Full-Chip Analysis of Leakage Power Under Process Variations, Including Spatial Correlations*

Hongliang Chang
Dept. of Computer Science and Engineering
University of Minnesota
hchang@cs.umn.edu

Sachin S. Sapatnekar
Dept. of Electrical and Computer Engineering
University of Minnesota
sachin@ece.umn.edu

ABSTRACT

In this paper, we present a method for analyzing the leakage current, and hence the leakage power, of a circuit under process parameter variations that can include spatial correlations due to intra-chip variation. A lognormal distribution is used to approximate the leakage current of each gate and the total chip leakage is determined by summing up the lognormals. In this work, Both subthreshold leakage and gate tunneling leakage are considered. The proposed method is shown to be effective in predicting the CDF/PDF of the total chip leakage. The average errors for mean and sigma values are -1.3% and -4.1%.

Categories and Subject Descriptors

B.8.2 [**Hardware**]: Performance and Reliability—*Performance Analysis and Design Aids*

General Terms

Algorithm, Design, Performance, Reliability

1. INTRODUCTION

Leakage power is increasing drastically with technology scaling, and has already become a substantial contributor to the total chip power dissipation. According to International Technology Roadmap for Semiconductors (ITRS) [16], leakage power is expected to increase to 50% of the total chip power and to dominate the switching power of a circuit over the next few generations. Consequently, it is important to accurately estimate leakage currents so that they can be accounted for during design, and so that it is possible to effectively optimize the total power consumption of a chip.

The major components of leakage in current CMOS technologies are due to subthreshold leakage and gate tunneling leakage. For a gate oxide thickness, T_{ox}, of over 20Å, the gate tunneling leakage current, I_{gate}, is typically very small [5], while the subthreshold leakage, I_{sub}, dominates other types of leakage in circuit. For this reason, there have been extensive studies on subthreshold leakage over the last ten years [4, 11]. However, the gate tunneling leakage is exponentially dependent on gate oxide thickness, e.g., a reduction in T_{ox} of 2Å will result in an order of magnitude increase in I_{gate}. Therefore, with the continuous scaling of gate oxide thickness, I_{gate} is no longer negligible and is likely to dominate other leakage mechanisms in future generations, at least until new high-K dielectrics are introduced. At this time, the earliest estimates of when these will be introduced is around 2007, and gate leakage is already seen to be very significant in 90nm technologies [16], so that analysis of gate leakage is of profound importance today.

In the literature, several research works on the analysis and minimization of total circuit leakage including the effect of I_{gate} have been conducted [5]. The analysis of total leakage power of circuit is complicated by the state dependency of subthreshold and gate tunneling leakage, and the interactions between these two leakage mechanisms.

An added complication, which has been less widely studied, arises due to the increasing importance of process variations in cutting-edge technologies. As a result of this the values of all process parameters can no longer be considered to be constants, but must be modeled as random variables that are described by probability density functions (PDFs). These variations translate into uncertainties in circuit performance metrics. Specifically, total circuit leakage also becomes a random variable that depends on the variations of fundamental process parameters that it is most sensitive to parameters such as the transistor effective gate length and the gate oxide thickness.

In general, process variations can be classified into the following categories: *inter-die variations* are the variations from die to die, while *intra-die variations* correspond to variability within a single chip. Inter-die variations affect all the devices on same chip in the same way, e.g., making the transistor gate lengths of devices on the same chip all larger or all smaller, while the intra-die variations may affect different devices differently on the same chip, e.g., making some devices have smaller transistor gate lengths and others larger transistor gate lengths. In addition, intra-die variations exhibit spatial correlation, i.e., devices located close to each other are more likely to have the similar characteristics than those placed far away. Mathematically, inter-die variations can be regarded as a special case of intra-die variations with a correlation value of one.

Under inter-die variations, if the leakage of all gates or devices are sensitive to the process parameters in similar ways, the circuit performance can be analyzed at multiple process corners using deterministic analysis methods. Otherwise, or with intra-die variations, statistical methods must be used to correctly predict the leakage. Specifically, the gate leakage can vary exponentially with these parameters, the simple use of worst-case values for all parameters can result in exponentially larger leakage estimates than

*This work was supported in part by the NSF under award CCR-0205227 and by the SRC under contract 2003-TJ-1092.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
DAC 2005, June 13–17, 2005, Anaheim, California, USA.
Copyright 2005 ACM 1-59593-058-2/05/0006 ...$5.00.

are actually obtained. While these will certainly be pessimistic, the inaccuracy in these values makes them practically useless.

