Full-chip sub-threshold leakage power prediction model for sub-0.18 μm CMOS

Siva Narendra[¶§], Vivek De[§], Shekhar Borkar [§], Dimitri Antoniadis[¶], and Anantha Chandrakasan[¶]
[¶]Microsystems Technology Laboratories
Massachusetts Institute of Technology, Cambridge, MA, 02139
{naren, anantha, daa}@mtl.mit.edu
[§]Microprocessor Research
Intel Laboratories, Hillsboro, OR, 97124
{vivek.de, shekhar.y.borkar}@ intel.com

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 sub-threshold leakage power is expected to become a significant portion of the total power in future CMOS systems. Therefore, it becomes crucial to predict sub-threshold leakage power of such systems. In this paper, we present a sub-threshold leakage power prediction model that takes into account within-die threshold voltage variation. Statistical measurements of 32-bit microprocessors in 0.18 μm CMOS confirms that the mean error of the model to be 4%. Comparisons of this model to two other existing models that do not take within-die threshold voltage variation into account are also presented.

1. 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 of 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, MOSFET 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 [1, 3, 4].

One such challenge is the expected increase in threshold voltage variation due to worsening short channel effects. With technology scaling the MOSFET's 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 to be less effective. As the band diagram illustrates in Figure 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 makes 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. Figure 2 shows measurements of 3σ threshold voltage variations for three device lengths in a 0.18 μm generation confirming this behavior.

With supply and threshold voltage scaling, control of threshold voltage variation becomes essential for achieving high yields and limiting worst-case sub-threshold 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 permittivity gate dielectric, metal gate, novel device structures and circuit-based techniques [10, 11, 12]. In the meanwhile, it is important to note that threshold voltage variation not only affects supply voltage scaling but also the accuracy of sub-threshold leakage power prediction. Accurate sub-threshold leakage power prediction is very critical for future CMOS systems since the sub-threshold leakage power is expected to be a significant portion of the total power due to threshold voltage scaling [1]. In this paper, sub-threshold 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 sub-threshold leakage power prediction model compared to other existing models.

2. PREDICTION OF FULL-CHIP SUB-THRESHOLD LEAKAGE
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 sub-threshold leakage power due to V_t scaling. Should the present scaling trend continue it is

ISLPED '02, August 12-14, 2002, Monterey, California, USA.

expected that the sub-threshold 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 sub-threshold leakage power more accurately. Present sub-threshold 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. A mathematical model for full-chip sub-threshold leakage current that considers within-die threshold voltage variation will be derived. Microprocessor measurements that verify the improvement in sub-threshold leakage current prediction with the new model are also presented. Calculation of sub-threshold leakage power is straightforward once the sub-threshold leakage current is known for a given V_{dd}.

3. PRESENT SUB-THRESHOLD LEAKAGE CURRENT PREDICTION TECHNIQUES

Due to the wide variation expected threshold voltage of MOS devices from die-to-die and within-die during the life time of a process, present sub-threshold leakage current prediction techniques provide lower and upper bounds on the sub-threshold leakage current. The sub-threshold leakage powers of most chips lie between the two bounds as shown in [13]. In older technology generations, basing system design on the two sub-threshold leakage current bounds was acceptable since sub-threshold leakage power was a negligible component of the total power.

In most current 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 sub-threshold 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 sub-threshold 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 sub-threshold leakage current ($I_{leak\text{-}l}$) prediction of a chip is given as follows,

$$I_{leak\text{-}l} = \frac{w_p}{k_p} I^o_p + \frac{w_n}{k_n} I^o_n$$

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^o_p and I^o_n are the nominally expected sub-threshold leakage currents per unit width of PMOS and NMOS devices in a particular chip. The nominal sub-threshold leakage current is obtained for devices with mean threshold voltage or channel length. The upper bound sub-threshold leakage current ($I_{leak\text{-}u}$) prediction of a chip is related to the device sub-threshold leakage as follows,

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

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

4. SUB-THRESHOLD 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 sub-threshold leakage currents, not just the mean sub-threshold leakage or the worst-case sub-threshold 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 sub-threshold leakage of the device with the mean threshold voltage or channel length. Then by performing the weighted sum of devices of different sub-threshold leakage, we can predict the total sub-threshold leakage of the chip. This is achieved by integrating the threshold voltage or channel length distribution multiplied by the sub-threshold leakage, as shown below.

