# Modeling and Analysis of Parametric Yield under Power and Performance Constraints

**Rajeev R. Rao, David Blaauw, and Dennis Sylvester**
University of Michigan, Ann Arbor

**Anirudh Devgan**
Magma Design Automation

Leakage current is a stringent constraint in today's ASIC designs. Effective parametric yield prediction must consider leakage current's dependence on chip frequency. The authors propose an analytical expression that includes both subthreshold and gate leakage currents. This model underlies an integrated approach to accurately estimating yield loss for a design with both frequency and power limits.

**Continued scaling** of device dimensions, combined with shrinking threshold voltages, has resulted in an exponential rise in IC power dissipation. This increase is primarily due to leakage, which is emerging as a significant portion of total power consumption. Kao, Narendra, and Chandrakasan estimate that subthreshold leakage power will account for more than 50% of total power for portable applications developed for the 65-nm technology node.[1] In future technologies, aggressive scaling of oxide thickness will lead to significant gate oxide tunneling current, further aggravating the leakage problem. Across successive technology generations, subthreshold leakage increases by about 5 times,[2] and gate leakage can increase by as much as 30 times.

At the same time, the increased presence of parameter variability in modern designs has intensified the need for designers to consider the impact of statistical leakage current variations. For a 10% variation in a transistor's effective channel length, there can be as much as a threefold difference in the amount of subthreshold leakage current.[3] Gate leakage current exhibits an even greater sensitivity to process variations, showing a 15× difference in current for a 10% variation in oxide thickness in a 100-nm Berkeley Predictive Technology Model (BPTM) process technology.[4] Hence, considerable variability in chip-level leakage current can be expected, and researchers have reported measured variations as high as 20×.[5]

For current designs, chip manufacturers typically calculate a lot's yield by characterizing chips according to their operating frequency. Manufacturers reject the subset of dies that don't meet the required performance constraint, making this aspect of the design process very important from a commercial point of view. However, Borkar et al. have observed that among the "good" chips that meet the performance constraint, a substantial portion dissipate very large amounts of leakage power and thus are unsuitable for commercial use.[5] Circuit delay and leakage current are inversely correlated so that devices with channel lengths smaller than the nominal value have smaller delay but produce significant amounts of leakage current. As a result, chips with high operating frequencies dissipate vast amounts of leakage power.

Figure 1 illustrates this inverse correlation, showing the distribution of chip performance and leakage based on silicon measurements over many samples of a high-end processor design.[5] As the figure shows, both the mean and the variance of the leakage distribution increase significantly for chips with higher frequencies. This trend is particularly troubling because it substantially reduces the yield of designs that are both performance and leakage constrained. Hence, designers need accurate leakage-yield-prediction methods to model this dependency and accurately map the yield value in the process and performance domains. The "Related

0740-7475/05/$20.00 © 2005 IEEE Copublished by the IEEE CS and the IEEE CASS

## Related work

Researchers have proposed several statistical methods for estimating full-chip leakage current. Narendra et al. consider within-die threshold voltage variability to estimate full-chip subthreshold leakage current.[1] Mukhopadhyay, Raychowdhury, and Roy use a compact current model to estimate total leakage current.[2] Rao et al. present analytical equations to model subthreshold leakage as a function of the transistor's channel length.[3] Srivastava et al., as well as Mukhopadhyay and Roy, use a moment-based approximation approach to estimate leakage current's mean and variance.[4,5] However, none of these methods provide exact mathematical equations to express chip leakage, and, furthermore, they don't consider the dependence of leakage on frequency.

More recently, researchers have suggested mathematical models to determine a lot's parametric yield. Najm and Menezes use principal component analysis to estimate timing yield.[6] Choi, Paul, and Roy propose a circuit-resizing algorithm that ensures the circuit's delay optimality while achieving a desired yield number.[7] Zhang, Wason, and Banerjee present a probabilistic framework for estimating full-chip subthreshold leakage power distribution as well as leakage-constrained yield under the impact of process variations.[8] Tsai et al. survey the impact of technology scaling and process variation on the efficacy of various leakage reduction schemes.[9]

### References

1. S. Narendra et al., "Full-Chip Subthreshold Leakage Power Prediction Model for Sub-0.18μm CMOS," *Proc. Int'l Symp. Low Power Electronics and Design* (ISLPED 02), IEEE Press, 2002, pp. 19-23.
2. S. Mukhopadhyay, A. Raychowdhury, and K. Roy, "Accurate Estimate of Total Leakage Current in Scaled CMOS Circuits Based on Compact Current Modeling," *Proc. 40th Design Automation Conf.* (DAC 03), ACM Press, 2003, pp. 169-174.
3. R. Rao et al., "Statistical Analysis of Subthreshold Leakage Current for CMOS Circuits," *IEEE Trans. Very Large Scale Integrated Systems*, vol. 12, no. 2, Feb. 2004, pp.131-139.
4. A. Srivastava et al., "Modeling and Analysis of Leakage Power Considering Within-Die Process Variations," *Proc. Int'l Symp. Low Power Electronics and Design* (ISLPED 02), IEEE Press, 2002, pp. 64-67.
5. S. Mukhopadhyay and K. Roy, "Modeling and Estimation of Total Leakage Current in Nano-Scaled CMOS Devices Considering the Effect of Parameter Variation," *Proc. Int'l Symp. Low Power Electronics and Design* (ISLPED 03), IEEE Press, 2003, pp. 172-175.
6. F. Najm and N. Menezes, "Statistical Timing Analysis Based on a Timing Yield Model," *Proc. 41st Design Automation Conf.* (DAC 04), ACM Press, 2004, pp. 460-465.
7. S. Choi, B. Paul, and K. Roy, "Novel Sizing Algorithm for Yield Improvement under Process Variation in Nanometer Technology," *Proc. 41st Design Automation Conf.* (DAC 04), ACM Press, 2004, pp. 454-459.
8. S. Zhang, V. Wason, and K. Banerjee, "A Probabilistic Framework to Estimate Full-Chip Subthreshold Leakage Power Distribution Considering Within-Die and Die-to-Die P-T-V Variations," *Proc. Int'l Symp. Low Power Electronics and Design* (ISLPED 04), IEEE Press, 2004, pp. 156-161.
9. Y.-F. Tsai et al., "Impact of Process Scaling on the Efficacy of Leakage Reduction Schemes," *Int'l Conf. Integrated Circuit Design and Technology* (ICICDT 04), IEEE Press, 2004, pp. 3-11.

work" sidebar summarizes work in this area.