Most of the previous works on statistical performance analysis has focused on statistical timing analysis, and only a few works have investigated the variation of leakage power under the effect of process variations [6,7,9,10,12]. In [7,12], analytical methods were proposed to estimate the mean and standard deviation of the total chip subthreshold leakage power under intra-die parameter variations. In [6], gate tunneling and the reverse biased source/drain junction band-to-band tunneling (BTBT) leakage, and the correlations among these components were included, in addition to subthreshold leakage, in the analysis of total leakage. In [10], the probability density function (PDF) of the total chip subthreshold leakage was derived. The authors of [9] presented an analytical framework that provides a closed form expression for the total chip leakage current as a function of process parameters that can be used to estimate yield under power and performance constraints. However, none of these have considered the effects of spatial correlations in intra-die process variations.

In this paper, we propose a method for predicting the distribution of total circuit leakage power, including subthreshold and gate tunneling leakage and their interactions, under both inter-die and intra-die variations of parameters. The spatial correlations in intra-die variations and the correlation between these two leakage mechanisms are also considered.

2. MODELING PARAMETER VARIATIONS

In general, a parameter variation can be modeled as

$$\delta_{total} \quad = \quad \delta_{inter} + \delta_{intra}, \tag{1}$$

where δ_{inter} is the inter-chip variation and δ_{intra} is the intra-chip variation. In this work, δ_{inter} and δ_{intra} are both modeled as Gaussian random variables. Due to global effect of inter-die variations, a single random variable δ_{inter} is used for all transistors [wires] in a chip to model the inter-die variation.

For intra-die variation δ_{intra}, we use the same model as in the work of [3], in which, under intra-die variation, the value of a parameter p located at (x, y) can be modeled as:

$$p \quad = \quad \bar{p} + \delta_x x + \delta_y y + \epsilon \tag{2}$$

where $\bar{p}$ is the nominal design parameter value at die location $(0, 0)$, and δ_x and δ_y are gradients of parameter indicating the spatial variations of parameter along the x and y directions respectively. The term, ϵ, stands for the random intra-chip variation, and the vector of all random components across the chip has a correlated multivariate normal distribution due to spatial correlations in the intra-chip variation $\vec{\epsilon} \sim N(0, \Sigma)$, where Σ is the covariance matrix of the spatially correlated parameters, as described in the remainder of this section.

In [3], the intra-die spatial correlations of parameters are modeled by partitioning the die region into $nrow \times ncol = n$ grids. Since devices close to each other are more likely to have more similar characteristics than those placed far away, we assume perfect correlations among the devices in the same grid, high correlations among those in close grids and low or zero correlations in far-away grids. For example, in Figure 1: gates a and b (whose sizes are shown to be exaggeratedly large) are located in the same grid square, and it is assumed that their parameter variations (such as the variations of their gate length), are always identical. Gate a and c lie in neighboring grids, and their parameter variations are not identical but highly correlated due to their spatial proximity (for example, when gate a has a larger than nominal gate length, it is highly probable that gate c will have a larger than nominal gate length, and less probable that it will have a smaller than nominal gate length). On the other hand, gates a and d are far away from each other, their parameters are uncorrelated (e.g., when gate a has a larger than nominal gate length, the gate length for d may be either larger or smaller than nominal).

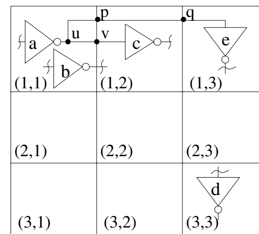

Figure 1: Grid model for spatial correlations

With this model, a parameter variation in a single grid at location (x, y) can be modeled using a single random variable $p(x, y)$. For each type of parameter, n random variables are needed, each representing the value of a parameter in one of the n grids. In addition, it is assumed that correlation exists only among the same type of parameters in different grids and there is no correlation between different types of parameters (however, this assumption is not critical to our framework and can easily be removed). For example, transistor gate length for transistors in a grid are correlated with those in nearby grids, but are uncorrelated with other parameters such as gate oxide thickness in any grid. For each type of parameter, a correlation matrix Σ of size $n \times n$ represents the spatial correlations of such a structure. Note that the number of grid partitions needed is determined by the process, but not the circuit. In other words, the same correlation model can be applied to different designs under the same process.

