *IEDM Tech. Dig.* 67–70.

INTEL. http://www.intel.com/research/silicon/mooreslaw.htm.

JACKSON, K. 2001. *Optimal MOSFET Design for Low Temperature Operation*. MIT EECS Doctoral Thesis.

KAO, J. AND CHANDRAKASAN, A. 2000. Dual-threshold voltage techniques for low power digital circuits. *IEEE J. Solid-State Circ. 35* (July), 1009–1018.

KAO, J., NARENDRA, S., AND CHANDRAKASAN, A. 1998. MTCMOS hierarchical sizing based on mutual exclusive discharge patterns. In *Proceedings of the 35th Design Automation Conference* (June). 495–500.

KAWAHARA, T., HORIGUCHI, M., KAWAJIRI, Y., KITSUKAWA, G., AND KURE, T. 1993. Sub-threshold current reduction for decoded-driver by self-reverse biasing. *IEEE J. Solid-State Circ. 28* (Nov.), 1136–1144.

KEMPF, K. G. 1998. Improving throughput across the factory life-cycle. *Intel Tech. J.* Q4.

KESHAVARZI, A., MA, S., NARENDRA, S., BLOECHEL, B., MISTRY, K., GHANI, T., BORKAR, S., AND DE, V. 2001. Effectiveness of reverse body bias for leakage control, in scaled dual Vt CMOS ICs. In *Proceedings of the International Symposium Low Power Electronics and Design*. (Aug.), 207–212.

KOHNO, I., SANO, T., KATOH, N., AND YANO, K. 2000. Threshold canceling logic (TCL): A post-CMOS logic family scalable down to 0.02 mm. In *Proceedings of the International Solid-State Circuits Conference*. 218–219.

KURD, N. A., BARKARULLAH, J. S., DIZON, R. O., FLETCHER, T. D., AND MADLAND, P. D. 2001. A multi-gigahertz clocking scheme for Intel® Pentium® 4 microprocessor. *IEEE J. Solid-State Circ. 36* (Nov.), 1647–1653.

KURODA, T., FUJITA, T., MITA, S., NAGAMATSU, T., YOSHIOKA, S., SUZUKI, K., SANO, F., NORISHIMA, M., MUROTA, M., KAKO, M., KINUGAWA, M., KAKUMU, M., AND SAKURAI, T. 1996. A 0.9-V, 150-MHz, 10-mW, 4-mm^2, 2-D discrete cosine transform core processor with variable threshold-voltage (VT) scheme. *IEEE J. Solid-State Circ. 31* (Nov.), 1770–1779.

KURSUN, V., NARENDRA, S. G., DE, V. K., AND FRIEDMAN, E. G. 2002. Efficiency analysis of a high frequency buck converter for on-chip integration with a dual-VDD microprocessor. In *Proceedings of the European Solid-State Circuits Conference* (Sept.), 743–746.

KURSUN, V., NARENDRA, S. G., DE, V. K., AND FRIEDMAN, E. G. 2003. Monolithic DC-DC converter analysis and MOSFET gate voltage optimization. In *Proceedings of the IEEE/ACM International Symposium on Quality Electronic Design* (March), 279–284.

KURSUN, V., SCHROM, G., DE, V. K., FRIEDMAN, E. G., AND NARENDRA S. G. 2005. Cascode buffer for monolithic voltage conversion operating at high input supply voltages. In *Proceedings of the IEEE International Symposium on Circuits and Systems* (May).

LARRSON, P. 1999. Power supply noise in future ICs: A crystal ball reading. In *Proceedings of the Custom Integrated Circuits Conference*. 467–474.

LEE, C. H., LEE, S. J., JEON, T. S., BAI, W. P., SENSAKI, Y., ROBERTS, D., AND KWONG, D. L. 2000. Ultra thin ZrO(2) and Zr(27)Si(10)O(63) gate dielectrics directly prepared on si-substrate by rapid thermal processing. *SRC Techcon* (Sept.). 46.

LEE, J., TARACHI, G., WEI, A., LANGDO, T. A., FITZGERALD, E. A., AND ANTONIADIS, D. 1999. Super self-aligned double-gate (SSDG) MOSFETs utilizing oxidation rate difference and selective epitaxy. In *Proceedings of the International Electron Devices Meeting*. 71–74.

