Full-Chip Subthreshold Leakage Power Prediction and Reduction Techniques for Sub-0.18-μm CMOS

Siva Narendra, *Member, IEEE*, Vivek De, *Member, IEEE*, Shekhar Borkar, *Member, IEEE*, Dimitri A. Antoniadis, *Fellow, IEEE*, and Anantha P. Chandrakasan, *Fellow, IEEE*

***Abstract*—The driving force for the semiconductor industry growth has been the elegant scaling nature of CMOS technology. In future CMOS technology generations, supply and threshold voltages will have to continually scale to sustain performance increase, control switching power dissipation, and maintain reliability. These continual scaling requirements on supply and threshold voltages pose several technology and circuit design challenges. With threshold voltage scaling, subthreshold leakage power is expected to become a significant portion of the total power in future CMOS systems. Therefore, it becomes crucial to predict and reduce subthreshold leakage power of such systems. In the first part of this paper, we present a subthreshold leakage power prediction model that takes into account within-die threshold voltage variation. Statistical measurements of 32-bit microprocessors in 0.18-μm CMOS confirm that the mean error of the model is 4%. In the second part of this paper, we present the use of stacked devices to reduce system subthreshold leakage power without reducing system performance. A model to predict the scaling nature of this stack effect and verification of the model through statistical device measurements in 0.18-μm and 0.13-μm are presented. Measurements also demonstrate reduction in threshold voltage variation for stacked devices compared to nonstack devices. Comparison of the stack effect to the use of high threshold voltage or longer channel length devices for subthreshold leakage reduction is also discussed.**

***Index Terms*—CMOS, leakage estimation, leakage reduction, subthreshold leakage, within-die variation.**

I. Introduction

CONVENTIONALLY, CMOS technology has been scaled 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 an additional at least 60% frequency increase with augmented die area, architectural enhancements, and novel circuit techniques. This complexity increase results in higher energy consumption, peak power dissipation, and power delivery requirements [1].

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 [2]. Along with supply voltage scaling, the MOSFET device threshold voltage will have to scale to sustain the traditional 30% gate delay reduction. These supply and threshold voltage scaling requirements pose several technology and circuit design challenges [1], [3], [4]. One such challenge is the expected increase in threshold voltage variation due to worsening short channel effects. With technology scaling, the MOSFETs 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 MOSFET gate-body voltage [5], rendering the gate and body terminals less effective. As the band diagram illustrates in Fig. 1, 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. However, as the channel length becomes shorter, both channel length and drain voltage reduce this barrier height. This two-dimensional short channel effect causes the barrier height to be modulated by channel length variation, resulting in threshold voltage variation. 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. Fig. 2 shows measurements of 3σ threshold voltage variations for three device lengths in a 0.18-μm generation, confirming this behavior.

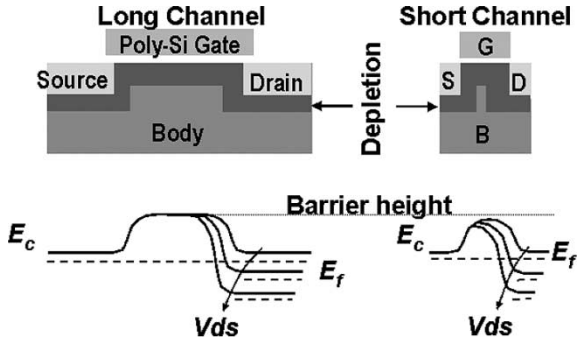


Fig. 1. Barrier height lowering due to channel length reduction and drain voltage increase.

With supply and threshold voltage scaling, control of threshold voltage variation becomes essential for achieving high yields and limiting worst case subthreshold leakage [6]. Maintaining good device aspect ratio by scaling gate-oxide thickness is important for controlling threshold voltage tolerances [7]. With the silicon dioxide gate dielectric thickness approaching scaling limits [8], [9], researchers have been exploring several alternatives, including the use of high

Manuscript received September 27, 2002; revised October 8, 2003.

S. Narendra, V. De, and S. Borkar are with the Microprocessor Research Laboratories, Intel Corporation, Hillsboro, OR 97124 USA (e-mail: siva.g.narendra@intel.com).

D. A. Antoniadis and A. P. Chandrakasan are with the Microsystems Technology Laboratories, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: anantha@mtl.mit.edu).

Digital Object Identifier 10.1109/JSSC.2003.821776

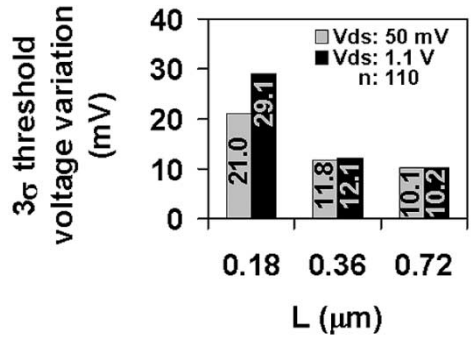


Fig. 2. Dependence of threshold voltage variation on channel length and drain voltage.

permittivity gate dielectric, metal gate, novel device structures, and circuit-based techniques [10]–[12]. Meanwhile, it is important to note that threshold voltage variation not only affects supply voltage scaling but also the accuracy of subthreshold leakage power prediction. Accurate subthreshold leakage power prediction is critical for future CMOS systems since the subthreshold leakage power is expected to be a significant portion of the total power due to threshold voltage scaling [1]. In this paper, a subthreshold leakage power prediction model that takes into account within-die threshold voltage variation due to short channel effect will be presented. We will also demonstrate through statistical measurements of 32-bit microprocessors in 0.18 μm CMOS the accuracy of the new subthreshold leakage power prediction model compared to other existing models. It should be pointed out that threshold voltage variation due to random doping fluctuation is not taken into account, which is expected to become significant in sub-100-nm technologies. In a subthreshold leakage dominant CMOS system, the need to identify techniques to reduce this variation and subthreshold leakage power becomes inevitable. Use of stacked devices to reduce system subthreshold leakage power without reducing system performance will be discussed in the paper. Analytical model to predict the scaling nature of this stack effect and verification of the model through statistical device measurements in 0.18-μm CMOS will be presented. Measurements also show reduction in threshold voltage variation for stacked devices compared to nonstack devices. Comparison of the stack effect to the use of high threshold voltage or longer channel length devices for subthreshold leakage reduction will also be discussed [13].