In this work, we consider the variations in the transistor gate length and gate oxide thickness[1], since I_{sub} and I_{gate} are most sensitive to these parameters [6, 14]. To reflect reality, we model spatial correlations in transistor gate length, while the gate oxide thickness values for different gates are taken to be uncorrelated.

3. MODEL FOR FULL-CHIP LEAKAGE CURRENT ANALYSIS

In this section, we will describe the empirical models for subthreshold and gate leakage currents, based on which the leakage current under process variations is estimated under lognormal distributions. The model for computing the full-chip leakage current is presented at the end of this section, where state dependencies of subthreshold and gate leakage currents and their interactions are considered. We will loosely use the terms "leakage current" and "leakage power" interchangeably, since the two terms are related by a multiplicative factor of V_{dd}.

3.1 Static Leakage Models

3.1.1 Subthreshold Leakage Model

The subthreshold leakage current, I_{sub}, is exponentially dependent on the threshold voltage, V_{th}, and V_{th} is observed to be most sensitive to gate oxide thickness T_{ox} and effective gate length L due to short-channel effects [14]. The precise relationship shows an exponential dependency, due to which a small change in L or T_{ox} will have a substantial effect on I_{sub}, with the effect of T_{ox} relatively weak. From this intuition, as in [6,9,10], we estimate the subthreshold leakage current by developing a empirical curve-fitted model:

[1] Although only transistor gate length and gate oxide thickness are considered in this work, the framework is general enough to consider effects of any other types of process variations such as channel dopant variation, etc.

$$I_{sub} = a_0 w e^{a_1 + a_2 L + a_3 L^2 + a_4 T_{ox}^{-1} + a_5 T_{ox}} \quad (3)$$

where w is the gate width of the transistor, and a_0 through a_5 are the fitting parameters. Observe that I_{sub} is proportional to w; we will later use a look-up table to store value of I_{sub} per unit width.

Subthreshold leakage current has a well-known state dependency due to the stack effect [11]. A simple look-up table (LUT) can be used to include this effect, and for a k-input gate, the size of the LUT is 2^k. In addition, by keeping only dominant states for I_{sub} [4,11], i.e., only one "off" transistor in a series transistor stack, the size of the table can be greatly reduced while maintaining a reasonable accuracy. For example, for a k-input NAND gate, the size of the LUT comes down from 2^k to only k entries.

In a circuit, the average subthreshold leakage of a gate can be computed as a weighted sum of I_{sub} under dominant input states,

$$I_{sub}^{avg} = \sum_{i \in dominant\ input\ states} P_i I_{sub_i} \quad (4)$$

where P_i is the probability of input state i and I_{sub_i} is the subthreshold leakage value at input state i.

3.1.2 Gate Tunneling Leakage Model

In [2], an analytical model was proposed for the gate oxide tunneling current I_{gate}. However, the formulation does not lend itself easily to the analysis of the effects of parameter variations. In this work, we again use a empirical model to estimate I_{gate} through curve fitting similarly to [6, 10]:

$$I_{gate} = w a_0 e^{a_1 + a_2 L + a_3 L^2 + a_4 T_{ox} + a_5 T_{ox}^2} \quad (5)$$

As in [2, 13], in this work, the gate-tunneling current of the PMOS device is neglected due to the larger effective mass and barrier height for holes compared to electrons at the $\mathrm{SiO_2/Si}$ interface. Only tunneling current in the gate-to-channel region is considered, and edge direct tunneling (EDT) in the gate-to-drain and gate-to-source overlap regions is ignored. This is because these overlap regions are significantly smaller than the gate-to channel region; moreover, EDT can be further reduced using process technologies. Therefore, the gate tunneling current is only computed when NMOS is at logic "1".

It was shown in [5] that interactions between I_{sub} and I_{gate} may exist, depending on the input vector state at the gate. At some states, their interactions can make either I_{sub} or I_{gate} diminished to a small value that can be ignored safely. The dominant states of I_{gate} can be determined from the analysis of [5]: in a transistor stack, the dominant states of a particular leaking transistor is when all transistors on the path from the source of this transistor to the source of the whole stack are "on" [13]. For details, the reader is referred to [5].