In this article, we develop a complete stochastic model for leakage current that includes effects from multiple sources of variability and captures the leakage current distribution's dependence on operating frequency. We consider the contribution of both interdie and intradie process variations, and we model total leakage as consisting of both subthreshold and gate-tunneling leakage. We derive a closed-form expression for total leakage as a function of all relevant process parameters. We also present an analytical equation to quantify yield loss for a design when a power limit is specified. This method precludes the need to use circuit simulation to

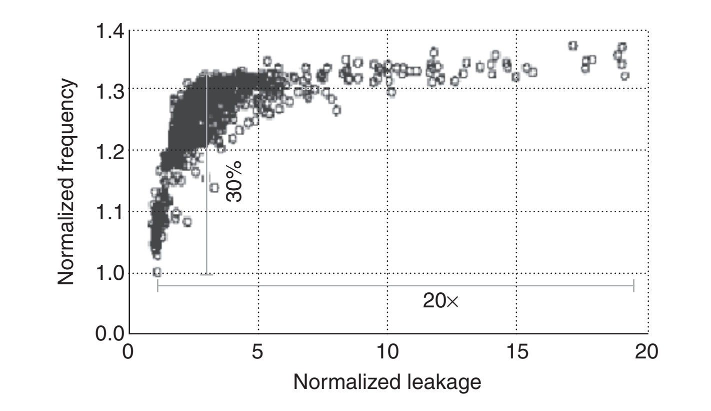

**Figure 1. Leakage and frequency variations (source: Intel).**

characterize a chip's leakage current and lets the designer budget for yield loss before the chip goes into production. We compare the proposed analytical expression with a Monte Carlo simulation, using Spice simulation of a large circuit block to demonstrate its accuracy. Finally, we present a sample yield calculation for a design while considering both power and frequency constraints.

## Full-chip leakage model

Now we present an analytical model to determine total leakage current expended by a chip. We model leakage current as a function of different process parameters. First, the total leakage is a sum of the subthreshold and gate leakages:

$$I_{\text{tot}} = I_{\text{sub}} + I_{\text{gate}}$$

Recently, researchers have noted that other types of leakage current, such as band-to-band tunneling, might become prominent in future process technologies.[3] Although we don't model other types of leakage in this article, our analysis can be easily extended to include additional leakage components.

In the following sections, we model each type of leakage separately. We express both types of leakage current as a product of the nominal value and a multiplicative function that represents the deviation from the nominal value caused by process variability:

$$I_{\text{leakage}} = (I_{\text{nominal}}) f(\Delta P)$$

where $P$ is the process parameter that affects leakage current $I_{\text{leakage}}$. In general, $f$ is a nonlinear function. Because estimation methods based directly on Berkeley simulation (BSIM) models are often too complex (http://www-device.eecs.berkeley.edu/~bsim3/bsim4.html), we use carefully chosen empirical equations in our analysis to provide efficiency and accuracy.

We further decompose parameter $P$ into two components:

$$\Delta P = \Delta P_{\text{global}} + \Delta P_{\text{local}} \tag{1}$$

where $\Delta P_{\text{global}}$ models global (die-to-die, or interdie) process variations, and $\Delta P_{\text{local}}$ represents local (within-die, or intradie) process variations. In a typical manufacturing process, both $\Delta P_{\text{global}}$ and $\Delta P_{\text{local}}$ are modeled as independent normal random variables, making $\Delta P$ also a normal random variable. Because we are dealing only with the deviation from the nominal value, $\Delta P$ is a zero-mean variable. If $P$ is the effective channel length, $\Delta P_{\text{local}}$ is the term for so-called across-chip line-width variations. For simplicity, we denote $\Delta P_{\text{global}}$ as $P_{\text{g}}$, and $\Delta P_{\text{local}}$ as $P_{\text{l}}$.

### Subthreshold leakage

Subthreshold leakage current $I_{\text{sub}}$ is the source-drain current in the transistor when the channel is turned off. $I_{\text{sub}}$ has an exponential relationship with the device's threshold voltage $V_{\text{th}}$, as Equation 2 shows:

$$I_{\text{sub}} = (I_{\text{nominal}}) e^{\left[f(\Delta V_{\text{th}})\right]} \tag{2}$$

For the 0.13-μm technology node, even small variations of $V_{\text{th}}$ can therefore result in leakage numbers that differ by a factor of 5 to 10 from the nominal value.

Threshold voltage is a technology-dependent variable that must be expressed as a function of several parameters. The standard BSIM4 description models device characteristics (such as short channel effect, drain-induced barrier lowering, and narrow-width effect) that influence $V_{\text{th}}$. This description expresses $V_{\text{th}}$ as a function of process parameters, including effective channel length $L_{\text{eff}}$, doping concentration $N_{\text{sub}}$, and oxide thickness $T_{\text{ox}}$ (http://www-device.eecs.berkeley.edu/~bsim3/bsim4.html). Among these parameters, $L_{\text{eff}}$ variation has the greatest impact.[6] A second-order but still significant portion of $V_{\text{th}}$ variation results from fluctuations in doping concentration that cause different values of flat-band voltage $V_{\text{fb}}$ for different transistors on the chip.[3] Finally, oxide thickness is a fairly well-controlled process parameter, and because subthreshold leakage current's sensitivity to oxide thickness is very small,[3,6] we don't include $T_{\text{ox}}$ in the list of parameters that influence $I_{\text{sub}}$. Hence, we empirically model $V_{\text{th}}$ variation as an algebraic sum of two terms—variation in the device's effective channel length $f(\Delta L_{\text{eff}})$, and $V_{\text{th}}$ variation due to doping concentration $f(\Delta V_{\text{th},N\text{sub}})$:

$$f(\Delta V_{\text{th}}) = f[\Delta L_{\text{eff}}] + f[\Delta(V_{\text{th},N\text{sub}})]$$

In our approach, we model $\Delta L_{\text{eff}}$ and $\Delta V_{\text{th},N\text{sub}}$ as independent normal random variables. Although there is a minor dependency between these two variables, we found the amount of error introduced by this independence assumption negligible.