II. Prediction of Full-Chip Subthreshold Leakage

It has been established that to limit the energy and power increase in future CMOS technology generations, the supply voltage $V_{\rm dd}$ will have to continually scale [1]. The amount of energy reduction depends on the magnitude of $V_{\rm dd}$ scaling. Along with $V_{\rm dd}$ scaling, the threshold voltage (V_t) of MOS devices will have to scale to sustain the traditional 30% gate delay reduction. These $V_{\rm dd}$ 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. Should the present scaling trend continue, it is expected that the subthreshold leakage power will become as much as 50% of the total power by the 90-nm generation [1]. Under this scenario, it is important to be able to predict subthreshold leakage power more accurately. Most subthreshold leakage current prediction techniques do not take into account the variation in within-die threshold voltage. It will be shown that this assumption leads to significant inaccuracies. There are unpublished models that take within-die threshold voltage variation into account empirically. In this paper, a rigorous mathematical model for full-chip subthreshold leakage current that considers within-die threshold voltage variation will be derived. Bulk substrate microprocessor measurements that verify the improvement in subthreshold leakage current prediction with the new model are also presented. Calculation of subthreshold leakage power is straightforward once the subthreshold leakage current is known for a given $V_{\rm dd}$.

A. Present Subthreshold Leakage Current Prediction Techniques

Due to the wide variation in expected threshold voltage of MOS devices from die to die and within die during the lifetime of a process, present subthreshold leakage current prediction techniques provide lower and upper bounds on the subthreshold leakage current. The subthreshold leakage power of most chips lies between the two bounds, as shown in [14]. In older technology generations, basing system design on the two subthreshold leakage current bounds was acceptable since subthreshold leakage power was a negligible component of the total power. In most systems, the worst case bound is assumed for the design. In future technology generations where as much as half of the system power during active mode can be due to subthreshold leakage, depending the worse case bound 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 predict subthreshold leakage current as accurately as possible. The upper and lower bound prediction equations and measurements are provided in the next part of this section. The lower bound subthreshold leakage current $(I_{{\rm leak}-l})$ prediction of a chip is given as follows:

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

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 the fraction of PMOS and NMOS device widths that are in the *off* state, and I_p^o and I_n^o are the nominally expected subthreshold leakage currents per unit width of PMOS and NMOS devices in a particular chip. The nominal subthreshold leakage current is obtained for devices with mean threshold voltage or channel length. The upper bound subthreshold leakage current $I_{{\rm leak}-u}$ prediction of a chip is related to the device subthreshold leakage as follows:

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

where $I_{{\rm off}-p}^{3\sigma}$ and $I_{{\rm off}-n}^{3\sigma}$ are the worst case subthreshold leakage current per unit width of PMOS and NMOS devices. The worst case subthreshold leakage current is obtained for devices with

threshold voltage or channel length 3σ lower than the mean subthreshold leakage currents per unit width of PMOS and NMOS devices in a particular chip.

B. Subthreshold Leakage Current Prediction Including Within-Die Variation

To include the impact of within-die threshold voltage or channel length variation, it is necessary to consider the entire range of subthreshold leakage currents, not just the mean subthreshold leakage or the worst case subthreshold 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 subthreshold leakage of the device with the mean threshold voltage or channel length. Then, by performing the weighted sum of devices of different subthreshold leakage, we can predict the total subthreshold leakage of the chip. This is achieved by integrating the threshold voltage or channel length distribution multiplied by the subthreshold leakage, as follows:

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

In the above equation, the first exponent predicts the fraction of the total width for the device subthreshold leakage predicted 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 $n\phi_t$. ϕ_t is the thermal voltage and n is $1 + (\mathrm{C_d}/\mathrm{C_{ox}})$ [7]. If the distribution considered is channel length, then x in the above equation will represent channel length l, and a will be equal to λ, which can be predicted for a technology by measuring the relationship between channel length and device subthreshold leakage. Note that device parameter variation due to random doping fluctuation is not considered. 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 subthreshold leakage is then given as follows:

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

Let

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

$$\therefore I_{\text{leak}} = \frac{I^o w}{k}\frac{1}{\sqrt{\pi}} e^{\frac{\sigma^2}{2\lambda^2}}\int_{\frac{l\,\min-\mu}{\sqrt{2}\sigma}+\frac{\sigma}{\sqrt{2}\lambda}}^{\frac{l\,\max-\mu}{\sqrt{2}\sigma}+\frac{\sigma}{\sqrt{2}\lambda}} e^{-[t]^2}\,dt.$$

The integral can be rewritten as

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

Since

$$\begin{aligned} &\mathrm{erf}(z) \to 1 \quad \text{if} \quad z > 1 \quad \text{and} \\ &\frac{l\,\max-\mu}{\sqrt{2}\sigma}+\frac{\sigma}{\sqrt{2}\lambda},\quad \frac{\mu-l\,\min}{\sqrt{2}\sigma}-\frac{\sigma}{\sqrt{2}\lambda} \gg 1 \\ &\Rightarrow I_{\text{leak}} = \frac{I^o w}{k} e^{\frac{\sigma^2}{2\lambda^2}}. \end{aligned}$$

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

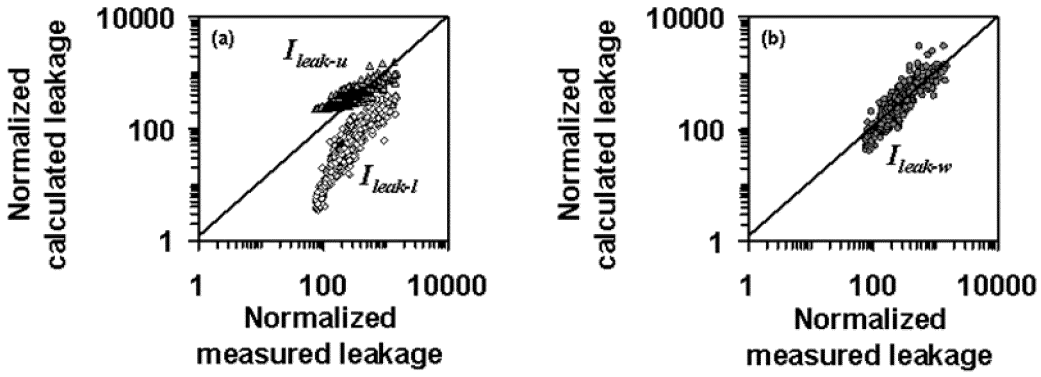
$$I_{\text{leak}-w} = \frac{I_p^o w_p}{k_p} e^{\frac{\sigma_p^2}{2\lambda_p^2}} + \frac{I_n^o w_n}{k_n} e^{\frac{\sigma_n^2}{2\lambda_n^2}}$$

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 the *off* state, I_p^o and I_n^o are the expected mean subthreshold 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, and 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}, a macroscopic standard deviation (σ) representing parameter variation in a chip can be determined if its I_{leak} is known:

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

C. Standby Subthreshold Leakage Measurement Results

Standby subthreshold leakage power measurements on 960 samples of a 0.18-μm 32-bit bulk substrate microprocessor were carried out. The subthreshold leakage current (with $V_{\text{gs}} = 0$ V and $V_{\text{ds}} = V_{\text{dd}}$) and effective channel length



(a) (b)

Fig. 3. Comparison of calculated subthreshold leakage current versus measured subthreshold leakage current for (a) existing subthreshold leakage current estimation techniques and (b) subthreshold leakage current estimation technique introduced in this paper.

measurements of 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 subthreshold leakage power was calculated using the $I_{\text{leak}-l}$, $I_{\text{leak}-u}$, and $I_{\text{leak}-w}$ formulas. In addition, we assumed that on an average half of the devices will be in the *off* state, that is, $k_p = k_n = 2$. This assumption is based on the fact that in high-performance designs, the majority of the device widths are in nonstacked buffers and domino logic. For designs that have the majority of the device width in complex gates, a better prediction of k_p and k_n based on stack-effect-related leakage reduction will be necessary. We will describe a stack-effect leakage reduction model based on device fundamentals in Section III.

