Parametric Yield Estimation Considering Leakage Variability

Rajeev Rao, *Anirudh Devgan, David Blaauw, Dennis Sylvester
University of Michigan, EECS Department, Ann Arbor, MI 48109
*IBM Corporation, 11501 Burnet Road, Austin, TX 78758
IBM Technical Contact: Sani Nassif, IBM Austin Research Lab, Austin, TX 78758

Abstract

Leakage current has become a stringent constraint in today's processor designs in addition to traditional constraints on frequency. Since leakage current exhibits a strong inverse correlation with circuit delay, effective parametric yield prediction must consider the dependence of leakage current on frequency. In this paper, we present a new chip-level statistical method to estimate the total leakage current in the presence of within-die and die-to-die variability. We develop a closed-form equation for total chip leakage that models the dependence of the leakage current distribution on different process parameters. The proposed analytical expression is obtained directly from pertinent design information and includes both subthreshold and gate leakage currents. Using this model, we then present an integrated approach to accurately estimate the yield loss when both a frequency and power limits are imposed on a design. Our method demonstrates the importance of considering both these limiters in calculating the yield of a lot.

1 Introduction

Continued scaling of device dimensions combined with shrinking threshold voltages has enabled designers to produce integrated circuits (ICs) that contain hundreds of millions of devices. However, this has also resulted in an exponential rise of IC power dissipation. This increase is primarily due to leakage which is emerging as a significant portion of the total power consumption. It is estimated that the subthreshold leakage power will account for over 50% of the total power for portable applications developed for the 65nm technology node [1]. In future technologies, aggressive scaling of the oxide thickness will lead to significant gate oxide tunneling current, further aggravating the leakage problem. It is estimated that across successive technology generations, subthreshold leakage increases by about 5X [2] while gate leakage can increase by as much as 30X.

At the same time, the increased presence of parameter variability in modern designs has accentuated the need to consider the impact of statistical leakage current variations during the design process. For ±10% variation in the effective channel length of a transistor, there can be up to a 3X difference in the amount of subthreshold leakage current [3]. Gate leakage current exhibits an even greater sensitivity to process variations, showing a 15X difference in current for a 10% variation in oxide thickness in a 100nm BPTM process technology [4]. Hence, considerable variability in chip level leakage current can be expected and measured variations as high as 20X have been reported in the literature [5].

In current designs, the yield of a lot is typically calculated by characterizing the chips according to their operating frequency. The subset of dies that do not meet the required performance constraint are rejected, making this aspect of the design process very important from a commercial point of view. However, it has been observed [5] that among the "good" chips that meet the performance constraint, a substantial fraction dissipate very large amounts of leakage power and thus are unsuitable for commercial usage. This is due to the inverse correlation between circuit delay and leakage current. Although the delay is reduced for devices with channel lengths smaller than the nominal value, it has the negative effect of vastly increasing the leakage current resulting in higher leakage dissipation for chips with high operating frequencies.

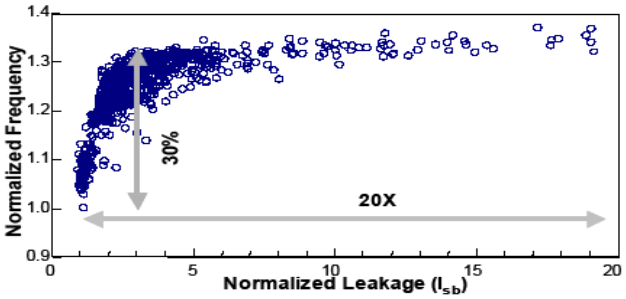

Figure 1. Leakage and frequency variations (Source: Intel)

This inverse correlation is illustrated in Figure 1, which shows the distribution of chip performance and leakage based on silicon measurements over a large number of samples of a high-end processor design [5]. As can be seen, both the mean and variance of the leakage distribution increase significantly for chips with higher frequencies. This trend is particularly troubling since it substantially reduces the yield of designs that are both performance and leakage constrained. Hence, there is a need for accurate leakage yield prediction methods that model this dependency.

Several statistical methods have been suggested to estimate the full chip leakage current. In [6], the authors consider within-die threshold voltage variability to estimate the full chip subthreshold leakage current. A compact current model is used in [7] to estimate the total leakage current. In [8], the authors present analytical equations to model subthreshold leakage as a function of the channel length of the transistor. A moment-based approximation approach is used to estimate the mean and variance of leakage current in [9] and [10]. However, none of these methods provide exact mathematical equations to express the chip leakage and furthermore, they do not consider the dependence of leakage on frequency.