The average gate tunneling leakage of a gate is computed as a weighted sum of the gate tunneling leakage corresponding to the dominant states:

$$I_{gate}^{avg} = \sum_{i \in dominant\ input\ states} P_i I_{gate_i} \quad (6)$$

where P_i is the probability of input state i and I_{gate_i} is the gate tunneling leakage value at input state i.

3.2 Distribution of Leakage Current

In the previous sections, I_{sub} and I_{gate} are both modeled as exponential functions αe^Y, where Y is a function of process parameters L and T_{ox}. Under process parameter variations, Y is a random variable. Since the parameter variations are in general around $10 - 20\%$ [8], using a first-order Taylor expansion at the nominal values of process parameters as in [3], Y can be approximated by a normal distribution. For instance, by expanding the exponent of I_{sub} as expressed in (3), we have,

$$I_{sub} = \alpha e^{Y^0 + \beta_0 \cdot dL + \beta_1 \cdot dT_{ox}} \quad (7)$$

where Y^0 is the nominal value of the exponent, β_0 and β_1 are the derivatives of the exponent to L and T_{ox} evaluated at their nominal values respectively.

Thus, αe^Y has a lognormal distribution, i.e., I_{sub} and I_{gate} can both be approximated by lognormal distributions.

3.3 Distribution of Full-Chip Leakage Current

The full-chip average leakage can be now computed by summing up leakage current of each gate in the circuit:

$$I_{total}^{avg} = \sum_{\forall\ gates\ i=1,\cdots,m} I_{sub_i}^{avg} + I_{gate_i}^{avg} \quad (8)$$

where m is the total number of gates in the circuit. $I_{sub_i}^{avg}$ and $I_{gate_i}^{avg}$ are subthreshold and gate tunneling leakage current computed using dominant states only. Note that since we use dominant states for both I_{sub} and I_{gate}, we ensure that the interaction between these two leakage mechanisms is included in the total leakage current estimation.

As each leakage component is approximated as a lognormal distribution, the full-chip leakage distribution can simply be found by summing up the distribution of the lognormals for all gates. Since spatial correlations are considered, the leakage distributions between any two gates may be correlated. Therefore, the computation of full-chip leakage current involves finding a sum of correlated lognormals:

$$S = \sum_{i=1}^{p} e^{Y_i} \quad (9)$$

where p is the total number of lognormals to sum, and Y_i is a Gaussian random variables with mean m_{y_i} and variance σ_{y_i}, and the vector of all Y_i's forms a multivariate normal distribution with covariance matrix Σ_Y.

4. COMPUTATION OF FULL-CHIP LEAKAGE CURRENT DISTRIBUTION

In this section, we will present our algorithm for computing the sum of correlated leakage components, so that the spatial correlations between parameters, and correlations between different leakage components can be taken into account in the sum. At this point, we consider only intra-die variations of parameters. The extension to handling inter-die variations is quite obvious, and will be described briefly in Section 5.

4.1 Sum of Correlated Lognormals

Theoretically, the sum of lognormal distribution is not known to have a closed form. However, it may be well approximated again as a lognormal using Wilkinson's method [1]. A sum of t lognormals, $S = \sum_{i=1}^{t} e^{Y_i}$, is approximated as the lognormal e^Z, where $Z = N(m_z, \sigma_z)$. In Wilkinson's approach, the mean and standard deviation of Z are obtained by matching the first two moments, u_1 and u_2 of $\sum_{i=1}^{t} e^{Y_i}$ as follows:

$$u_1 = E(S) = e^{m_z + \sigma_z^2/2} = \sum_{i=1}^{t} e^{m_{y_i} + \sigma_{y_i}^2/2} \quad (10)$$

$$u_2 = E(S^2) = e^{2m_z + 2\sigma_z^2} = \sum_{i=1}^{t} e^{2m_{y_i} + 2\sigma_{y_i}^2} + 2\sum_{i=1}^{t-1} \sum_{j=i+1}^{t} e^{m_{y_i} + m_{y_j}} e^{(\sigma_{y_i}^2 + \sigma_{y_j}^2 + 2r_{ij}\sigma_{y_i}\sigma_{y_j})/2}$$