The calculated subthreshold leakage currents based on the three methods with the $k_p = k_n = 2$ assumption are compared to the measured subthreshold leakage current. Fig. 3(a) clearly illustrates that the upper bound technique overpredicts the subthreshold leakage current of the chips, while the lower bound techniques underpredicts the subthreshold leakage current. However, the prediction technique introduced in this paper that includes within-die variation matches the measurement better, as illustrated in Fig. 3(b). The data shown in Fig. 3 is summarized in Fig. 4. As the figure indicates, the subthreshold leakage power for most of the samples are underpredicted by 6.5× if the lower bound technique is used and over predicted by 1.5× if the upper bound technique is used. The measured-to-calculated subthreshold leakage ratio for the majority of the device samples is 1.04 for the new technique described in this paper. The calculated subthreshold leakage is within ±20% of the measured subthreshold leakage for more than 50% of the samples, if the new $I_{\text{leak}-w}$ technique is used. Only 11% and 0.2% of the samples fall into this range for the $I_{\text{leak}-u}$ and $I_{\text{leak}-l}$ techniques, respectively. $I_{\text{leak}-w}$ technique can be used to predict full-chip standby subthreshold leakage with better accuracy once device level subthreshold leakage, parameter variation, and total transistor widths are known. This technique can also be used to estimate bulk substrate full-chip active leakage power by dividing the entire chip into multiple iso-temperature regions and using the $I_{\text{leak}-w}$ leakage estimation formula separately for each region. $I_p^o, I_n^o, w_p, w_n\lambda_p$, and λ_n will have to be determined for each iso-temperature region. For partially depleted SOI substrate designs, active leakage estimation would require proper modeling of floating body and the resulting fluctuations in the body voltage.

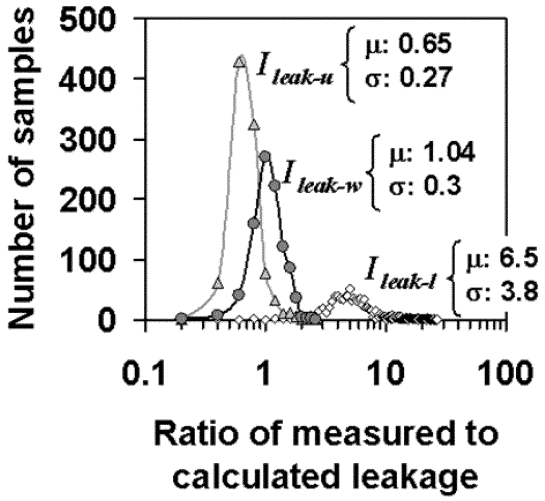


Fig. 4. Ratio of measured to calculated subthreshold leakage current ratio distribution for $I_{\text{leak}-u}$, $I_{\text{leak}-l}$, and $I_{\text{leak}-w}$ techniques. (Sample size: 960.)

III. Subthreshold Leakage Reduction

To reiterate, should the present scaling trend continue, it is expected that the subthreshold leakage power will become as much as 50% of the total power in the 90-nm generation [1]. Under this scenario, it is not only important to be able to predict subthreshold leakage power more accurately, as discussed in the previous section, but it becomes crucial to identify techniques to reduce this subthreshold leakage power component. It has been shown previously that the stacking of two *off* devices has significantly reduced subthreshold leakage compared to a single *off* device [15]–[17]. This concept of stack effect is illustrated in Fig. 5.

In this section, a model is derived that predicts the stack-effect factor, which is defined as the ratio of the subthreshold leakage current of one *off* device to the subthreshold leakage current of two *off* devices in series. Model derivation based on

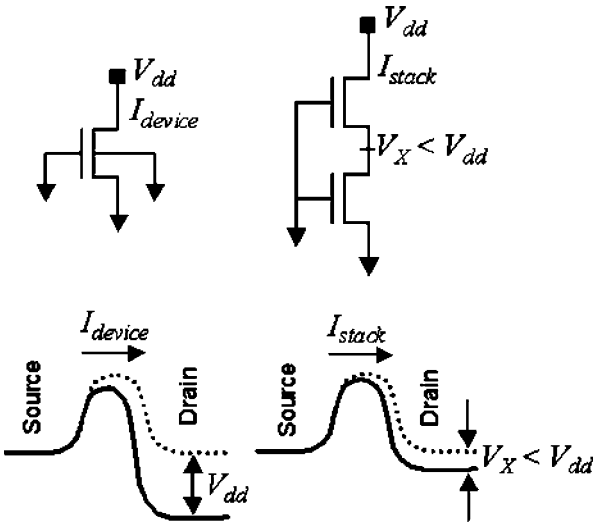


Fig. 5. Subthreshold leakage current difference between a single *off* device and a stack of two *off* devices. As illustrated by the energy band diagram, the barrier height is modulated to be higher for the two-stack due to smaller drain-to-source voltage resulting in reduced subthreshold leakage.

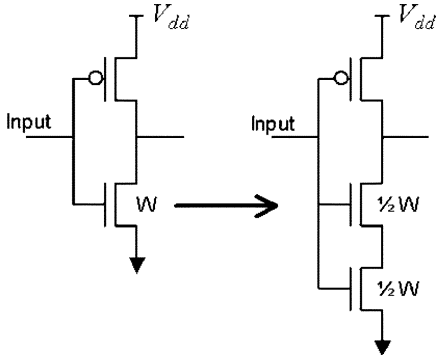


Fig. 6. Tradeoff between standby subthreshold leakage and performance by forcing a two-stack under iso-input load. An NMOS two-stack will reduce subthreshold leakage when input stays at logic 0.

device fundamentals and verification of the model through statistical device measurements from 0.18-μm and 0.13-μm technology generations and the scaling nature of the stack-effect subthreshold leakage reduction factor will also be discussed.