In this paper we develop a complete stochastic model for leakage current that includes the effects from multiple sources of variability and captures the dependence of the leakage current distribution on operating frequency. We consider the contribution from both inter- and intra-die process variations and model total leakage as consisting of both subthreshold and gate tunneling leakage. We derive a closed-form expression for the total leakage as a function of all relevant process parameters. We also present an analytical equation to quantify the yield loss when a power limit is imposed. This method precludes the need to use circuit simulation to characterize the leakage current of a chip and enables the designer to budget for yield loss before the chip is sent to production. The proposed analytical expression is then compared with Monte-Carlo simulation using SPICE simulation of a large circuit block to demonstrate its accuracy. Finally, we construct yield curves to accurately estimate the number of chips that satisfy both power and frequency constraints.

The remainder of this paper is organized as follows. In Section 2, we present the model for full chip total leakage. The models for subthreshold and gate leakages are presented separately. In Section 3

we derive analytical equations to describe the yield prediction of a lot based on our leakage model. In Section 4 we present results and in Section 5 we conclude the paper.

2 Full Chip Leakage Model

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

$$I_{tot} = I_{sub} + I_{gate} \quad \text{(EQ 1)}$$

Recently, it has been noted that other types of leakage current, such as Band-to-Band Tunneling (BTBT), may become prominent in future process technologies [7]. Although we do not model other types of leakage in this paper, our analysis can be easily extended to include these additional leakage components.

In the subsequent 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 due to process variability.

$$I_{leakage} = I_{nominal} \cdot f(\Delta P), \quad \text{(EQ 2)}$$

where P is the process parameter that affects the leakage current $I_{leakage}$. In general, f is a non-linear function. Since estimation methods based directly on BSIM models [11] are often overly complex, we use carefully chosen empirical equations in our analysis to provide both efficiency and accuracy.

We further decompose parameter P into two components:

$$\Delta P = \Delta P_{global} + \Delta P_{local}, \quad \text{(EQ 3)}$$

where ΔP_{global} models the global (die-to-die or inter-die) process variations while ΔP_{local} represents the local (within-die or intra-die) process variations. In a typical manufacturing process, both ΔP_{global} and ΔP_{local} are generally modeled as independent normal random variables making ΔP also a normal random variable. Since we are only dealing with the deviation from nominal, ΔP is a zero mean variable. If P is the effective channel length, then ΔP_{local} is the term for the so-called Across Chip Length Variations (ACLV). For simplicity of notation, we let $\Delta P_{global}=P_g$ and $\Delta P_{local}=P_l$.

2.1 Subthreshold Leakage

Subthreshold leakage current (I_{sub}) refers to the source-to-drain current when the transistor has been turned "off". As is well known, I_{sub} has an exponential relationship with the threshold voltage V_{th} of the device as shown below in EQ4.

$$I_{sub} = I_{nominal} \cdot e^{[f(\Delta V_{th})]} \quad \text{(EQ 4)}$$

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

Threshold voltage is a technology-dependent variable that must be expressed as a function of a number of parameters. The standard BSIM4 description [11] models several device characteristics (short channel effect, drain-induced barrier lowering (DIBL), narrow-width effect, etc.) that influence V_{th} and expresses it as a function of several process parameters including effective channel length (L_{eff}), doping concentration (N_{sub}), and oxide thickness (T_{ox}). Among these parameters, the variation in L_{eff} has the greatest impact as noted in [8]. A second order, but still significant portion of the variation in V_{th} occurs due to fluctuations in doping concentration that result in different values of the flat band voltage V_{fb} for different transistors on the chip [10]. Finally, oxide thickness is a fairly well-controlled process parameter and does not influence subthreshold leakage significantly. Hence, we empirically model the variation in V_{th} as an algebraic sum of two terms:

1. $f(\Delta L_{eff})$ = the variation in effective channel length of the device
2. $f(\Delta V_{th,\,Nsub})$ = the variation in V_{th} due to doping concentration

$$f(\Delta V_{th}) = f[\Delta L_{eff}] + f[\Delta (V_{th,\,Nsub})] \quad \text{(EQ 5)}$$

In our approach we model ΔL_{eff} and $\Delta V_{th,Nsub}$ as independent normal random variables. While there is a minor dependency between these two variables the amount of error introduced as a result of this independence assumption was found to be negligible.