Previously, researchers modeled leakage as a single exponential function of effective channel length,[7] but a polynomial exponential model is far more accurate in capturing the dependency of leakage on effective channel length.[6] Hence, we use a quadratic exponential model to express $\Delta L_{\text{eff}}$. On the other hand, for $f(\Delta V_{\text{th},N\text{sub}})$, we determined from circuit simulations that a linear exponential model is sufficient. For simplicity, we denote $\Delta L_{\text{eff}}$ as $L$, and $\Delta V_{\text{th},N\text{sub}}$ as $V$. Hence, we rewrite Equation 2 as

$$f\left(\Delta L_{\text{eff}}\right) = \frac{-\left(L + c_2 L^2\right)}{c_1} \quad f\left(\Delta V_{\text{th},N_{\text{sub}}}\right) = -\left(\frac{c_3}{c_1}\right)V$$

$$I_{\text{sub}} = (I_{\text{sub,nom}})e^{-\left[\frac{L + c_2 L^2 + c_3 V}{c_1}\right]}$$

Here, $c_1$, $c_2$, and $c_3$ are fitting parameters, and $I_{\text{sub,nom}}$ is the device's subthreshold leakage in the absence of any variability. The negative sign in the exponent indicates that transistors with shorter channel length and lower threshold voltage produce higher leakage current.

Using Equation 1, we decompose $L$ and $V$ into local ($L_{\text{l}}$, $V_{\text{l}}$) and global ($L_{\text{g}}$, $V_{\text{g}}$) components and write the $I_{\text{sub}}$ equation as follows:

$$I_{\text{sub}} = (I_{\text{sub,nom}})e^{-\left[\frac{L_{\text{g}} + c_2 L_{\text{g}}^2 + c_3 V_{\text{g}}}{c_1}\right]} e^{-\left[\frac{L_{\text{l}} + \lambda_2 L_{\text{l}}^2 + \lambda_3 V_{\text{l}}}{\lambda_1}\right]} \tag{3}$$

$I_{\text{sub}}$ is the subthreshold leakage of a single device with unit width. The mapping from $c_i$ to $\lambda_i$ (for $i = 1, 2, 3$) is $\lambda_i = \psi c_i$, where $\psi = 1/(1 + 2c_2 L_{\text{g}})$. By definition, $L_{\text{l}}$, the within-die channel length variation, is a zero-mean random variable. A process parameter's within-die variation typically consists of a systematic (correlated, layout-dependent) component and a random (independent) component. For designs consisting of at least 250 individual clusters, with a die area of 2.5 mm$^2$, we can ignore the effect of spatial correlation on the total chip leakage current.[6] We use this result in our subsequent analyses and assume that a process parameter's within-die variation is entirely due to the random (independent) component.

To calculate a chip's total subthreshold leakage, we must add the leakages device by device, considering that each device has unique random variables $L_{\text{l}}$ and $V_{\text{l}}$ and shares the same random variables $L_{\text{g}}$ and $V_{\text{g}}$ with all other devices. From Equation 3, $I_{\text{sub}} = g(L_{\text{l}})$, a single transistor's subthreshold leakage distribution expressed as a function of $L_{\text{l}}$ can be considered a random variable. The total subthreshold leakage is then a sum of all these individual random variables (RVs). If the number of RVs is large enough, the variance of their sum approaches 0. Consequently, we use the central limit theorem to approximate this sum's distribution with a single deterministic number. Furthermore, if the number of RVs is large enough, the single number will be the mean of the sum's distribution. Because modern CMOS designs contain millions of devices distributed over a relatively large chip area, we use the independence assumption and substitute the sum of leakages over all devices with the mean value of $I_{\text{sub}}$ over the complete range of $L_{\text{l}}$. This mean value is a simple scaling factor that describes the relation between $I_{\text{sub}}$ and $L_{\text{l}}$. Local variations are often spatially correlated, meaning devices positioned close together have a positive correlation. However, as long as the die has sufficient independent regions (typically the case for gigascale designs), the central limit theorem is applicable. We use a similar method to calculate the scaling factor for each process parameter's local variability.

To calculate the scaling factor, we must find an exact expression for the expected value (mean) of $I_{\text{sub}}$. Because $I_{\text{sub}}$ is a function of two independent random variables, ($V_{\text{l}}$ and $L_{\text{l}}$), we write a double integral to calculate the mean:

$$E\left[I_{\text{sub}}\right] = \int_{-\infty}^{\infty}\left(\left[\int_{-\infty}^{\infty} g\left(L_{\text{l}}\right)PDF\left(L_{\text{l}}\right)dL_{\text{l}}\right] g\left(V_{\text{l}}\right)PDF\left(V_{\text{l}}\right)dV_{\text{l}}\right)$$

$$g\left(L_{\text{l}}\right) = (I_{\text{sub,nom}})e^{-\left[\frac{L_{\text{g}} + c_2 L_{\text{g}}^2}{c_1}\right]} e^{-\left[\frac{L_{\text{l}} + \lambda_2 L_{\text{l}}^2}{\lambda_1}\right]}$$

$$g\left(V_{\text{l}}\right) = e^{-\left[\frac{c_3 V_{\text{g}}^2}{c_1}\right]} e^{-\left[\frac{\lambda_3 V_{\text{l}}^2}{\lambda_1}\right]}$$