One solution to the problem of ever-increasing subthreshold leakage is to force a nonstack device to a stack of two devices without affecting the input load, as shown in Fig. 6. By ensuring iso-input load, the previous gate's delay and the switching power will remain unchanged. Logic gates after stack forcing will reduce subthreshold leakage power, but incur a delay penalty, similar to replacing a low-V_t device with a high-V_t device in a dual-V_t design [18]. In a dual-V_t design, the low-V_t devices are used in performance critical paths and the high-V_t devices in the rest [19]. Usually a significant fraction of the devices can be high-V_t or forced-stack since a large number of the paths are noncritical. This will reduce the overall subthreshold leakage power of the chip without impacting operating clock frequency. A stack-forcing method to reduce subthreshold leakage in paths that are not performance critical will be discussed. This stack-forcing technique either can be

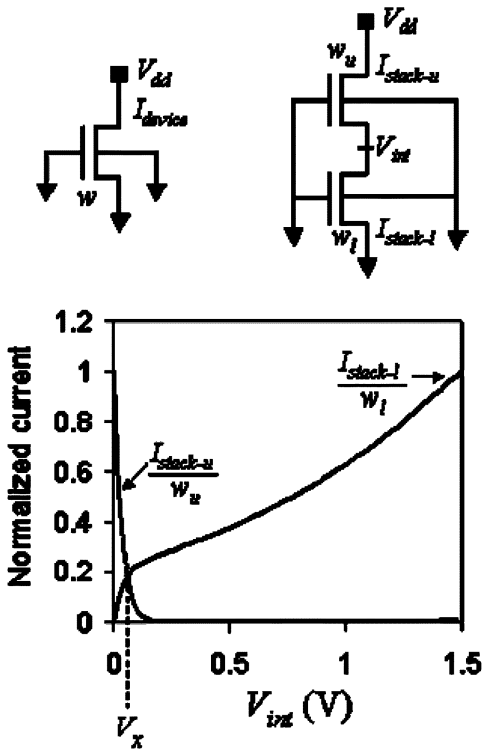


Fig. 7. Load line analysis showing the subthreshold leakage reduction in a two-stack.

used in conjunction with dual-V_t or can be used to reduce the subthreshold leakage in a single-V_t design. Differences between achieving subthreshold leakage reduction through forced stacks and channel length increase are discussed. It is necessary to note that the reduced device width for forced stacking might result in 1) higher V_t than the nominal devices due to narrow width effects or 2) underflow of device width below the minimum allowed by the process technology. In the first case, the leakage reduction will be higher than just forced stack with a corresponding penalty in delay. In the second case, underflow avoidance will either mean forced stacking cannot be implemented for such gates or minimum device width will have to be used for the stacked devices if the corresponding leakage reduction is more than the switching power increase.

A. Model for Stack Effect Factor

Let I_1 be the subthreshold leakage of a single device of unit width in the *off* state with its $V_{\rm gs} = V_{\rm bs} = 0$ V and $V_{\rm ds} = V_{\rm dd}$. If the gate-drive, body bias, and drain-to-source voltages reduce by $\Delta V_g, \Delta V_b$, and ΔV_d, respectively, from the above-mentioned conditions, the subthreshold leakage will reduce to

$$I_1' = I_1 \quad 10^{-\frac{1}{S}[\Delta V_g + \lambda_d \Delta V_d + k_\gamma \Delta V_b]}$$

where S is the subthreshold swing, λ_d is the drain-induced barrier lowering (DIBL) factor, and k_γ is the body effect coefficient. The above equation assumes that the resulting $V_{\rm ds} > 3kT/q$ [20]. For the two-device stack shown in Fig. 7, a steady-state condition will be reached when the intermediate node voltage $V_{\rm int}$ approaches V_x such that the subthreshold leakage currents in the upper and lower devices are equal.

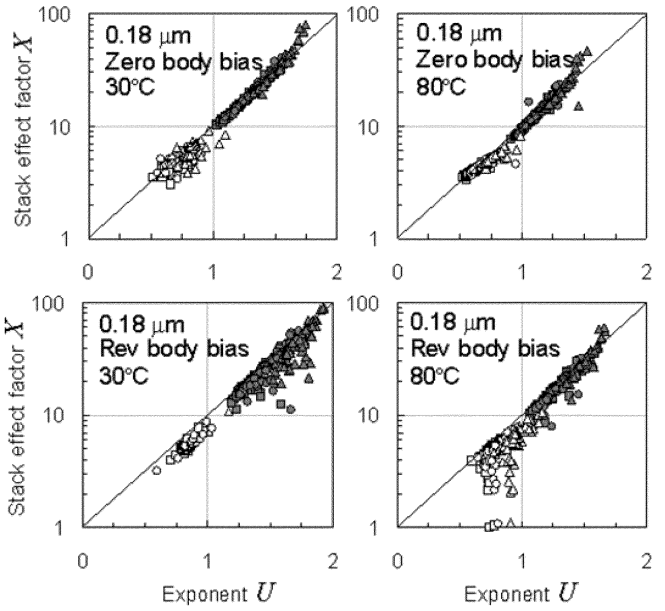


Fig. 8. Measurement results showing the relationship between stack-effect factor X for a two-stack to the exponent U. Lines indicate the relationship as per the analytical model and symbols are from measurement results. White symbols are for nominal channel devices and gray symbols are for devices smaller than the nominal channel length. Triangle, circle, and square symbols are for $V_{\rm dd}$ of 1.5, 1.2, and 1.1 V, respectively. Zero body bias is when the body-to-source diode of the device closet to the power supply is zero biased and reverse body bias is when the diode is reverse biased by 0.5 V.

Under this condition, the subthreshold leakage currents in the upper and lower devices can be expressed as

$$I_{\rm stack-u} = w_u I_1 \quad 10^{\frac{-(1+\lambda_d+k_\gamma)V_x}{S}}$$
$$I_{\rm stack-l} = w_l I_1 \quad 10^{\frac{-\lambda_d(V_{\rm dd}-V_x)}{S}}$$

and the intermediate node voltage will be

$$V_x = \frac{\lambda_d V_{\rm dd} + S\log\frac{w_u}{w_l}}{1+k_\gamma+2\lambda_d.}$$

For short channel devices, the body terminal's control on the channel is negligible compared to gate and drain terminals, implying $k_\gamma \ll 1+2\lambda_d$. Hence, the steady-state value V_x of the intermediate node voltage can be approximated as

$$V_x \approx \frac{\lambda_d V_{\rm dd} + S\log\frac{w_u}{w_l}}{1+2\lambda_d}.$$

Substituting V_x in either $I_{\rm stack-u}$ or $I_{\rm stack-l}$ will yield the subthreshold leakage current in a two-stack given by

$$I_{\rm stack} = w_u^\alpha w_l^{1-\alpha} I_1 \quad 10^{\frac{-\lambda_d V_{\rm dd}}{S}(1-\alpha)}$$

where

$$\alpha \approx \frac{\lambda_d}{1+2\lambda_d}.$$

The subthreshold leakage reduction achievable in a two-stack comprising devices with widths w_u and w_l compared to a single device of width w is given by

$$X = \frac{I_{\rm device}}{I_{\rm stack}} = \frac{w}{w_u^\alpha w_l^{1-\alpha}} 10^{\frac{\lambda_d V_{\rm dd}}{S}(1-\alpha)}$$
$$= 10^{\frac{\lambda_d V_{\rm dd}}{S}(1-\alpha)} \quad \text{when} \quad w_u = w_l = w$$