Previously, leakage had been modeled as a single exponential function of the effective channel length [6], but as the authors show in [8] a polynomial exponential model is much more accurate in capturing the dependency of leakage on effective channel length. Hence, we use a quadratic exponential model to express $f(\Delta L_{eff})$. On the other hand, for $f(\Delta V_{th,Nsub})$ we determined from circuit simulations that a linear exponential model is sufficient. For simplicity of notation, let $\Delta L_{eff}=L$ and $\Delta V_{th,Nsub}=V$. Using this we can rewrite EQ4:

$$f(\Delta L_{eff}) = \frac{-(L + k_2 L^2)}{k_1} \qquad f(\Delta V_{th,\,Nsub}) = -\left(\frac{k_3}{k_1}\right) V$$
$$I_{sub} = I_{sub,\,nom} \cdot e^{-\left[\frac{L + k_2 L^2 + k_3 V}{k_1}\right]} \quad \text{(EQ 6)}$$

Here, k_1, k_2, k_3, are fitting parameters and $I_{sub,nom}$ is the subthreshold leakage of the device in the absence of any variability. The negative sign in the exponent is indicative of the fact that transistors with shorter channel lengths and lower threshold voltage produce higher leakage current.

Using EQ3 we decompose L and V into local (L_l, V_l) and global (L_g, V_g) components and we write the I_{sub} equation as follows:

$$I_{sub} = I_{sub,\,nom} \cdot e^{-\left[\frac{L_g + \lambda_2 L_g^2 + \lambda_3 V_g}{\lambda_1}\right]} \cdot e^{-\left[\frac{L_l + \lambda_2 L_l^2 + \lambda_3 V_l}{\lambda_1}\right]} \quad \text{(EQ 7)}$$

I_{sub} is the subthreshold leakage of a single device with unit width. The mapping from k_i to λ_i (for i=1,2,3) is given by $\lambda_i=\psi k_i$ where $\psi = 1/(1+2k_2L_g)$. By definition L_l, which is the within-in chip channel length variation, is a zero mean random variable. It is characterized by its standard deviation σ_{Ll}. Similarly, V_l is a zero mean random variable and is characterized solely by its standard deviation σ_{Vl}. The equations for the probability density function (PDF) of L_l and V_l are the standard equations for PDF of a normal random variable and are given in EQ20 in the Appendix.

To calculate the total subthreshold leakage for a chip we need to add the leakages device-by-device, considering that each device has unique random variables L_l and V_l, while sharing the same random variables L_g and V_g with all other devices. EQ7 suggests the well-known fact that the subthreshold leakage distribution of a single transistor has a lognormal distribution. The total subthreshold leakage is then a sum of all these individual (possibly dependent) lognormals. However, if the number of lognormals is large enough the variance of their sum decreases gradually to zero. Consequently, we use the Central Limit Theorem to approximate the distribution of

this sum of lognormals by a single number. Further, if the number of lognormals is large enough, the single number will be the mean of the distribution of this sum. Modern CMOS designs contain millions of devices that are distributed over a relatively large area on the chip. Thus, we use the independence assumption and substitute the sum of leakages over all devices with the mean value of I_{sub} over the complete range of L_l. This mean value is a simple scaling factor that describes the relation between I_{sub} and L_l. We use a similar method to calculate the scaling factor for each process parameter that possesses local variability.

To calculate the scaling factor, we need to find an exact expression for the expected value (mean) of I_{sub}. Since I_{sub} is a function of two independent random variables, (V_l, L_l), we write a double integral to calculate the mean:

$$E[I_{sub}] = \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} g(L_l)\cdot PDF(L_l)\cdot dL_l\right]\bullet g(V_l)\cdot PDF(V_l)\cdot dV_l$$

$$g(L_l) = I_{sub,nom}\cdot e^{-\left[\frac{L_g+\lambda_2 L_g^2}{\lambda_1}\right]}\bullet e^{-\left[\frac{L_l+\lambda_2 L_l^2}{\lambda_1}\right]} \qquad \text{(EQ 8)}$$

$$g(V_l) = e^{-\left[\frac{\lambda_3 V_g^2}{\lambda_1}\right]}\bullet e^{-\left[\frac{\lambda_3 V_l^2}{\lambda_1}\right]}$$

In this equation, the terms containing L_g and V_g are constant for a given chip. Although the above integrals look quite complicated, they can be solved in closed-form and result in the expressions given in EQ21 and EQ23 of the Appendix. We obtain $E[I_{sub}] = \mu_{Isub} = S_L.S_V.I_{Lg,Vg}$ where S_L, S_V are the scale factors introduced due to local variability in L and V. $I_{Lg,Vg}$ corresponds to the subthreshold leakage as a function of global variations.