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

In the above equation, the first exponent predicts the fraction of the total width for the device sub-threshold 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 $1+(C_d/C_{ox})$ [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. λ can be predicted for a technology by measuring the relationship between channel length and device sub-threshold 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 sub-threshold leakage is then given as follows,

$$I_{leak} = \frac{I^o w}{k} \frac{1}{\sigma\sqrt{2\pi}} \int_{l\,min}^{l\,max} e^{\frac{-(l-\mu)^2}{2\sigma^2}} e^{\frac{(\mu-l)}{\lambda}} dl$$

$$= \frac{I^o w}{k} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{\sigma^2}{2\lambda^2}} \int_{l\,min}^{l\,max} e^{\frac{-(l-\mu)^2}{2\sigma^2}} e^{\frac{(\mu-l)}{\lambda}} e^{\frac{-\sigma^2}{2\lambda^2}} dl$$

$$= \frac{I^o w}{k} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{\sigma^2}{2\lambda^2}} \int_{l\,min}^{l\,max} e^{-\left[\frac{l-\mu}{\sqrt{2}\sigma}+\frac{\sigma}{\sqrt{2}\lambda}\right]^2} dl$$

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,

$$I_{\text{leak}} = \frac{I^o w}{2k}\ e^{\frac{\sigma^2}{2\lambda^2}} \left[\frac{2}{\sqrt{\pi}} \int_o^{\frac{l\,max-\mu}{\sqrt{2}\sigma}+\frac{\sigma}{\sqrt{2}\lambda}} e^{-t^2}\,dt \;-\; \frac{2}{\sqrt{\pi}} \int_o^{\frac{l\,min-\mu}{\sqrt{2}\sigma}+\frac{\sigma}{\sqrt{2}\lambda}} e^{-t^2}\,dt \right]$$

$$= \frac{I^o w}{2k}\ e^{\frac{\sigma^2}{2\lambda^2}} \left[erf\left(\frac{l\,max-\mu}{\sqrt{2}\sigma}+\frac{\sigma}{\sqrt{2}\lambda} \right) - erf\left(\frac{l\,min-\mu}{\sqrt{2}\sigma}+\frac{\sigma}{\sqrt{2}\lambda} \right) \right] \quad \because erf(z) = \frac{2}{\sqrt{\pi}} \int_o^z e^{-t^2} dt$$

$$= \frac{I^o w}{2k}\ e^{\frac{\sigma^2}{2\lambda^2}} \left[erf\left(\frac{l\,max-\mu}{\sqrt{2}\sigma}+\frac{\sigma}{\sqrt{2}\lambda} \right) + erf\left(\frac{\mu-l\,min}{\sqrt{2}\sigma}-\frac{\sigma}{\sqrt{2}\lambda} \right) \right] \quad \because erf(-z) = -erf(z)$$

where, *erf (z)* is the error function.

Since,

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

$$\Rightarrow I_{leak} = \frac{I^o w}{k}\ e^{\frac{\sigma^2}{2\lambda^2}}$$

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

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

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 off state; ${I^o}_p$ and ${I^o}_n$ are the expected mean sub-threshold 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 sub-threshold 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,

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

5. STANDBY SUB-THRESHOLD LEAKAGE MEASUREMENT RESULTS

Standby sub-threshold leakage power measurements on 960 samples of a 0.18 µm 32-bit microprocessor were carried out. The sub-threshold leakage current (with V_{gs} = 0 V and $V_{ds} = V_{dd}$) and effective channel length measurements of test devices that accompany each microprocessor were measured to determine ${I^o}_p$, ${I^o}_n$, 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 sub-threshold leakage power was calculated using the I_{leak-l}, I_{leak-u}, and I_{leak-w} formulae.

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 sub-threshold leakage currents are then compared with the measured sub-threshold leakage current.

Figure 3(a) clearly illustrates that the upper bound technique over predicts the sub-threshold leakage current of the chips while the lower bound techniques under predicts the sub-threshold leakage current. However, the prediction technique introduced in this paper that includes within-die variation matches the measurement better, as illustrated in Figure 3(b).

Data shown in Figure 3 is summarized in Figure 4. As the figure indicates the sub-threshold leakage power for most of the samples are under predicted by 6.5X if the lower bound technique is used and over predicted by 1.5X if the upper bound technique is used. The measured-to-calculated sub-threshold leakage ratio for majority of the device samples is 1.04 for the new technique described in this paper. The calculated sub-threshold leakage is within ±20% of the measured sub-threshold leakage for more than 50% of the samples, if the new I_{leak-w} technique is used. Only 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 full-chip standby sub-threshold leakage with better accuracy once device level sub-threshold leakage, parameter variation, and total transistor widths are known.

This technique can also be used to estimate full-chip active leakage power by dividing the entire chip into multiple iso-temperature regions and using the I_{leak-w} leakage estimation formula separately for each region. ${I^o}_p$, ${I^o}_n$, w_p, w_n, λ_p, and λ_n will have to be determined for each iso-temperature region.

6. CONCLUSIONS

We showed that threshold voltage variation not only affects supply voltage scaling but also the accuracy of sub-threshold leakage power prediction. Accurate sub-threshold leakage current prediction is very critical for future CMOS systems since the sub-threshold leakage power is expected to be a significant portion of the total power due to threshold voltage scaling. A sub-threshold 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. Steps to extend this technique to estimate active leakage power were described.