LEE, W., LANDMAN, P. E., BARTON, B., ABIKO, S., TAKAHASHI, H., MIZUNO, H., MURAMATSU, S., TASHIRO, K., FUSUMADA, M., PHAM, L., BOUTAUD, F., EGO, E., GALLO, G., TRAN, H., LEMONDS, C., SHIH, A., NANDAKUMAR, M., EKLUND, R. H., AND CHEN I.-C. 1997. A 1V DSP for wireless communications. In *Proceedings of the IEEE International Solid-State Circuits Conference* (Feb.). IEEE Computer Society Press, Los Alamitos, Calif. 92–93.

MIYAZAKI, M., ONO, G., HATTORI, T., SHIOZAWA, K., UCHIYAMA, K., AND ISHIBASHI, K. 2000. A 1000-MIPS/W microprocessor using speed adaptive threshold-voltage CMOS with forward bias. In *Proceedings of the International Solid-State Circuits Conference*. 420–421.

MIYAZAKI, M., MIZUNO, H., AND ISHIBASHI, K. 1998. A delay distribution squeezing scheme with speed-adaptive threshold-voltage CMOS (SA-Vt CMOS) for low voltage LSIs. In *Proceedings of the International Symposium Low Power Electronics and Design*. (Aug.). 48–53.

MOHAPATRA, N. R., DESAI, M. P., NARENDRA, S., AND RAO, V. R. 2002. The effect of high-K gate dielectrics on deep submicrometer CMOS device and circuit performance. *IEEE Trans. Elect. Dev. 49* (May), 826–831.

MOORE, G. 2003. No exponential is forever: but "forever" can be delayed. In *Proceedings of the International Solid-State Circuits Conference 1* (Feb.), 20–23.

MOORE, G. E. 1965. Cramming more components onto integrated circuits. *Electronics 38*, 8, April 19.

MULLER, D. A., SORSCH, T., MOCCIO, S., BAUMANN, F. H., EVANS-LUTTERODT, K., AND TIMP, G. 1999. The electronic structure at the atomic scale of ultrathin gate oxides. *Nature 399* (June), 758–761.

MUTOH, S., DOUSEKI, T., MATSUYA, Y., AOKI, T., SHIGEMATSU, S., AND YAMADA, J. 1995. 1-V power supply high-speed digital circuit technology with multithreshold-voltage CMOS. *IEEE J. Solid-State Circ. 30* (Aug.), 847–854.

NARENDRA, S., ANTONIADIS, D., AND DE, V. 1999. Impact of using adaptive body bias to compensate die-to-die Vt variation on within-die Vt variation. In *Proceedings of the International Symposium Low Power Electronics and Design* (Aug.). 229–232.

NARENDRA, S., BORKAR, S., DE, V., ANTONIADIS, D., AND CHANDRAKASAN, A. 2001. Scaling of stack effect and its application for leakage reduction. In *Proceedings of the International Symposium Low Power Electronics and Design* (Aug.). 195–200.

NARENDRA, S., KESHAVARZI, A., BLOECHEL, B., BORKAR, S., AND DE, V. 2003. Forward body bias for microprocessors in 130nm technology generation and beyond. *IEEE J. Solid-State Circ. 38* (May), 696–601.

NARENDRA, S., DE, V., BORKAR, S., ANTONIADIS, D., AND CHANDRAKASAN, A. 2004. Full-chip subthreshold leakage power prediction and reduction techniques for sub-0.18-um CMOS. *IEEE J. Solid-State Circ. 39* (Sept.), 501–510.

POON, H. C., YAU, L. D., JOHNSTON, R. L., AND BEECHAM, D. 1973. DC model for short-channel IGFET's. In *Proceedings of the International Electron Devices Meeting*. (Dec.). 156–159.

SAKATA, T., ITOH, K., HORIGUCHI, H., AND AOKI, M. 1994. Sub-threshold-current reduction circuits for multi-gigabit DRAM's. *IEEE J. Solid-State Circ. 29* (July), 761–769.

SCHULZ, M. 1999. The end of the road for silicon. *Nature 399* (June), 729–730.