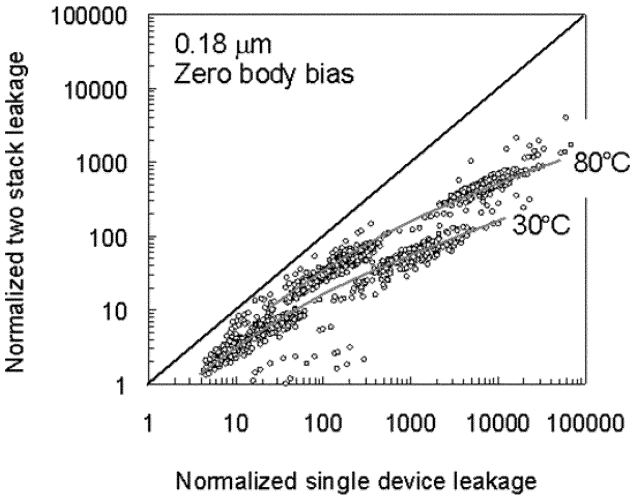


Fig. 9. Measurement results indicate a slower rate of increase in subthreshold leakage of two-stack compared to that of a single device.

The stack-effect factor, when $w_u = w_l = w$, can be rewritten as

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

where U is the two-stack exponent which depends only on the process parameters λ_d and S and the design parameter $V_{\rm dd}$. Once these parameters are known, the reduction in subthreshold 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 $3\,kT/q$.

To confirm the model's accuracy, we performed device measurements on test structures fabricated in 0.18-μm and 0.13-μm process technologies. Results discussed in the rest of the section are from NMOS device measurements, but similar results hold true for PMOS devices as well.

Fig. 8 shows NMOS device measurements under different temperature, $V_{\rm dd}$, body bias, and channel length conditions for 0.18-μm technology generations, which prove the accuracy of the theoretical model. It is important to note that the model discussed above does not include the impact of diode junction subthreshold leakages that originate at the intermediate stack node. In Fig. 8, the model's accuracy deviates the most under reverse body bias for nominal channel length devices, where the ratio of diode junction subthreshold leakage to subthreshold leakage current increases.

It is known that the stack-effect factor strongly depends on λ_d as suggested by the model. In addition, a decrease in the channel length L will increase λ_d in a given technology [41]. So, any increase in the subthreshold leakage of a single device due to decrease in L will not increase subthreshold leakage of a two-stack at the same rate. This is illustrated in Fig. 9, where the increase in two-stack subthreshold leakage is at a slower rate than that of a single device. Therefore, variation in L will result in smaller effective threshold voltage variation for a two-stack compared to a single device.

Fig. 10 illustrates the average stack-effect factor for the nominal channel devices in both 0.18-μm and 0.13-μm technology generations obtained from both the measurements and the model. The increase in stack-effect factor at a given $V_{\rm dd}$ with technology scaling is attributed to increase in λ_d, which

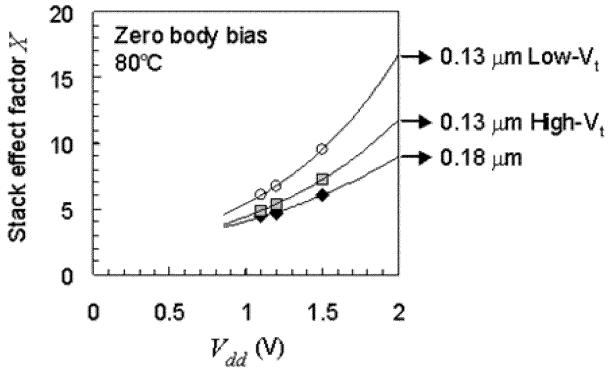


Fig. 10. Nominal channel length device measurement results showing stack-effect factor across two technology generations. The increase in stack effect factor is attributed to worsening of short channel effect λ_d, which is predicted by the analytical model. The higher stack effect factor for the low-V_t device in 0.13-μm technology generation is attributed to the same reason. Lines are from analytical model and symbols are from measurement.

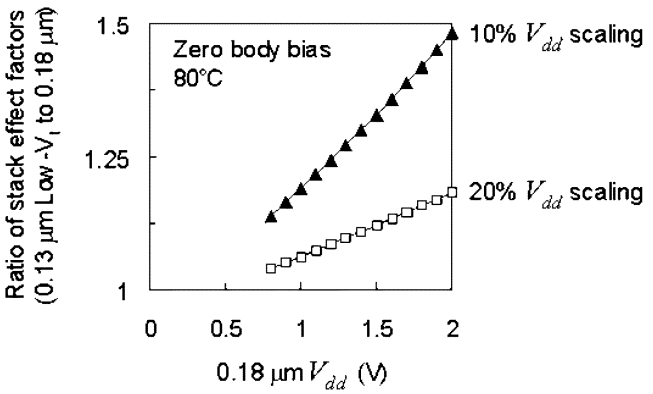


Fig. 11. Nominal channel length device measurement results indicating the scaling of stack effect factor from 0.18-μm to 0.13-μm low-V_t under different $V_{\rm dd}$ scaling conditions. The low-V_t device typically contributes majority (over 70%) of subthreshold leakage in high-performance 0.13-μm designs, so the comparison is made with the low-V_t device.

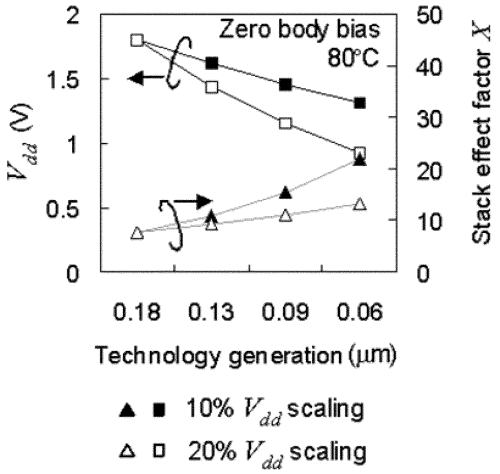


Fig. 12. Prediction in the scaling of stack effect factor for two $V_{\rm dd}$ scaling scenarios in nominal channel length devices. $V_{\rm dd}$ for 0.18-μm is assumed to be 1.8 V.