where r_{ij} is the correlation coefficient of Y_i and Y_j. Solving (10) for m_z and σ_z yields:

$$m_z = 2\ln u_1 - \frac{1}{2}\ln u_2 \quad (11)$$

$$\sigma_z^2 = \ln u_2 - 2\ln u_1$$

The above formula shows the need for a pair-by-pair computation for all correlated pairs of variables, i.e., for all i, j such that $r_{ij} \neq 0$. It is easy to see that this could lead to a prohibitive amount of computation. Firstly, due to spatial correlation of L, leakage currents of different gates are correlated. Secondly, the subthreshold leakage and gate leakage associated with the same NMOS transistor are correlated, and thirdly, subthreshold leakage currents in the same transistor stack are also correlated. If there are N_{gate} gates in the circuit, the complexity for computing the sum will be $O(N_{gate}^2)$ which is far from practical for large circuits. We will introduce a mechanism to reduce this complexity in Section 4.2.

4.2 Reducing the Number of Lognormals to be Summed

As described in section 3, only leakage at dominant states of I_{sub} and I_{gate} are considered. Consider the dominant states of subthreshold leakage current in two transistor stacks, each has only one "off" transistor in the stack. It is observed that the values of subthreshold leakage currents *per unit width* are almost the same for any two transistor stacks that have the same number of "on" transistors between the drain of the only "off" transistor and the output of the stack. For example, it is observed that the subthreshold leakage current per unit width is the same for the pulldown of a NAND4 in state 0111, a NAND3 in state 011, a NAND2 in state 01, and an INV in state 0. Therefore, we can create a look-up table that stores the subthreshold leakage current per unit width, for various numbers of "on" transistors above the leaking transistor, and the size of this table is small. If q is the length of the longest stack in the library, then the number of entries in the look-up table is $2q$ for I_{sub} (q each for I_{sub} for the PMOS and the NMOS).

For gate tunneling leakage, the size of the look-up table can be similarly reduced. For a dominant state of I_{gate}, it is observed that the value of I_{gate} for a particular transistor does not depend much on the number of transistors in the stack, since all transistors below it are "on" which make a conducting path from the leaking transistor to the source of the stack. Therefore, only one model is needed to characterize the gate tunneling leakage of a transistor.

In addition, under our spatial correlation model, gates in the same grid have the same parameter values. For example, let I_{sub}^i be the subthreshold leakage currents for gates $i = 1, \cdots, t$, under the same input vector, and assume that these gates are all in the same grid k. Then,

$$I_{sub}^i = \alpha_i e^{Y_i^0 + \beta_0 \cdot dL_k + \beta_1 \cdot dTox_i} \qquad (12)$$

Note that all of the I_{sub}^i's in the same grid use the same variable dL_k, but different $dTox$ values since the gate oxide thickness is uncorrelated from gate to gate.

Then, the sum of the leakage terms I_{sub}^i in grid k is given by:

$$e^{Y_i^0 + \beta_0 \cdot dL_k} \cdot \sum_{i=1}^{t} \alpha_i \cdot e^{\beta_1 \cdot dTox_i} \qquad (13)$$

The second part of the expression above is a sum of independent lognormal variables, which is a special case of sum of correlated lognormal variables, and this can be computed in linear time using Wilkinson's method. Therefore, for gates of the same type with the same input state in the same grid, only a linear time complexity is needed and the sum of leakage of these gates is finally approximated by a lognormal variable that can be superposed in the original expression.

Similarly, the gate tunneling leakages of different gates in the same grid can be summed up in linear time and approximated by a lognormal variable.

At this point, the number of correlated leakage components in each grid is reduced to a small constant c in our library and if the chip is divided into n grids, the total number of correlated lognormals to sum is no more than $c \cdot n$. In general, the number of grids is substantially smaller than the number of gates in the circuit and can be regarded as a constant number. Therefore, we have reduced the complexity required for the sum of lognormals from $O(N_{gate}^2)$ to a substantially smaller constant $O(n^2)$.