In this equation, the terms containing $L_{\text{g}}$ and $V_{\text{g}}$ are constant for a given chip. PDF($L_{\text{l}}$) refers to the probability density function of the parameter $L_{\text{l}}$. We can solve the integrals in closed form, and we obtain $I_{\text{sub}} \approx E[I_{\text{sub}}] = S_{\text{L}} S_{\text{V}} I_{\text{Lg,Vg}}$:

$$S_{\text{L}} = \left[1/\left(\sqrt{1 + \frac{2\lambda_2}{\lambda_1}\sigma_{Ll}^2}\right)\right] e^{\left[\sigma_{Ll}^2/\left(2\lambda_1^2 + 4\sigma_{Ll}^2\lambda_1\lambda_2\right)\right]}$$

$$S_{\text{V}} = e^{\left(\lambda_3^2\sigma_{\text{Vl}}^2/2\lambda_1^2\right)}$$

$$I_{\text{Lg,Vg}} = I_{\text{sub,nom}} e^{-\left[\frac{L_{\text{g}} + c_2 L_{\text{g}}^2}{c_1}\right]} e^{-\left[\frac{c_3 V_{\text{g}}}{c_1}\right]} \tag{4}$$

where $S_{\text{L}}$ and $S_{\text{V}}$ are scale factors introduced by local variability in $L$ and $V$. $I_{\text{Lg,Vg}}$ corresponds to subthreshold leakage as a function of global variation. Equation 4 provides the average value of subthreshold leakage for a unit width device. To compute total chip subthreshold leakage, we must perform a weighted sum of the leakages of all devices by considering the device widths as weights. For complex gates (transistor stacks and registers), we use a scale factor ($k$) model[7] to predict the effect of total device width:

$$I_{\text{c,sub}} = S_{\text{L}} S_{\text{V}} \left[ \sum_d \left( W_d / k \right) I_{\text{Lg,Vg}} \right] \tag{5}$$

Here the term

$$\sum_d \left( W_d / k \right) I_{\text{Lg,Vg}}$$

represents chip-level subthreshold leakage as a function of global process parameters $L_g$ and $V_g$.

### Gate leakage

As the oxide thickness of a transistor is scaled, the number of carriers that can tunnel through the thin gate oxide increases. This phenomenon leads to the presence of gate leakage current ($I_{\text{gate}}$) between the gate and the substrate, as well as between the gate and the channel. $I_{\text{gate}}$ is linearly dependent on the device's area and has a highly exponential relationship with oxide thickness $T_{\text{ox}}$. Because the $T_{\text{ox}}$ variation has by far the greatest impact on gate leakage, we model $I_{\text{gate}}$ as

$$I_{\text{gate}} = (I_{\text{nom}}) e^{\left[ f(\Delta T_{\text{ox}}) \right]}$$

From circuit simulations, we found that expressing $f(\Delta T_{\text{ox}})$ as a simple linear function is sufficient. A suitable value for a single parameter $\beta_1$ efficiently captures the highly exponential relationship. We let $\Delta T_{\text{ox}} = T$. Using Equation 1, we decompose $T$ into global ($T_{\text{g}}$) and local ($T_{\text{l}}$) components:

$$f\left( \Delta T_{ox} \right) = -\left( \frac{T}{\beta_1} \right)$$

$$I_{\text{gate}} = (I_{\text{gate,nom}}) e^{-\left( T_{\text{g}} / \beta_1 \right)} e^{-\left( T_l / \beta_1 \right)}$$

$I_{\text{gate}}$ is the gate leakage current of a single device with unit width. $I_{\text{gate,nom}}$ is the nominal gate leakage, and both $T_{\text{g}}$ and $T_{\text{l}}$ are zero-mean random variables. The relationship between $I_{\text{gate}}$ and $T_{\text{l}}$ is similar to the single exponential relationship between $I_{\text{sub}}$ and $V_{\text{l}}$. Similar to the calculation for $S_{\text{V}}$, we compute scale factor $S_{\text{T}}$, which is caused by $T_{\text{l}}$:

$$I_{\text{gate}} \approx E\left[ I_{\text{gate}} \right] = S_{\text{T}} I_{\text{Tg}}$$
$$S_{\text{T}} = e^{\left( \sigma_{\text{Tl}}^2 / 2\beta_1^2 \right)}$$
$$I_{\text{Tg}} = (I_{\text{gate,nom}}) e^{-\left( T_{\text{g}} / \beta_1 \right)} \tag{6}$$

Using the device widths, we calculate chip-level gate leakage in a manner similar to how we calculate subthreshold leakage:

$$I_{\text{c,gate}} = S_{\text{T}} \left[ \sum_d \left( W_d / k \right) I_{\text{Tg}} \right] \tag{7}$$

### Total leakage

Total leakage is the sum of all the devices' subthreshold and gate leakage currents. In Equations 5 and 7, all the devices on the chip share $I_{\text{Lg,Vg}}$ and $I_{\text{Tg}}$. Hence, we write the equation for total chip leakage as

$$I_{\text{c,tot}} = \left[ \sum_d \left( W_d / k \right) \right] \left[ S_{\text{L}} S_{\text{V}} (I_{\text{Lg,Vg}}) + S_{\text{T}} I_{\text{Tg}} \right] \tag{8}$$

We can use this equation to calculate total leakage for different types of devices, such as NMOS/PMOS and low/high-$V_{\text{th}}$ transistors. The differences will be in the fitting parameters and scale factor $k$.

## Yield analysis

Traditional parametric yield analysis of high-performance ICs uses the frequency (or speed) binning method.[8] For a given lot, each chip is characterized according to its operating frequency and figuratively placed in a particular bin according to this value. A frequency limit is specified, and chips that operate at frequencies below this limit are discarded. As Figure 1 showed, chips in the "fast" corner produce far more leakage current than the other chips because of the inverse correlation between leakage and circuit delay. In current technologies, this is a major concern because many of these chips leak more than the acceptable value and must be discarded.[5] Thus, the frequency-binning method exacerbates parametric yield loss, because dies are lost at both the low- and high-speed bins, further narrowing the acceptable process window.

Here, we describe a method of calculating a lot's yield when both frequency and power limits are imposed. We first show that chip frequency is most strongly influenced by global gate length variability, and, hence, as in standard industry practice, each frequency bin corresponds to a specific $L_{\text{g}}$ value. We then compute the yield caused by the imposed leakage limit on a bin-by-bin basis.