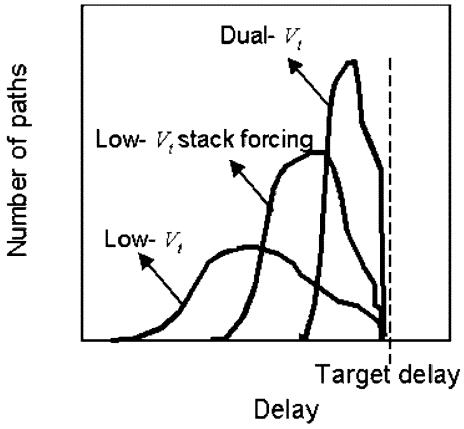


Fig. 13. Stack forcing and dual-V_t can reduce subthreshold leakage of gates in paths that are faster than required.

is predicted by the analytical model. The higher stack-effect factor for the low-V_t device in 0.13-μm technology generation is attributed to the same reason. In the 0.13-μm generation, the low-V_t device will dominate chip subthreshold leakage. Fig. 11 shows the scaling of stack effect from a 0.18-μm device to a 0.13-μm low-V_t device based on device measurements under different $V_{\rm dd}$ scaling scenarios. Since λ_d is expected to increase due to worsening device aspect ratio and since $V_{\rm dd}$ scaling will slow down due to related challenges [21], the stack-effect subthreshold leakage reduction factor is expected to increase with technology scaling. The predicted scaling of stack-effect factor from 0.18-μm to 0.06-μm is depicted in Fig. 12. This scaling nature of the stack-effect factor makes it a powerful technique for subthreshold leakage reduction in future technologies. In the next section, we describe a circuit technique for taking advantage of the stack effect to reduce subthreshold leakage at a functional block level.

B. Subthreshold Leakage Reduction Using Forced Stacks for Logic Gates

As shown earlier, stacking of two devices that are *off* has significantly reduced subthreshold leakage compared to a single *off* device. However, due to the iso-input load requirement and due to stacking of devices, the drive current of a forced-stack gate will be lower, resulting in increased delay. So stack forcing can be used only for paths that are noncritical, just like using high-V_t devices in a dual-V_t design. Forced-stack gates will have slower output edge rate similar to gates with high-V_t devices. Fig. 13 illustrates the use of techniques that provide delay–subthreshold-leakage tradeoff. As demonstrated in the figure, paths that are faster than required can be slowed down, which will result in subthreshold leakage savings. Such tradeoffs are valid only if the resulting path still meets the target delay. Simulation results show that the delay increases by $\sim$100% for about 10$\times$ reduction in leakage by stack forcing in the 0.13-μm technology generation. By properly employing forced stack, one can reduce standby and active subthreshold leakage of noncritical paths even if a dual-V_t process is not available. This method can also be used in conjunction with dual-V_t. Stack forcing provides wider coverage in the delay–subthreshold-leakage tradeoff space. Functional blocks have naturally stacked gates such as NAND, NOR, or other complex gates. By maximizing the number of natural stacks in

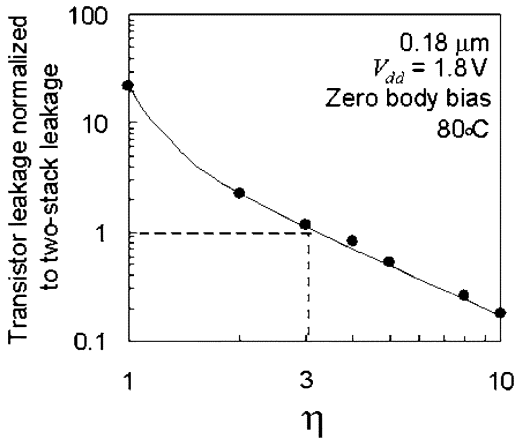


Fig. 14. Comparing device subthreshold leakage reduction due to channel length increase with two-stack subthreshold leakage. The channel length is given by $\eta \times 0.18\ \mu$m. Stack subthreshold leakage is a two-stack of devices with $\eta = 1$ and $wu = wl = 1/2\,w$. Subthreshold leakage numbers are obtained from simulation under iso-input load.

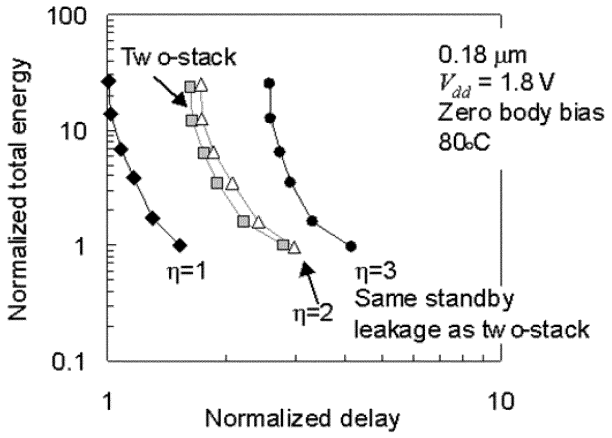


Fig. 15. Energy–delay tradeoff of inverter under different configurations with fanout of 1 and iso-input load. The simulation-based comparison clearly shows that the two-stack configuration's delay is less than increasing channel length, especially when compared to iso-standby subthreshold leakage ($\eta = 3$) configuration.

the *off* state during standby by setting proper input vectors, the standby subthreshold leakage of the functional block can be reduced. Since it is not possible to force all natural stacks in the functional block to be in the *off* state, the overall subthreshold leakage reduction at the block level will be far less than the stack-effect subthreshold leakage reduction possible at a single logic gate level [16]. With stack forcing, the potential for subthreshold leakage reduction will be higher.

C. Stack Effect Versus Channel Length Increase

It is possible to facilitate delay–subthreshold-leakage tradeoff by increasing the channel length of devices [22] that are in noncritical paths. To maintain iso-input load, the channel width will have to be reduced along with the increase in the channel length. Fig. 14 shows the mean subthreshold leakage reduction achievable by increasing the channel length. In Fig. 14, the channel length of interest is given by $\eta \times 0.18\ \mu$m and stack subthreshold leakage is for a stack of two devices with η of 1 and $w_u = w_l = 1/2\,w$. Subthreshold leakage calculation includes within-die channel length variation as described in Section II. As is clear from Fig. 14, the channel length has to be increased to 3 × that of the nominal channel length to match the mean subthreshold leakage of a two-stack of 0.18-μm devices. The main reason for such a large increase is attributed to the reverse short channel effect that is present due to halo doping [21], where V_t reduces with increase in channel length.

Fig. 15 shows the energy–delay tradeoff of an inverter under different configurations with fanout of 1 and iso-input load. The simulation-based comparison clearly shows that the two-stack configuration's delay is less than delay due to increasing channel length, especially when compared to iso-standby subthreshold leakage ($\eta \approx 3$) configuration. As summarized in Fig. 16, η of 2 has about the same delay as that of the two-stack with η of 1 but with a 2.3× higher mean subthreshold leakage. On the other hand, η of 3 provides about the same mean subthreshold leakage as the two-stack but with 60% higher delay.