4.3 Handling Correlations Between Leakage Mechanisms

In different gates, leakage currents are correlated due to spatially correlated parameters such as transistor gate length. Within the same gate, the subthreshold and gate tunneling leakage currents are correlated, and leakage currents under different input vectors are correlated because they are sensitive to the same parameters of the gate, regardless of whether these are spatially correlated or not. In order to correctly predict the distribution of total leakage in the circuit, the correlations of these leakage currents must be considered when they are summed up.

In Section 4.2, in order to reduce the number of correlated leakage components to sum, the leakage currents that arise from the same leakage mechanisms in the same grid from the same entry of the look-up table are merged into a single lognormally distributed leakage component. Let I_1^{sum} and I_2^{sum} be two merged sums, corresponding to subthreshold leakage and gate leakage components in the same grid, respectively. These are calculated as:

$$I_1^{sum} = e^{Y_1^0 + \beta_0 dL} \left(\sum_{i=1}^{t} \alpha_i e^{\beta_1 dTox_i} \right) = e^{Y_1^0 + \beta_0 dL} e^{\xi} \qquad (14)$$

$$I_2^{sum} = e^{Y_2^0 + \beta_0' dL} \left(\sum_{i=1}^{t'} \alpha_i' e^{\beta_1' dTox_i'} \right) = e^{Y_2^0 + \beta_0' dL} e^{\gamma} \qquad (15)$$

where e^{ξ} and e^{γ} are the lognormals approximating the sum of independent lognormals, $\sum_{i=1}^{t} \alpha_i e^{\beta_1 dTox_i}$ and $\sum_{i=1}^{t'} \alpha_i' e^{\beta_1' dTox_i}$ in I_1^{sum} and I_2^{sum} respectively, as described in Section 4.2.

Note that $\sum_{i=1}^{t} \alpha_i e^{\beta_1 dTox_i}$ and $\sum_{i=1}^{t'} \alpha_i' e^{\beta_1' dTox_i}$ may be correlated, since the same gate may contribute both subthreshold and gate leakage. Therefore, e^{ξ} and e^{γ} are correlated and the correlation between ξ and σ has to be derived.

Since the T_{ox} values are independent in different gates, the correlation, $cov(\sum_{i=1}^{t} \alpha_i e^{\beta_1 dTox_i}, \sum_{i=1}^{t'} \alpha_i' e^{\beta_1' dTox_i})$, is easily computed as:

$$\sum \alpha_i \alpha_i' e^{(\beta_i^2 + \beta_i'^2)\sigma_{Tox_i}^2/2} (e^{\beta_i \beta_i' \sigma_{Tox_i}^2} - 1) \qquad (16)$$

The correlation between e^{ξ} and e^{γ} is then found as:

$$\begin{aligned} cov(e^{\xi}, e^{\gamma}) &= E(e^{\xi+\gamma}) - E(e^{\xi})E(e^{\gamma}) \qquad (17) \\ &= e^{M_\xi + M_\gamma + (\sigma_\xi^2 + \sigma_\gamma^2)/2} (e^{cov(\xi,\gamma)/2} - 1) \end{aligned}$$

where M_ξ $[M_\gamma]$ and M_γ $[\sigma_\gamma]$ are the mean and standard deviation of ξ $[\gamma]$, respectively. Solving Equation (17) for $cov(\xi, \gamma)$, we have:

$$cov(\xi, \gamma) = 2 \log \left(1 + \frac{cov(e^{\xi}, e^{\gamma})}{e^{M_\xi + M_\gamma + (\sigma_\xi^2 + \sigma_\gamma^2)/2}} \right) \qquad (18)$$

Since e^{ξ} and e^{γ} are approximations of $\sum_{i=1}^{t} \alpha_i e^{\beta_1 dTox_i}$ and $\sum_{i=1}^{t'} \alpha_i' e^{\beta_1' dTox_i}$ respectively, we can reasonably assume that

$$cov(e^{\xi}, e^{\gamma}) = cov \left(\sum_{i=1}^{t} \alpha_i e^{\beta_i \Delta Tox_i}, \sum_{i=1}^{t'} \alpha_i' e^{\beta_i' \Delta Tox_i} \right) \qquad (19)$$

Moreover, the mean and standard deviation of ξ and γ are already known from the approximation, and therefore, the computation of $cov(\xi, \gamma)$ is easily possible.