$$\mu_{Isub} = E[I_{sub}] = S_L\cdot S_V\cdot I_{Lg,Vg}$$

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

$$S_V = e^{\left(\lambda_3^2\sigma_{Vl}^2/2\lambda_1^2\right)} \qquad \text{(EQ 9)}$$

$$I_{Lg,Vg} = I_{sub,nom}\cdot e^{-\left[\frac{L_g+\lambda_2 L_g^2}{\lambda_1}\right]}\cdot e^{-\left[\frac{\lambda_3 V_g}{\lambda_1}\right]}$$

EQ9 provides the average value of subthreshold leakage for a unit width device. To compute the total chip subthreshold leakage, we need to perform a weighted sum of the leakages of all devices by considering the device widths to be the weights. For complex gates (transistor stacks, registers), a scale factor (k) model [6] is used to predict the effect of the total device width.

$$I_{c,sub} = S_L\cdot S_V\cdot\left[\sum_d (W_d/k)\cdot I_{Lg,Vg}\right] \qquad \text{(EQ 10)}$$

Here the term $\sum_d (W_d/k)\cdot I_{Lg,Vg}$ represents the chip level subthreshold leakage as a function of the global process parameters (L_g, V_g).

2.2 Gate Leakage

When the oxide thickness of a device is reduced there is an increase in the amount of carriers that can tunnel through the gate oxide. This phenomenon leads to the presence of gate leakage current (I_{gate}) between the gate and substrate as well as the gate and channel. I_{gate} is linearly dependent on the area of the device and has a super-exponential relationship with the oxide thickness (T_{ox}). Since the variation in T_{ox} has by far the greatest impact on gate leakage, we model I_{gate} as:

$$I_{gate} = I_{nominal}\cdot e^{[f(\Delta T_{ox})]} \qquad \text{(EQ 11)}$$

From circuit simulations, we found that it is sufficient to express $f(\Delta T_{ox})$ as a simple linear function. A suitable value for a single parameter β_1 efficiently captures the highly exponential relationship. Let $\Delta T_{ox}=T$. Using EQ3 we decompose T into global (T_g) and local (T_a) components.

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

$$I_{gate} = I_{gate,nom}\cdot e^{-(T_g/\beta_1)}\cdot e^{-(T_l/\beta_1)} \qquad \text{(EQ 12)}$$

I_{gate} is the gate leakage current of a single device with unit width. $I_{gate,nom}$ is the nominal gate leakage in the absence of any T_{ox} variability. Both T_g and T_l are zero mean random variables. Comparing this equation with EQ6 and EQ7, we see that the relationship between I_{gate} and T_l is similar to the single exponential relationship between I_{sub} and V_l. Similar to S_V, we compute the scale factor due to T_a (S_T) and μ_{Igate}:

$$E[I_{gate}] = \mu_{Igate} = S_T\cdot I_{Tg}$$

$$S_T = e^{\left(\sigma_{Tl}^2/2\beta_1^2\right)} \qquad \text{(EQ 13)}$$

$$I_{Tg} = I_{gate,nom}\cdot e^{-(T_g/\beta_1)}$$

Based on the widths of the devices, the chip level gate leakage can be calculated in a similar manner as the subthreshold leakage:

$$I_{c,gate} = S_T\cdot\left[\sum_d (W_d/k)\cdot I_{Tg}\right] \qquad \text{(EQ 14)}$$

2.3 Total Leakage

The total leakage is the sum of the subthreshold and gate leakage currents of all the devices. In EQ10 and EQ14 we note that $I_{Lg,Vg}$ and I_{Tg} are shared by all the devices on the chip. Hence we can write the equation for total chip leakage as:

$$I_{c,tot} = \left[\sum_d (W_d/k)\right]\left[S_L\cdot S_V\cdot I_{Lg,Vg} + S_T\cdot I_{Tg}\right] \qquad \text{(EQ 15)}$$

This equation can be used to calculate the total leakage for different types of devices such as NMOS/PMOS and low/high-V_{th} transistors. The differences will be in the fitting parameters and the scale factor k. The sum total over all devices gives the total leakage of the chip.

3 Yield Analysis

Traditional parametric yield analysis of high-performance integrated circuits is done using the frequency (or speed) binning method [12]. 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 was illustrated in Figure 1, due to the inverse correlation between leakage and circuit delay chips in the "fast" corner produce vast amounts of leakage current compared to the other chips. In current technologies this is a major concern since a significant number of these chips leak more than the acceptable value and must be discarded. Thus, parametric yield loss is exacerbated since dies are now being lost at both

the low and high speed bins, further narrowing the acceptable process window.

In this section we describe a method to calculate the yield of a lot 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 value of L_g. We then compute the yield due to the imposed leakage limit on a bin-by-bin basis.