### Frequency dependence on process parameters

In principle, IC or processor frequency depends on many process parameters, such as gate length, doping concentration, and oxide thickness. However, our Spice simulations demonstrate that circuit delay is primarily affected by gate length variations. We simulated a 17-stage ring oscillator for different process conditions using the BPTM 100-nm process technology. Figure 2 shows the results. From this plot, we see that variations in $L_g$ significantly influence (±15%) the ring oscillator's delay. Variations in $T_g$ and $V_g$ have little or no impact on delay and thus can be ignored. By a similar argument, it follows that variations in local variability components $V_l$ and $T_l$ also have negligible impact on circuit delay and can be neglected during chip performance analysis.

From this analysis, we've determined that only the global ($L_g$) and local ($L_l$) variability components of effective channel length affect the chip frequency value. Bowman, Duvall, and Meindl pointed out that within-die variations primarily impact the mean of the maximum frequency (*FMAX*), whereas die-to-die variations affect the variance of the maximum frequency distribution,[9] implying that both $L_g$ and $L_l$ determine the chip frequency value. Recently, however, Samaan has pointed out that even for a well-debugged product, in which 50% of all critical paths are within 2.5% of the worst path delay, the effect of within-die variation on maximum chip frequency is minimal.[10] Samaan uses an analytical model on industrial circuits to show that for within-die variation with 3σ = 10%, the *FMAX* value is no worse than 5% of its original value.[6] Although local variations also affect circuit delay, their effect tends to average out over a circuit path, lessening the impact as compared with global variations. Moreover, within-die variation will remain under relative control (according to the ITRS[11]). Thus, it will not affect chip performance by more than a small percentage for at least the next four technology generations. Therefore, we ignore local variations' effect in our yield computation.

We assume a one-to-one correspondence between chip frequency (performance) and the global effective-channel-length variability value $L_g$. This is consistent with current practices, which often assume a one-to-one correspondence between frequency bins and specific gate length values.

## Yield estimate computation

We now discuss the method for computing the expected yield for a particular frequency bin with an imposed leakage limit. For a particular bin, the frequency value and the corresponding value of $L_g$ are available, and using the expressions for $I_{Lg,Vg}$ (Equation 4) and $I_{Tg}$ (Equation 6), we rewrite the equation for total chip leakage (Equation 8) as follows:

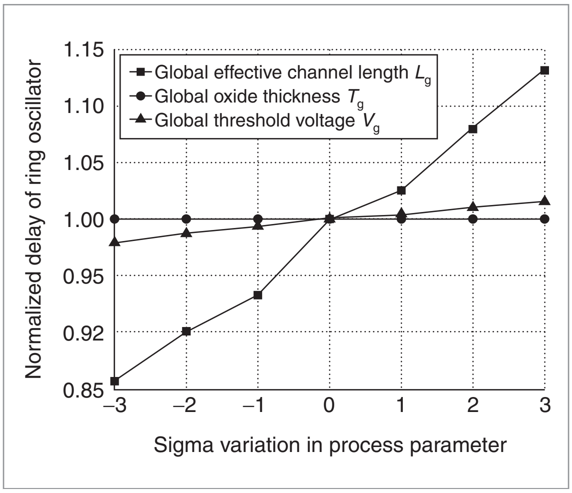

**Figure 2. Relative contribution of parameter variations on ring oscillator delay.**

$$
\begin{aligned}
I_{\text{tot}} &= (A_S)e^{\left(V_g/k_v\right)} + (A_g)e^{\left(T_g/k_t\right)} \\
A_s &= \left(\sum_d \left(W_d / k\right)\right) S_L S_V (I_{\text{sub,nom}})e^{-\left[\left(L_g + c_2 L_g^2\right)/c_1\right]} \\
A_g &= \left(\sum_d \left(W_d / k\right)\right) S_T (I_{\text{gate,nom}}) \\
k_v &= -\left(c_1 / c_3\right) \qquad k_t = -\beta_1
\end{aligned}
\tag{9}
$$

Here we simplified the notation for the fitting parameters and expressed this equation in terms of the new constants $k_v$ and $k_t$. The $k_v$ and $k_t$ values are generally expressed in terms of $\sigma_{V_g}$ and $\sigma_{T_g}$. $A_s$ represents total chip subthreshold leakage at a value of $L_g$ and includes the scale factors due to local variability. Similarly, $A_g$ represents total chip gate leakage at a given $L_g$ value. However, $I_{c,gate}$ is independent of $L_g$, so $A_g$ is not influenced by changes in the $L_g$ value. To plot total leakage versus $L_g$, we first compute $A_s$ and $A_g$ at particular $L_g$ values and then calculate $I_{tot}$ distribution at each of these points.

For every device type, $I_{tot}$ is the sum of two lognormal variables, each representing leakage current. For a particular device, by our formulation, no parameter affects both terms simultaneously. Thus, we can consider these terms as independent random variables. For a given circuit design, total leakage will then be the sum of a small set of lognormals, with each device type contributing exactly two lognormals to the total leakage set. We use Wilkinson's method[12] to model this sum of lognormals as another lognormal random variable (http://mathworld.wolfram.com/LogNormalDistribution.html). Using the independence condition, we set the sums of

**Table 1. Value of $I_{\text{tot}}$ for an $n$-sigma point. $I_{\text{tot}}$ is the sum of two lognormal variables, each representing leakage current.**

| $n$ | $F_x(I_{\text{tot}})$ | $I_{\text{tot}}$ |
|---|---|---|
| 0 | 0.500 | $\exp(\mu_{\text{N,Itot}})$ |
| 1 | 0.682 | $\exp(\mu_{\text{N,Itot}} + 0.473\sigma_{\text{N,Itot}})$ |
| 2 | 0.954 | $\exp(\mu_{\text{N,Itot}} + 1.685\sigma_{\text{N,Itot}})$ |
| 3 | 0.998 | $\exp(\mu_{\text{N,Itot}} + 2.878\sigma_{\text{N,Itot}})$ |

the means and variances as equal to the mean and variance of the new lognormal:

$$
\begin{gathered}
I_{\text{tot}} = X_1 + X_2 \\
X_1 \sim LN\left(\log(A_s),(\sigma_{\text{Vg}}/k_v)^2\right) \quad X_2 \sim LN\left(\log(A_g),(\sigma_{\text{Tg}}/k_t)^2\right) \\
\mu_{I\text{tot}} = \exp\left[\log(A_s)+\frac{1}{2}\left(\frac{\sigma_{\text{Vg}}}{k_v}\right)^2\right] + \exp\left[\log(A_g)+\frac{1}{2}\left(\frac{\sigma_{\text{Tg}}}{k_t}\right)^2\right] \\
\sigma^2_{I\text{tot}} = \begin{array}{l} \exp\left[2\log(A_s)+\left(\dfrac{\sigma_{\text{Vg}}}{k_v}\right)^2\right]\left[\exp\left(\sigma^2_{\text{Vg}}/k_v^2\right)-1\right]+ \\ \exp\left[2\log(A_g)+\left(\dfrac{\sigma_{\text{Tg}}}{k_t}\right)^2\right]\left[\exp\left(\sigma^2_{\text{Tg}}/k_t^2\right)-1\right] \end{array}
\end{gathered}
\tag{10}
$$