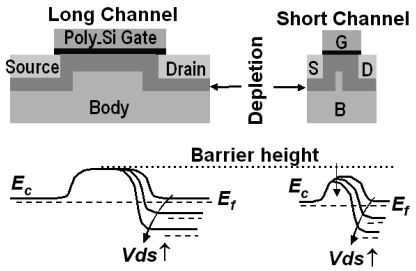

Figure 1: Barrier height lowering due to channel length reduction and drain voltage increase.

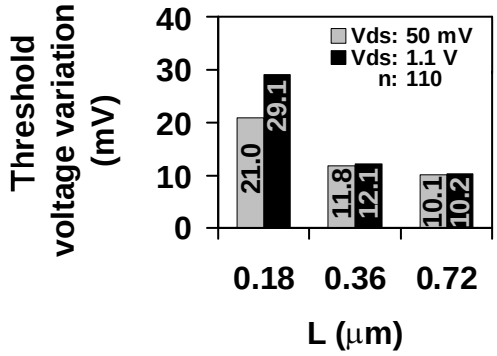

Figure 2: Dependence of threshold voltage variation (3σ) on channel length and drain voltage.

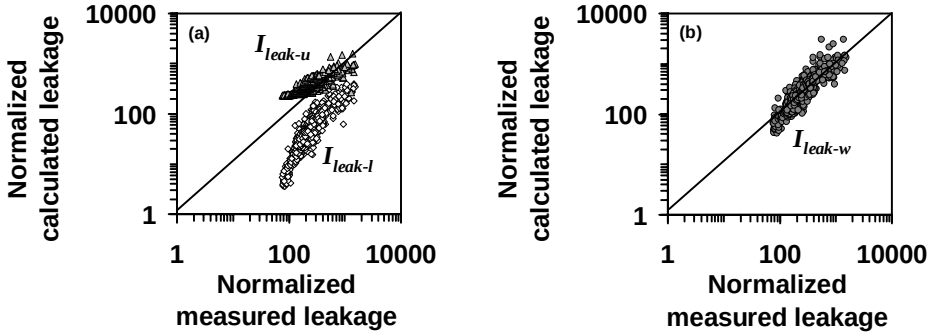

Figure 3: Comparison of calculated sub-threshold leakage current versus measured sub-threshold leakage current for (a) existing sub-threshold leakage current estimation techniques and (b) sub-threshold leakage current estimation technique introduced in this work.

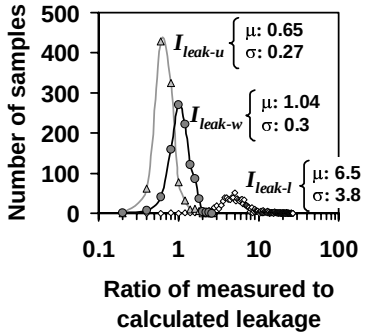

Figure 4: Ratio of measured to calculated sub-threshold leakage current distribution for I_{leak-u}, I_{leak-l}, and I_{leak-w} techniques (Sample size: 960).

7. ACKNOWLEDGEMENTS

The first author would like to acknowledge DARPA and Intel Labs for financial assistance, Dinesh Somasekhar and Yibin Ye of Intel Labs for technical discussions, and Adam Brand of Intel Corporation for the measurements.

8. REFERENCES

[1] V. De and S. Borkar, "Technology and Design Challenges for Low Power & High Performance," *Intl. Symp. Low Power Electronics and Design*, pp. 163-168, Aug. 1999.

[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," *Intl. Electron devices Meeting*, pp. 21-24, 1991.

[4] Z. Chen, J. Shott, J. Burr, and J. D. Plummer, "CMOS Technology Scaling for Low Voltage Low Power Applications," *IEEE Symp. Low Power Elec*., pp. 56-57, 1994.

[5] H.C. Poon, L.D. Yau, R.L. Johnston, D. Beecham, "DC Model for Short-Channel IGFET's," *IEEE Intl. Electron Devices Meeting*, pp. 156-159, Dec. 1973.

[6] S. W. Sun and P. G. Y. Tsui, "Limitation of Supply Voltage Scaling by MOSFET Threshold-Voltage variation," *IEEE Custom Integrated Circuits Conf*., pp. 267-270, 1994.

[7] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*, Cambridge University 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," *IEEE Intl. Electron Devices Meeting*, Paper 30.1, Dec. 2001.

[11] J. Lee, G. Tarachi, A. Wei, T. A. Langdo, E. A. Fitzgerald, D. Antoniadis, "Super self-aligned double-gate (SSDG) MOSFETs utilizing oxidation rate difference and selective epitaxy," *Intl. Electron Devices Meeting*, pp. 71-74, 1999.

[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," *Intl. Solid-State Circuits Conf*., pp. 218-219, 2000.

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