3.1 Frequency Dependence on Process Parameters

In principle IC or processor frequency depends on all major process parameters, such as gate length, doping concentration, and oxide thickness. However, we demonstrate from SPICE simulations that circuit delay is primarily impacted by gate length variations. For this purpose, we simulated a 17-stage ring oscillator for different process conditions using the BPTM 100nm process technology as shown in Figure 2. We assume that only global variations in the parameters are likely to influence the delay. We support this assumption by pointing out that ring oscillators are very small structures and thus we can safely ignore the effects of local or within-die variations. From this plot, we see that variations in L_g significantly influence (about $\pm$15%) the delay of the ring oscillator. Variations in T_g and V_g have little or no impact on the delay and thus can be ignored. This is consistent with current practices where a one-to-one correspondence is often assumed between frequency bins and specific gate length values.

3.2 Yield Estimate Computation

We now discuss the method to compute the expected yield for a particular frequency bin based on an imposed leakage limit. For a particular bin, the value of L_g is available and using the expressions for $I_{Lg,Vg}$ (EQ9) and I_{Tg} (EQ13) we rewrite the equation for total chip leakage EQ15 as follows:

$$I_{tot} = A_s \cdot e^{(V_g/k_v)} + A_g \cdot e^{(T_g/k_t)}$$

$$A_s = \left(\sum_d (W_d/k)\right) \cdot S_L \cdot S_V \cdot I_{sub,nom} \cdot e^{-\left[\left(L_g + \lambda_2 L_g^2\right)/\lambda_1\right]} \qquad \text{(EQ 16)}$$

$$A_g = \left(\sum_d (W_d/k)\right) \cdot S_T \cdot I_{gate,nom}$$

$$k_v = -(\lambda_1/\lambda_3) \qquad k_t = -\beta_1$$

Here we simplified the notation for the fitting parameters and expressed this equation in terms of the new constants k_v and k_t. We

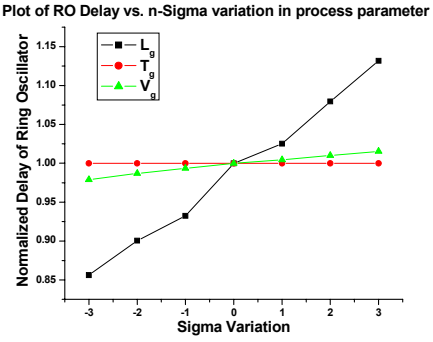

Figure 2. Comparison of the relative contribution of parameter variations on ring oscillator delay

again note that V_g and T_g are both zero mean random variables. The values for k_v and k_t are generally expressed in terms of σ_{Vg} and σ_{Tg}. A_s represents the total chip subthreshold leakage at a value of L_g and includes the scale factors due to the local variability. Similarly A_g represents the total chip gate leakage at a given value of L_g. However, since $I_{c,gate}$ is independent of L_g, A_g is not influenced by changes in the value of L_g. In a plot of total leakage vs. L_g, we first compute A_s and A_g at particular values of L_g and then calculate the distribution of I_{tot} at each of these points.

For every device type, I_{tot} is the sum of two lognormal variables each of which represents a type of leakage current. By our formulation there is no parameter that affects both these terms simultaneously. Thus, we can consider these terms as independent random variables. We model the sum of this pair of lognormals as another lognormal random variable. Using the independence condition, we set the sums of the means and variances to be equal to the mean and variance of the new lognormal. From EQ21 in the Appendix we get:

$$I_{tot} = X_1 + X_2$$

$$X_1 \sim LN(\log(A_s), (\sigma_{Vg}/k_v)^2) \qquad X_2 \sim LN(\log(A_g), (\sigma_{Tg}/k_t)^2)$$

$$\mu_{Itot} = \exp\left[\log(A_s) + \frac{1}{2}\cdot\left(\frac{\sigma_{Vg}}{k_v}\right)^2\right] + \exp\left[\log(A_g) + \frac{1}{2}\cdot\left(\frac{\sigma_{Tg}}{k_t}\right)^2\right] \qquad \text{(EQ 17)}$$

$$\sigma^2_{Itot} = \exp\left[2\log(A_s) + \left(\frac{\sigma_{Vg}}{k_v}\right)^2\right]\cdot[\exp(\sigma^2_{Vg}/k_v^2) - 1] + \exp\left[2\log(A_g) + \left(\frac{\sigma_{Tg}}{k_t}\right)^2\right]\cdot[\exp(\sigma^2_{Tg}/k_t^2) - 1]$$

EQ22 in the Appendix is then used to obtain the mean and variance ($\mu_{N,Itot}$, $\sigma_{N,Itot}^2$) 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.