We then obtain the mean and variance ($\mu_{\text{N,Itot}}$, $\sigma_{\text{N,Itot}}{}^2$) of the normal random variable corresponding to this lognormal. From these values, we can express the *PDF* of the total leakage using the standard expression for the *PDF* of a lognormal random variable:

$$
\begin{gathered}
\mu_{\text{N,Itot}} = \frac{1}{2}\log\left[\mu^4_{\text{Itot}}/\left(\mu^2_{\text{Itot}}+\sigma^2_{\text{Itot}}\right)\right] \\
\sigma^2_{\text{N,Itot}} = \log\left[1+\left(\sigma^2_{\text{Itot}}/\mu^2_{\text{Itot}}\right)\right] \\
PDF(I_{\text{tot}}) = \left(\frac{1}{I_{\text{tot}}\sqrt{2\pi\sigma^2_{\text{N,Itot}}}}\right)\exp\left[-\left(\frac{\log(I_{\text{tot}})-\mu_{\text{N,Itot}}}{\sigma_{\text{N,Itot}}\sqrt{2}}\right)^2\right]
\end{gathered}
\tag{11}
$$

Finally, to obtain exact yield estimates, we require the quantile numbers for the lognormal distribution described by $I_{\text{tot}}$ (that is, the $I_{\text{tot}}$ confidence points that correspond to the specified leakage limit). The exponential function that relates lognormal distribution $LN(\mu_{\text{Itot}}, \sigma_{\text{Itot}}{}^2)$ to normal distribution $N(\mu_{\text{N,Itot}}, \sigma_{\text{N,Itot}}{}^2)$ is a monotone increasing function, so we map the quantiles of the normal random variable directly to the quantiles of the lognormal random variable. Hence, we write the expression for cumulative distribution function *CDF* of a lognormal variable as

$$
CDF(I_{\text{tot}}) = F_x(I_{\text{tot}}) = \frac{1}{2}\left[1+erf\left(\frac{\log(I_{\text{tot}})-\mu_{\text{N,Itot}}}{\sigma_{\text{N,Itot}}\sqrt{2}}\right)\right]
\tag{12}
$$

In Equation 12, *erf*() is the error function. By setting $F_x(.)$ to a particular confidence point on the normal distribution, we obtain the corresponding value on the lognormal distribution, as Table 1 shows. In Table 1, the 0-sigma point corresponds to the distribution's median.

Conversely, given a limit for $I_{\text{tot}}$, we can use Equation 12 to compute $CDF(I_{\text{tot}})$ and determine the number of chips that meet the leakage limit in a particular performance bin. Thus, in a given frequency bin and for a given leakage limit, $[CDF(I_{\text{tot}})100]\%$ is the fraction of chips that meet both the speed and power criteria. By repeating this computation for each frequency bin that meets the frequency specification, we find the total percentage of chips that meet both the leakage and performance constraints.

## Results

Next, we use our analytical method to predict a lot's yield. Our circuit of choice is a 64-bit adder written for the Alpha architecture. We assume that all dies in the lot consist of this circuit, and we use a small ring oscillator circuit to characterize the chip's frequency with the $L_g$ variation. We use the 100-nm ($L_{\text{eff}}$ = 60 nm) Berkeley Predictive Technology Model for our Spice Monte Carlo simulations.[4] We also use a gate leakage model based on the BSIM4 equations (http://www-device.eecs.berkeley.edu/~bsim3/bsim4.html). The variability numbers for $\Delta L_{\text{eff}}$, $\Delta V_{\text{th,Nsub}}$, and $\Delta T_{\text{ox}}$ are based on estimates obtained from an industrial 90-nm process.

Table 2 presents a quantitative comparison between Spice data and our analytical method. We consider three cases: no variability in any parameter, die-to-die variability only, and both within-die and die-to-die variability in all three parameters. The middle three columns list the sigma variation values corresponding to each parameter. Thus, for the case in which both types of variability are present, the global variability values for all three parameters are set to $-1\sigma$ from the nominal value, and the local variability is set to $\pm3\sigma$. The table shows that for all cases, the difference between the experimental Monte Carlo simulation data (Exp) and the analytical expressions (Ana) is less than 5%. Moreover, the presence of local variability increases the amount of total chip leakage by about 15%.

Figure 3 shows a scatter plot for 2,000 samples of the Spice-generated total circuit leakage. The $y$-axis is normalized to the sample mean of the leakage currents. For a ±3$\sigma$ variation in $L_g$, there is a 14× spread in leakage. Additionally, for a given $L_g$, there is a wide local distribution in leakage. For instance, given $L_g = 0\sigma$, the normalized value of total circuit leakage is between 0.5 and 1.7. In Equation 9, we observe that even for small $(V/k_v)$ and $(T/k_t)$ values, the exponential terms increase rapidly and contribute a larger portion to the total leakage value. As a result, the distribution in $V_g$ and $T_g$ (for each $L_g$ value) produces the bandlike curve we see in Figure 3 (instead of a single curve). This is significant because for a given $L_g$ value (and hence a given operating frequency), a large portion of chips can dissipate about three times the nominal leakage. A chip that operates at an acceptable frequency might still have to be discarded because the variability in $V_g$ and $T_g$ pushes its leakage consumption over the tolerable limit. Thus, we see that the secondary variations $V_g$ and $T_g$ play a significant role in determining a lot's yield.