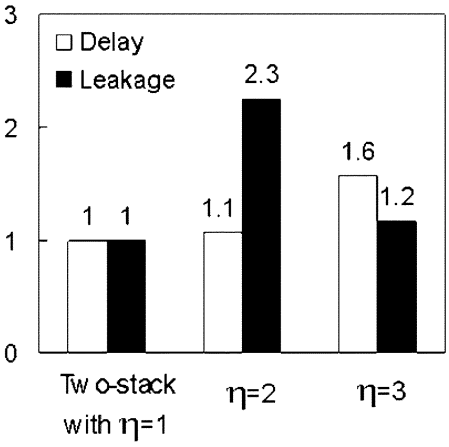


Fig. 16. Summary of delay–subthreshold-leakage tradeoff comparison between two-stack and channel length.

D. Case Study

Two-stack assignment of low-V_t transistors was applied to a 32-bit microprocessor's instruction decode block in 0.13-μm technology. Stack assignment was done so that the all-low-V_t maximum frequency of 1 GHz is preserved at 1.4 V. Switching power of 45.9 mW at 1.4 V was also preserved since iso-input load was maintained during stack assignment. All low-V_t subthreshold leakage power was 39.1 mW. Iso-frequency stack assignment allowed conversion about 70% of transistor width to two-stack, resulting in subthreshold leakage power reduction of 3×. If high-V_t assignment instead of forced stack is used, then about 95% of the transistor width becomes high-V_t, resulting in 4.3× subthreshold leakage reduction. Although using high-V_t is more effective, the forced stack method does not need a dual V_t process. Additionally, if it is available it can used in conjunction with forced stacks.

IV. Conclusion

We showed that threshold voltage variation not only affects supply voltage scaling but also the accuracy of subthreshold

leakage power prediction. Accurate subthreshold leakage current prediction is very critical for future CMOS systems since the subthreshold leakage power is expected to be a significant portion of the total power due to threshold voltage scaling. A subthreshold leakage current prediction technique that takes into account within-die threshold voltage variation was presented. Standby leakage measurement results from 960 samples of a 0.18-μm 32-bit microprocessor verified the model's accuracy. A step to extend this technique to estimate active leakage current was described. In a subthreshold leakage dominant CMOS system, the need to identify techniques to reduce this variation and subthreshold leakage power also becomes inevitable. A model based on device fundamentals that predicted the scaling nature of stack-effect-based subthreshold leakage reduction was presented. Device measurements verified the model's accuracy across different temperatures, channel lengths, body bias values, supply voltages, and process technologies. Measurements also demonstrate reduction in threshold voltage variation for stacked devices compared to nonstack devices. Using stack forcing to reduce standby and active subthreshold leakage components was discussed and the advantage of stack forcing over channel length increase for delay–subthreshold-leakage tradeoff was demonstrated. A case study for stack forcing showed 3$\times$ subthreshold leakage reduction at the block level without reduction of the maximum frequency of operation.

REFERENCES

[1] V. De and S. Borkar, "Technology and design challenges for low power and high performance," in *Int. Symp. Low Power Electronics and Design*, Aug. 1999, pp. 163–168.

[2] A. Chandrakasan, S. Sheng, and R. W. Brodersen, "Low-power CMOS digital design," *IEEE J. Solid-State Circuits*, vol. 27, pp. 473–484, Apr. 1992.

[3] D. Antoniadis and J. E. Chung, "Physics and technology of ultra short channel MOSFET devices," in *Int. Electron Devices Meeting Tech. Dig.*, 1991, pp. 21–24.

[4] Z. Chen, J. Shott, J. Burr, and J. D. Plummer, "CMOS technology scaling for low voltage low power applications," in *Proc. IEEE Symp. Low Power Electronics*, 1994, pp. 56–57.

[5] H. C. Poon, L. D. Yau, R. L. Johnston, and D. Beecham, "DC model for short-channel IGFETs," in *Int. Electron Devices Meeting Tech. Dig.*, Dec. 1973, pp. 156–159.

[6] S. W. Sun and P. G. Y. Tsui, "Limitation of supply voltage scaling by MOSFET threshold-voltage variation," in *Proc. IEEE Custom Integrated Circuits Conf.*, 1994, pp. 267–270.

[7] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[8] D. A. Muller, T. Sorsch, S. Moccio, F. H. Baumann, K. Evans-Lutterodt, and G. Timp, "The electronic structure at the atomic scale of ultrathin gate oxides," *Nature*, vol. 399, pp. 758–761, June 1999.

[9] M. Schulz, "The end of the road for silicon," *Nature*, vol. 399, pp. 729–730, June 1999.

[10] K. Reid, B. Taylor, L. Dip, L. Hebert, R. Garcia, R. Hegde, J. Grant, D. Gilmer, A. Franke, V. Dhandapani, M. Azrak, L. Prabhu, R. Rai, S. Bagchi, J. Conner, S. Backer, F. Dumbuya, B. Nguyen, and P. Tobin, "80 nm poly-Si gate CMOS with HfO2 gate dielectric," in *Int. Electron Devices Meeting Tech. Dig.*, Dec. 2001, pp. 30.1.1–30.1.4.

[11] J. Lee, G. Tarachi, A. Wei, T. A. Langdo, E. A. Fitzgerald, and D. Antoniadis, "Super self-aligned double-gate (SSDG) MOSFET's utilizing oxidation rate difference and selective epitaxy," in *Int. Electron Devices Meeting Tech. Dig.*, 1999, pp. 71–74.

[12] I. Kohno, T. Sano, N. Katoh, and K. Yano, "Threshold canceling logic (TCL): A post-CMOS logic family scalable down to 0.02 μm," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, 2000, pp. 218–219.

[13] S. Narendra, S. Borkar, V. De, D. Antoniadis, and A. Chandrakasan, "Scaling of stack effect and its application for leakage reduction," in *Proc. Int. Symp. Low Power Electronics and Design*, Aug. 2001, pp. 195–200.

[14] A. Keshavarzi, S. Ma, S. Narendra, B. Bloechel, K. Mistry, T. Ghani, S. Borkar, and V. De, "Effectiveness of reverse body bias for leakage control, in scaled dual Vt CMOS ICs," in *Proc. Int. Symp. Low Power Electronics and Design*, Aug. 2001, pp. 207–212.

[15] J. P. Halter and F. Najm, "A gate-level leakage power reduction method for ultra-low-power CMOS circuits," in *Proc. IEEE Custom Integrated Circuits Conf.*, 1997, pp. 475–478.

[16] Y. Ye, S. Borkar, and V. De, "A new technique for standby leakage reduction in high-performance circuits," in *Symp. VLSI Circuits Dig. Tech. Papers*, 1998, pp. 40–41.

[17] Z. Chen, M. Johnson, L. Wei, and K. Roy, "Estimation of standby leakage power in CMOS circuits considering accurate modeling of transistor stacks," in *Proc. Int. Symp. Low Power Electronics and Design*, 1998, pp. 239–244.

[18] L. Su *et al.*, "A high-performance sub-0.25 μm CMOS technology with multiple thresholds and copper interconnects," in *Symp. VLSI Technology Dig. Tech. Papers*, 1998, pp. 18–19.