$$\mu_{N,Itot} = \frac{1}{2}\cdot\log[\mu^4_{Itot}/(\mu^2_{Itot} + \sigma^2_{Itot})]$$

$$\sigma^2_{N,Itot} = \log[1 + (\sigma^2_{Itot}/\mu^2_{Itot})] \qquad \text{(EQ 18)}$$

$$PDF(I_{tot}) = \frac{1}{I_{tot}\cdot\sqrt{2\pi\sigma^2_{N,Itot}}}\cdot\exp\left[-\left(\frac{\log(I_{total}) - \mu_{N,Itot}}{\sqrt{2}\sigma_{N,Itot}}\right)^2\right]$$

Finally, to obtain exact yield estimates we require the quantile numbers for the lognormal distribution described by I_{tot}, i.e., the confidence points of I_{tot} that correspond to the specified leakage limit. Since the exponential function that relates LN(μ_{Itot}, σ_{Itot}^2) with N($\mu_{N,Itot}$, $\sigma_{N,Itot}^2$) is a monotone increasing function, the quantiles of the normal random variable are mapped directly to the quantiles of the lognormal random variable. For instance, the 95th-percentile point on the lognormal distribution will be the exponential of the 95th-percentile point on the distribution of the corresponding normal variable. Using this fact, we can write the expression for the CDF of a lognormal variable:

$$CDF(I_{total}) = F_x(I_{total}) = \frac{1}{2}\cdot\left[1 + erf\left(\frac{\log(I_{total}) - \mu_{N,Itot}}{\sqrt{2}\cdot\sigma_{N,Itot}}\right)\right] \qquad \text{(EQ 19)}$$

Here *erf(.)* is the error function. By setting $F_x(.)$ to a particular confidence point on the normal distribution, we can obtain the corresponding value on the lognormal distribution (see Table 1). In Table 1 the 0-sigma point corresponds to the median of the distribution

Table 1. Value of I_{tot} for an n-sigma point

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

Conversely, if we are given a limit for I_{tot}, we can use EQ19 to compute CDF(I_{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_{tot})*100]% is the fraction of chips that meet both the speed and power criteria. Hence, by repeating this computation for each frequency bin that meets the frequency specification, the total percentage of chips that meet both the leakage and performance constraints can be found.

4 Results

In this section we use our analytical method from the previous section to predict the yield of a lot. Our circuit of choice is a fairly large 64-bit adder written for the Alpha architecture. We assume that all dies in the lot consist of this circuit and a small ring oscillator circuit used to characterize the frequency of the chip with the variation in L_g. We use the 100nm (L_{eff}=60nm) Berkeley Predictive Technology model [4] for our SPICE Monte Carlo simulations. We also employ a gate leakage model based on the BSIM4 equations [11]. The variability numbers for ΔL_{eff}, $\Delta V_{th,Nsub}$ and ΔT_{ox} are based on estimates obtained from an industrial 90nm process.

Figure 3 gives the scatter plot of the total circuit leakage. The y-axis in the plot has been normalized to the sample mean of the leakage currents. We see that for a $\pm 3\sigma$ variation in L_g there is a 7X spread in the leakage. Additionally, for a given L_g, there is a wide distribution in the total circuit leakage. For instance, given $L_g = 0\sigma$, the normalized value of total circuit leakage is between 0.5 and 1.7. This large distribution is due to the variation in V_g and T_g. In EQ16, we see that if V_g=0 and T_g=0 the total leakage is just $I_{tot} = A_s + A_g$, the sum of the subthreshold and gate leakage. However, for small values of (V/k_v) and (T/k_t), the exponential terms increase rapidly and contribute a larger portion to the total leakage value. For instance, if $(V/k_v) = (T/k_t) = 1$, $I_{tot} = (A_s.e + A_g.e) \sim 2.7(A_s + A_g)$, a large increase compared to the nominal value.

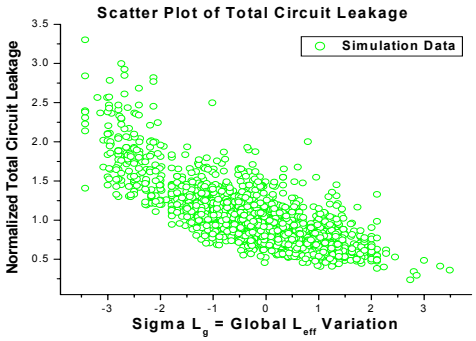

Figure 3. Scatter plot showing the distribution of the total circuit leakage

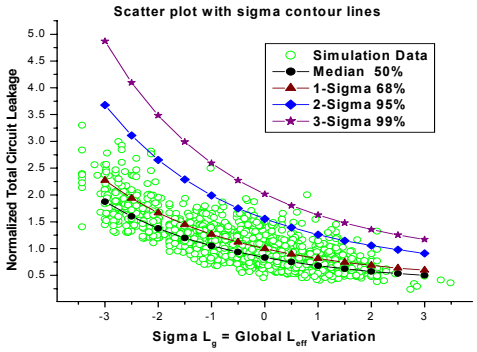

Figure 4. Scatter plot of total circuit leakage with the sigma contour lines added