We now present an example yield calculation. For the lot presented here, we impose frequency limit +1$\sigma$ and normalized power limit ($P_{\text{lim}}$) = 1.75, as Figure 4 shows. We specify the frequency bins at the $L_g$ $n$-sigma boundaries as $f_{(Lg=-3)} > f_{(Lg=-2)} > \dots > f_{(Lg=+1)}$. We assign all chips with frequencies in the range [$f_{(Lg=i)}$, $f_{(Lg=i-1)}$] to the bin characterized by frequency $f_{(Lg=i)}$. This is a conservative estimate because we assume that all devices in a bin are operating at that bin's minimum possible frequency value. First, because of the performance (frequency) limit, we discard all chips that operate at frequencies less than +1$\sigma$. As the plot shows, although these chips meet the power criterion, they are discarded (or "rejected") because they are too slow.

Next, we calculate each bin's yield, proceeding bin by bin. To illustrate this computation, we present the numbers for cases in which $L_g = [-3\sigma, -2\sigma, \dots, +1\sigma]$.

Table 2. Comparison of Spice simulation and analytical data. Exp refers to experimental data and Ana refers to analytical values.

| | Parameter sigma (σ) values | | | Mean leakage (μA) | |
|---|---|---|---|---|---|
| Case | ($L_g$, $L_l$) | ($V_g$, $V_l$) | ($T_g$, $T_l$) | Exp | Ana |
| No variation | (0, 0) | (0, 0) | (0, 0) | 14.97 | 15.22 |
| Die-to-die only | (–1, 0) | (–1, 0) | (–1, 0) | 20.82 | 21.32 |
| Both variations | (–1, ±3) | (–1, ±3) | (–1, ±3) | 24.01 | 24.95 |

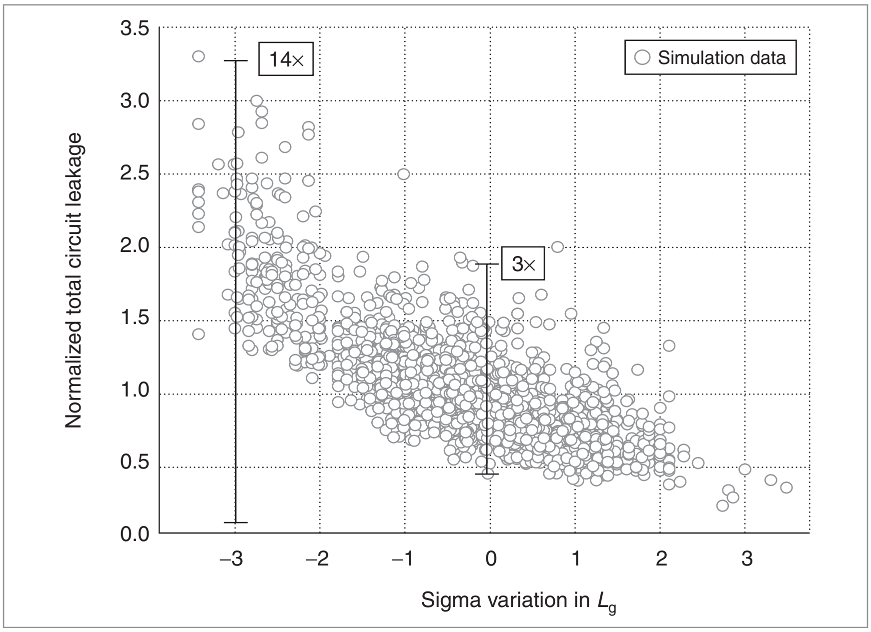

Figure 3. Scatter plot of total circuit leakage distribution.

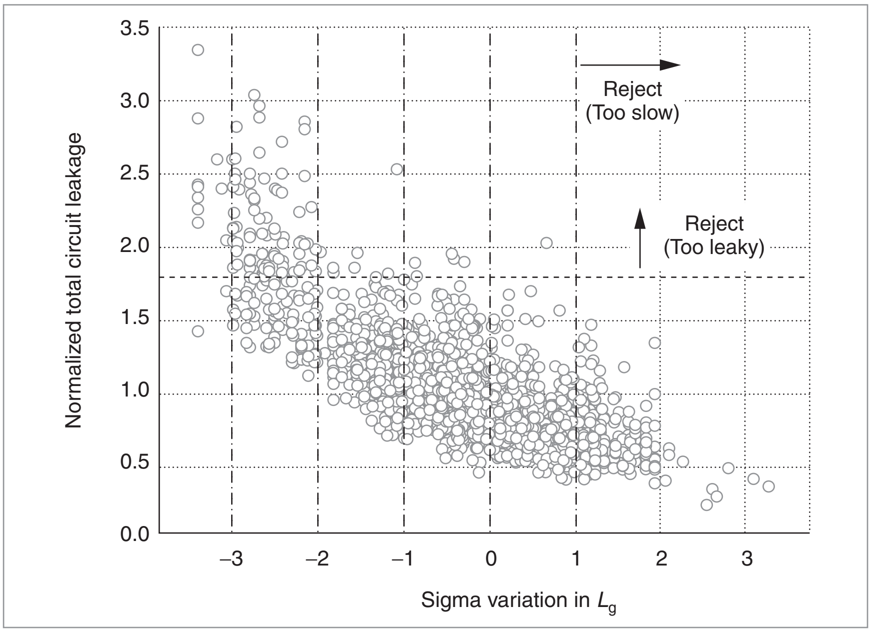

Figure 4. Scatter plot for lot with specified power and performance limits.

**Table 3.** ***CDF*[($I_{tot}$)100]% numbers for $P_{lim}$ = 1.75 and the given range of $L_g$ values.**

|  | $L_g$ *n*-sigma |  |  |  |  |
|---|---|---|---|---|---|
| $P_{lim}$ | –3 | –2 | –1 | 0 | 1 |
| 1.75 | 43.6 | 72.6 | 90.5 | 97.5 | 99.4 |

For each such $L_g$, we calculate the *CDF* values using Equations 9 through 12. Table 3 summarizes these *CDF* numbers for $P_{lim} = 1.75$.