[19] D. T. Blaauw, A. Dharchoudhury, R. Panda, S. Sirichotiyakul, C. Oh, and T. Edwards, "Emerging power management tools for processor design," in *Proc. Int. Symp. Low Power Electronics and Design*, 1998, pp. 143–148.

[20] A. Chandrakasan, W. J. Bowhill, and F. Fox, *Design of High Performance Microprocessor Circuits*. New York, NY: IEEE Press, 2000, pp. 46–47.

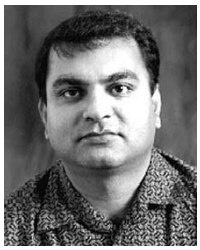
[21] Y. Taur, "CMOS scaling beyond 0.1 μm: How far can it go?," in *Int. Symp. VLSI Technology, Systems, and Applications Dig. Tech. Papers*, 1999, pp. 6–9.

[22] D. Dobberpuhl, "The design of a high performance low power microprocessor," in *Proc. Int. Symp. Low Power Electronics and Design*, 1996, pp. 11–16.

Siva Narendra (M'99) received the B.E. degree from the Government College of Technology, Coimbatore, India, in 1992, the M.S. degree from Syracuse University, Syracuse, NY, in 1994, and the Ph.D. degree from Massachusetts Institute of Technology, Cambridge, in 2002.

He has been with Intel Laboratories, Hillsboro, OR, since 1997, where his research areas include low voltage MOS analog and digital circuits and impact of MOS parameter variation on circuit design. He is also an Adjunct Faculty with the Department of Electrical and Computer Engineering, Oregon State University, Corvallis. He has authored 40 papers and has 40 issued patents in these areas.

Dr. Narendra is an Associate Editor for the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION SYSTEMS and a member of the ISLPED, ISQED, and DAC/ISSCC Student Design Contest program committees.



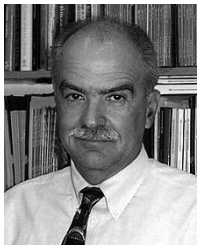
Vivek De (M'89) received the Ph.D. in electrical engineering from Rensselaer Polytechnic Institute, Troy, NY, in 1992.

He is currently a Senior Principal Engineer and Manager of Low Power Circuit Technology at the Microprocessor Research Laboratories, Intel Corporation, Hillsboro, OR. He is also an Adjunct Faculty Member with the Department of Electrical and Computer Engineering, Oregon State University, Corvallis. He has authored 82 technical papers in refereed international conferences and journals, and two book chapters on low power design. He has 23 issued patents and 45 patents pending.

Dr. De served as Technical Program Chair of 2001 ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED'01), General Chair of ISLPED'02, and Technical Program Chair of 2002 ACM Great Lakes Symposium on VLSI. He served on technical program committees of ARVLSI and ISQED conferences. He was the guest editor of a special issue on low power electronics for IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION SYSTEMS. He was the recipient of a Best Paper Award at the 1996 IEEE International ASIC Conference, Portland, OR.

Shekhar Borkar (M'97) received the B.S. and M.S. degrees in physics from the University of Bombay, Bombay, India, in 1979 and the M.S. degree in electrical engineering from the University of Notre Dame, Notre Dame, IN, in 1981.

He joined Intel Corporation, Hillsboro, OR, in 1981, where he worked on the design of the 8051 family of microcontrollers, high-speed communication links for the iWarp multicomputer, and Intel Supercomputers. He is an Intel Fellow and Director of Circuit Research in Intel Laboratories, researching low-power high-performance circuits and high-speed signaling. He is also an Adjunct Faculty Member of the Oregon Graduate Institute, Beaverton, and teaches digital CMOS VLSI.



Dimitri A. Antoniadis (M'79–SM'83–F'90) received the B.S. degree in physics from the National University of Athens, Athens, Greece, in 1970, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, in 1976. His initial research activities were in the area of measurement and modeling of the earth's ionosphere and thermosphere ranging from instrument design to computer simulation.

In 1978, Dr. Antoniadis joined the faculty at Massachusetts Institute of Technology (MIT), Cambridge, where he currently holds the Ray and Maria Stata Chair in Electrical Engineering. He was co-founder and first Director of the MIT Microsystems Technology Laboratories, and from 1993 to 2000, he was Director of the SRC MIT Center of Excellence for Microsystems Technology. Currently, he is Director of the National, Multi-University Focus Research Center for Materials, Structures and Devices, centered at MIT. His initial research activities covered the area of measurement and modeling of the earth's ionosphere and thermosphere ranging from instrument design to computer simulation. He led the development of the first two generations of the SUPREM process simulator and since then, his technical activity has been in the area of semiconductor devices and integrated circuit technology. He has worked on the physics of diffusion in silicon, thin-film technology and devices and quantum-effect semiconductor devices. His current research focuses on the physics and technology of extreme-submicron Si, SOI and Si/SiGe MOSFETs.

Dr. Antoniadis is the recipient of the Solid State Science and Technology Young Author Award of the Electrochemical Society in 1979, the Paul Rappaport Award of the IEEE in 1998, and the 2002 Andrew Grove Award of the IEEE.

Anantha P. Chandrakasan (S'87–M'95–SM'01–F'04) received the B.S, M.S., and Ph.D. degrees in electrical engineering and computer sciences from the University of California, Berkeley, in 1989, 1990, and 1994, respectively.

Since September 1994, he has been with the Massachusetts Institute of Technology, Cambridge, where he is currently a Professor of electrical engineering and computer science. His research interests include low-power digital integrated circuit design, distributed wireless microsensors, ultra wideband radios, and emerging technologies. He is a coauthor of *Low Power Digital CMOS Design* (Norwell, MA: Kluwer, 1995) and *Digital Integrated Circuits* (Upper Saddle River, NJ: Pearson Prentice Hall, 2002, 2nd ed.). He is also a coeditor of *Low Power CMOS Design* (Piscataway, NJ: IEEE Press, 1997) and *Design of High-Performance Microprocessor Circuits* (Piscataway, NJ: IEEE Press, 2000).

Dr. Chandrakasan has received several Best Paper Awards, including the 1993 IEEE Communications Society's Best Tutorial Paper Award, the IEEE Electron Devices Society's 1997 Paul Rappaport Award for the Best Paper in an EDS publication during 1997, and the 1999 Design Automation Conference Design Contest Award. He has served as a technical program co-chair for the 1997 International Symposium on Low-Power Electronics and Design (ISLPED), VLSI Design '98, and the 1998 IEEE Workshop on Signal Processing Systems. He was the Signal Processing Subcommittee Chair for ISSCC 1999–2001, the Program Vice-Chair for ISSCC 2002, and the Program Chair for ISSCC 2003. He was an Associate Editor for the IEEE JOURNAL OF SOLID-STATE CIRCUITS from 1998 to 2001. He serves on the SSCS AdCom. He is the Technology Directions Chair for ISSCC 2004.