As a result, the distribution in V_g and T_g (for each value of L_g), produces a band-like curve for the scatter plot of total circuit leakage instead of a single curve. This is very significant since for a given value of L_g (and hence a given operating frequency), a large portion of the chips may be over 3X the nominal leakage value. A chip that operates at an acceptable frequency may still have to be discarded because the variability in V_g, T_g pushes its leakage consumption over the tolerable limit. Thus, we see that the secondary variations V_g, T_g play a major role in determining the yield of a lot.

In Figure 4 we superimpose the sigma contour lines on top of the same leakage scatter plot. For each value of L_g, we calculate ($\mu_{N,Itot}$, $\sigma_{N,Itot}$) and then use Table 1 to construct the contour lines. From the plot we see that there are a fair number of samples "outside" the 2σ range. This is especially true for gate lengths close to the nominal. For shorter channel lengths, since the contour value is quite large there are only a small number of chips outside this range. For larger channel lengths, since the absolute value of the leakage is quite small, there are practically no chips outside the 2σ range.

We now present an example calculation for the yield. For the lot presented here, we impose a frequency limit of $+1\sigma$ and a normalized power limit (P_{lim}) of 1.75. This is indicated in Figure 5. Further, the frequency bins are specified to be at the L_g n-sigma boundaries. First, we see that due to the performance (frequency) limit all chips that operate at frequencies smaller than the $+1\sigma$ value are discarded. As we can see from the plot, although all of these chips meet the power criteria they are discarded since they are "too slow". Next, we proceed bin-by-bin and calculate the yield for each bin. To illustrate the yield computation, we present the numbers for

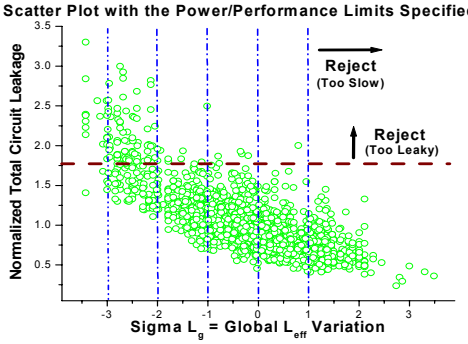

Figure 5. Scatter plot with power and performance limits specified

Table 2. CDF[(I_{tot})*100]% numbers for three different values of P_{lim} for the our range of L_g values.

P_{lim}	L_g *n*-sigma				
	-3	-2	-1	0	1
1.00	6.4	21.1	44.8	68.5	84.8
1.75	43.6	72.6	90.5	97.5	99.4
2.50	76.0	93.3	98.7	99.8	99.9

only the cases when L_g = [-3σ,-2σ,...,+1σ]. For each such L_g, we calculate A_s, A_g in EQ16. We then use these values to calculate ($\mu_{N,Itot}$, $\sigma_{N,Itot}^2$) from EQ17 and EQ18. From the power limiter number, we calculate the unscaled value of I_{tot} and use it in EQ19 to obtain CDF values. [CDF(I_{tot})*100]% is the number of good chips that satisfy both the power and performance criteria. Table 2 summarizes these CDF numbers for three different values of P_{lim}.

Traditional parametric yield analysis does not consider power as a criterion and hence overestimates the number of chips that are actually good/sellable. For instance, if P_{lim}=1.75, we see from Table 2 that for L_g=-2σ 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=0σ, about 2.5% of the chips are lost since they lie outside the power limit. While a typical frequency binning method would predict that 100% of the chips with L_g=-2σ are good, our method captures the fact that over 25% cannot be marketed. This is particularly important since fast bin devices are considered to be highly profitable. Hence, there is a need to adopt an integrated approach that accounts for the compounded loss due to both limiting factors. We find that our approach always predicts a lower yield percentage compared to the method that assumes independence of the limiting factors of power and performance. By preserving the correlation between frequency and leakage we are able to obtain more accurate estimates for the yield.

5 Conclusions

In this paper we presented an analytical framework that provides a closed-form expression for the total chip leakage current as a function of relevant process parameters. Using this expression we estimate the yield of a lot when both power and performance constraints are imposed. We presented an example calculation for yield that shows the compounded loss that occurs due to chips that operate at low frequencies as well as chips that produce excessive amounts of leakage. Our method exemplifies the need to consider both limiters when calculating the yield of a lot.