Traditional parametric yield analysis does not consider power as a criterion and hence overestimates the number of chips that are actually good or sellable. For instance, if $P_{lim} = 1.75$, we see from Table 3 that for $L_g = -2\sigma$, only 72.6% of the chips meet the power criterion. Thus, even if the chip designer budgets for 1.75 times the nominal power, there is a loss of 27.4% of the chips operating in the fast corner. Furthermore, even for the nominal value of $L_g = 0\sigma$, about 2.5% of the chips are lost because they lie outside the power limit. Whereas a typical frequency-binning method would predict that 100% of the chips with $L_g = -2\sigma$ are good, our method captures the fact that over 25% cannot be marketed. This is particularly important because fast bin devices are highly profitable and determine the pricing model for a chip company. Hence, ASIC designers need to adopt an integrated approach that accounts for the compounded loss caused by both the power and the performance constraints. We find that our approach always predicts a lower yield percentage than methods that assume independence of these limiting factors. By preserving the correlation between frequency and leakage, we obtain more accurate yield estimates.

**THIS ARTICLE PRESENTS** a systematic methodology for chip-level leakage power estimation and analytical yield prediction while considering multiple sources of parameter variability. The yield equations presented here enable chip designers to accurately estimate the parametric yield of digital circuits. ■

## Acknowledgments

This work was supported in part by the National Science Foundation, the Semiconductor Research Corporation, the Defense Advanced Research Projects Agency/Gigascale Silicon Research Center (DARPA/GSRC), IBM, and Intel. We thank Vivek De of Intel for providing us with Figure 1.

## ■ References

1. J. Kao, S. Narendra, and A. Chandrakasan, “Subthreshold Leakage Modeling and Reduction Techniques,” *Proc. IEEE/ACM Int’l Conf. Computer-Aided Design* (ICCAD 02), IEEE Press, 2002, pp. 141-148.
2. S. Borkar, “Design Challenges of Technology Scaling,” *IEEE Micro*, vol. 19, no. 4, Jul.-Aug. 1999, pp. 23-29.
3. S. Mukhopadhyay, A. Raychowdhury, and K. Roy, “Accurate Estimate of Total Leakage Current in Scaled CMOS Circuits Based on Compact Current Modeling,” *Proc. 40th Design Automation Conf.* (DAC 03), ACM Press, 2003, pp. 169-174.
4. “Berkeley Predictive Technology Model (BPTM),” http://www-device.eecs.berkeley.edu/~ptm/.
5. S. Borkar et al., “Parameter Variations and Impact on Circuits and Microarchitecture,” *Proc. 40th Design Automation Conf.* (DAC 03), ACM Press. 2003, pp. 338-342.
6. R. Rao et al., “Statistical Analysis of Subthreshold Leakage Current for CMOS Circuits,” *IEEE Trans. Very Large Scale Integrated Systems*, vol. 12, no. 2, Feb. 2004, pp. 131-139.
7. S. Narendra et al., “Full-Chip Subthreshold Leakage Power Prediction Model for Sub-0.18μm CMOS,” *Proc. Int’l Symp. Low Power Electronics and Design* (ISLPED 02), IEEE Press, 2002, pp. 19-23.
8. B. Cory, R. Kapur, and B. Underwood, “Speed Binning with Path Delay Test in 150-nm Technology,” *IEEE Design & Test*, vol. 20, no. 5, Oct. 2003, pp. 41-45.
9. K. Bowman, S. Duvall, and J. Meindl, “Impact of Die-to-Die and Within-Die Parameter Fluctuations on the Maximum Clock Frequency Distribution for Gigascale Integration,” *IEEE J. Solid-State Circuits*, vol. 37, no. 2, Feb. 2002, pp. 183-190.
10. S. Samaan, “The Impact of Device Parameter Variations on the Frequency and Performance of VLSI Chips,” *Proc. ACM/IEEE Int’l Conf. Computer Aided Design* (ICCAD 04), IEEE Press, Nov. 2004, pp. 343-346.
11. *International Technology Roadmap for Semiconductors*, Semiconductor Industry Assoc., 2001.
12. N.C. Beaulieu, A.A. Abu-Dayya, and P.J. Mclane, “Estimating the Distribution of a Sum of Independent Lognormal Random Variables,” *IEEE Trans. Communications*, vol. 43, no. 12, Dec. 1995, pp. 2869-2873.

**Rajeev R. Rao** is pursuing a PhD in computer science and engineering at the University of Michigan, Ann Arbor. His research interests include modeling and analysis of robust, low-power VLSI designs and variability-aware circuit approaches. Rao has a BS in electrical and computer engi-

neering from Rutgers University, and an MS in computer science and engineering from the University of Michigan, Ann Arbor.

**David Blaauw** is an associate professor of computer science and engineering at the University of Michigan. His research interests include VLSI design and CAD with particular emphasis on circuit design and optimization for high-performance and low-power designs. Blaauw has a BS in physics and computer science from Duke University and an MS and a PhD, both in computer science, from the University of Illinois, Urbana-Champaign.

**Dennis Sylvester** is an assistant professor of electrical engineering and computer science at the University of Michigan, Ann Arbor. His research interests include low-power circuit design and design automation, design for manufacturability, and on-chip interconnect modeling. Sylvester has a PhD in electrical engineering from the University of California, Berkeley. He is a senior member of the IEEE.

**Anirudh Devgan** is vice president of product development at Magma Design Automation in Austin, Texas. His research interests include CAD of ICs with emphasis on electrical analysis and simulation, physical design, low-power design, and design for manufacturability and variability. Devgan has a BTech degree in electrical engineering from the Indian Institute of Technology, Delhi, and an MS and a PhD in electrical and computer engineering from Carnegie Mellon University. He is a senior member of the IEEE.

■ Direct questions and comments about this article to Rajeev R. Rao, University of Michigan, Dept. of Electrical Engineering and Computer Science, Ann Arbor, MI 48109; rrrao@eecs.umich.edu.

**For further information on this or any other computing topic, visit our Digital Library at http://www.computer.org/publications/dlib.**