SMOLAN, R. AND ERWITT, J. 1998. *One Digital Day—How the Microchip is Changing Our World*. Random House, New York.

SUN, S. W. AND TSUI, P. G. Y. 1994. Limitation of supply voltage scaling by MOSFET threshold-voltage variation. In *Proceedings of the Custom Integrated Circuits Conference*. 267–270.

TAKASHIMA, D., OOWAKI, Y., WATANABE, S., AND OHUCHI, K. 1998. Noise suppression scheme for gigabit-scale and gigabyte/s data-rate LSI's. *IEEE J. Solid-State Circ. 33* (Feb.), 260–267.

TAUR, Y. AND NING, T. H. 1998. *Fundamentals of Modern VLSI Devices*. Cambridge University Press, Cambridge, Mass.

THOMPSON, S., PACKAN, P., AND BOHR, M. 1998. MOS scaling: Transistor challenges for the 21st century. *Intel Tech. J.* Q3.

TSCHANZ, J., KAO, J., NARENDRA, S., NAIR, R., ANTONIADIS, D., CHANDRAKASAN, A., AND DE, V. 2002. Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage. *IEEE J. Solid-State Circ. 37* (Nov.), 1396–1402.

TSCHANZ, J., NARENDRA, S., NAIR, R., AND DE, V. 2003. Effectiveness of adaptive supply voltage and body bias for reducing impact of parameter variations in low power and high performance microprocessors. *IEEE J. Solid-State Circ. 38* (May), 826–829.

TSCHANZ, J., NARENDRA, S., YE, Y., BLOECHEL, B., BORKAR, S., AND DE, V. 2003. Dynamic sleep transistor and body bias for active leakage power control of microprocessors. *IEEE J. Solid-State Circ. 38* (Nov.), 1838–1845.

TSIVIDIS, Y. P. 1987. *Operation and Modeling of the MOS Transistor*. McGraw-Hill, New York.

VANGAL, S., ANDERS, M. A., BORKAR, N., SELIGMAN, E., GOVINDARAJULU, V., ERRAGUNTLA, V., WILSON, H., PANGAL, A., VEERAMACHANENI, V., TSCHANZ, J. W., YE, Y., SOMASEKHAR, D., BLOECHEL, B. A., DERMER, G. E., KRISHNAMURTHY, R. K., SOUMYANATH, K., MATHEW, S., NARENDRA, S. G., STAN, M. R. THOMPSON, S., DE, V., AND BORKAR, S. 2002. 5GHz 32b integer-execution core in 130nm dual-VT CMOS. *IEEE J. Solid-State Circ. 37* (Nov.), 1421–1432.

WANN, C., HARRINGTON, J., MIH, R., BIESEMANS, S., HAN, K., DENNARD, R., PRIGGE, O., LIN, C., MAHNKOPF, R., AND CHEN, B. 2000. CMOS with active well bias for low-power and RF/analog applications. In *Proceedings of the Symposium on VLSI Technology*. 158–159.

WEI, L., CHEN, Z., ROY, K., JOHNSON, M., YE, Y., AND DE, V. 1999. Design and optimization of dual-threshold circuits for low-voltage low-power applications. *IEEE Trans. VLSI Syst. 7* (Mar.), 16–24.

YAMASHITA, T., YOSHIDA, N., SAKAMOTO, M., MATSUMOTO, T., KUSUNOKI, M., TAKAHASHI, H., WAKAHARA, A., ITO, T., SHIMIZU, T., KURITA, K., HIGETA, K., MORI, K., TAMBA, N., KATO, N., MIYAMOTO, K., YAMAGATA, R., TANAKA, H., AND HIYAMA, T. 2000. A 450 Mhz 64b RISC processor using multiple threshold voltage CMOS. In *Proceedings of the IEEE International Solid-State Circuits Conference* (Feb.). IEEE Computer Society Press, Los Alamitos, Calif. 414–415.

YE, Y., BORKAR, S., AND DE, V. 1998. A new technique for standby leakage reduction in high-performance circuits. In *Proceedings of the 1998 Symposium on VLSI Circuits* (June). 40–41.

Received February 2005; accepted February 2005