6 Appendix

Given a Normal (Gaussian) random variable $X \sim N(\mu_x,\sigma_x^2)$, the PDF of X is given by [13]:

$$PDF(x) = f_X(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \cdot \exp\left[-\left(\frac{x-\mu_x}{\sqrt{2}\sigma_x}\right)^2\right] \quad \text{(EQ 20)}$$

The function $Y = g(X) = e^{-(X/a_1)}$ is a Lognormal random variable. Using the values for (μ_x,σ_x^2) we can express the mean and variance of Y in closed form.

$$\mu_y = e^{\left[-(\mu_x/a_1)+\left(\sigma_x^2/2a_1^2\right)\right]}$$
$$\sigma_y^2 = e^{\left[-(2\mu_x/a_1)+\left(\sigma_x^2/a_1^2\right)\right]} \cdot \left[e^{\left(\sigma_x^2/a_1^2\right)} - 1\right] \quad \text{(EQ 21)}$$

Conversely, given the values for the mean and variance of the Lognormal random variable (μ_y,σ_y^2), we can compute the mean and variance of the corresponding Normal random variable to obtain (μ_x,σ_x^2). (We have normalized Y by setting a_1=-1).

$$\mu_x = \frac{1}{2} \cdot \log[\mu_y^4/(\mu_y^2+\sigma_y^2)]$$
$$\sigma_x^2 = \log[1+(\sigma_y^2/\mu_y^2)] \quad \text{(EQ 22)}$$

For the random variable $Z = h(X) = e^{-\left[\left(X+a_2X^2\right)/a_1\right]}$ where X is a zero mean normal random variable, it is possible to obtain a closed-form expression for the mean and variance.

$$E[Z] = \mu_z = \left[1/\left(\sqrt{1+\frac{2a_2}{a_1}\sigma_x^2}\right)\right] \cdot e^{\left[\sigma_x^2/\left(2a_1^2+4\sigma_x^2a_1a_2\right)\right]}$$
$$E[Z^2] = \left[1/\left(\sqrt{1+\frac{4a_2}{a_1}\sigma_x^2}\right)\right] \cdot e^{\left[\sigma_x^2/\left(a_1^2/2+2\sigma_x^2a_1a_2\right)\right]} \quad \text{(EQ 23)}$$
$$\sigma_z^2 = E[Z^2] - \mu_z^2$$

7 Acknowledgements

We would like to thank Vivek De from Intel for providing us with Figure 1. This work was supported in part by NSF, SRC, GSRC/ DARPA, IBM, and Intel.

8 References

[1] S. Narendra, D. Blaauw, A. Devgan and F. Najm, "Leakage issues in IC design: Trends, estimation and avoidance", Tutorial, *ICCAD* 2003.
[2] S. Borkar, "Design challenges of technology scaling", *IEEE Micro*, July 1999.
[3] S. Mukhopadhyay, K. Roy, "Modeling and estimation of total leakage current in nano-scaled CMOS devices considering the effect of parameter variation", *ISLPED* 2003.
[4] http://www-device.eecs.berkeley.edu/~ptm/
[5] S. Borkar, et.al, "Parameter variations and impact on circuits and microarchitecture", *DAC* 2003.
[6] S. Narendra, et.al, "Full-chip subthreshold leakage power prediction model for sub-0.18um CMOS", *ISLPED* 2002.
[7] S. Mukhopadhyay, A. Raychowdhury, K. Roy, "Accurate estimate of total leakage current in scaled CMOS circuits based on compact current modeling", *DAC* 2003.
[8] R. Rao, A. Srivastava, D. Blaauw, D. Sylvester, "Statistical estimation of leakage current considering inter- and intra-die process variation", *ISLPED* 2003.
[9] A. Srivastava, R. Bai, D. Blaauw, D. Sylvester, "Modeling and analysis of leakage power considering within-die process variations", *ISLPED* 2002.
[10] S. Mukhopadhyay, K. Roy, "Modeling and estimation of total leakage current in nano-scaled CMOS devices considering the effect of parameter variation", *ISLPED* 2003.
[11] http://www-device.eecs.berkeley.edu/~bsim3/bsim4.html
[12] B. Cory, R. Kapur, B. Underwood, "Speed binning with path delay test in 150-nm technology", *IEEE Design and Test of Computers*, October 2003.
[13] A. Papoulis, Probability, Random Variables and Stochastic Processes, McGraw-Hill Inc., New York 1991.