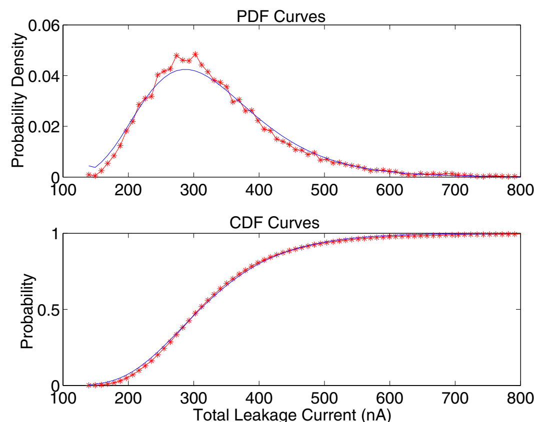

Figure 2: Distributions of the total leakage using the proposed method against Monte Carlo simulation method for circuit c7552. (Our method: solid line, Monte Carlo: starred line)

5. HANDLING INTER-DIE VARIATIONS

The framework for statistical computation of full-chip leakage considering spatial correlations in intra-die variations of parameters can easily be extended to handle inter-die variation. To include the effects of inter-die variations in leakage estimation, for each type of parameter, a global random variable can be applied to all gates in the circuit to model this effect. For spatially correlated parameters, this is reflected as an update of the covariance matrix by adding to all entries the variance of the global random variable, and thus the same framework for leakage estimation proposed in earlier sections can be applied. For spatially uncorrelated parameters, the global random variable only introduces a correlation term between the leakage currents of different gates, and the algorithm for total leakage estimation remains the same.

6. EXPERIMENTAL RESULTS

Our method for full-chip statistical leakage estimation was tested on circuits in the ISCAS85 benchmark set. The circuits were synthesized with SIS with cell library consisting of an inverter, and NAND, NOR, AND, and OR gates with 2, 3 and 4 input pins. The designs were placed using Capo [15]. The technology parameters that were used correspond to the 100nm Berkeley Predictive Technology model, and the 3σ value of parameter variations for L and T_{ox} were set to 20% of the nominal parameter values, of which inter-die variations constitute 40% and intra-die variations, 60%. The spatial correlation was modeled so that the correlation coefficient value diminishes equally with the distance between any two grids, and the numbers of grid partitions of spatial correlation model used for the benchmarks are given in Table 1.

For comparison purposes, we performed Monte Carlo simulations with 10,000 runs on the benchmarks. The results of the comparison are shown in Table 1. The average errors for mean and sigma values are -1.3% and -4.1%, respectively. In Figure 2, we show the distribution of leakage achieved from the proposed method and Monte Carlo simulation for circuit c7552: it is easy to see that the curve achieved by the proposed method matches well with the Monte Carlo simulation result. For all testcases, the runtimes of the proposed method are less than one second, while the Monte Carlo simulation takes considerably longer: for the largest test case, c7552, Monte Carlo simulation takes 3 hours.

To show the importance of considering spatial correlations, we run another set of Monte Carlo simulations (*MCNoCorr*) on the same set of benchmarks, assuming correlation coefficients of zero between the effective gate length L of all transistors on the chip. The comparison between the data is also shown in Table 1. It can be observed that although the mean values are close, on average, the variances of *MCNoCorr*, where spatial correlations are ignored, show an average underestimation of 16.5% compared to *MC*, where the spatial correlations are taken into account. This is because the leakage values of different gates are less correlated when spatial correlations are ignored, and thus different gates have lower probabilities of taking larger values of leakage simultaneously, which results in smaller overall variations.

To visualize the difference, in Figures 3 and 4, for circuit c432, we show the scatter plots for 2000 samples of full-chip leakage generated by Monte Carlo simulations with and without consideration of spatial correlations of L. The x-axis marks the multiples of the standard deviation value of inter-die variations in the effective gate length, ΔL^{inter}, which ranges from -3 to $+3$ since a Gaussian distribution is assumed. For each specific value of L^{inter}, the scatter plot shows the various values for leakage due to variations in T_{ox} and intra-die variations of L. The plots also show a set of contours lines that correspond to different percentage points for the CDF of leakage current, with spatial correlation considered at different values of ΔL^{inter}. In Figure 3, where spatial correlations are considered, nearly all points generated from Monte Carlo simulation fall between the contours of the 1% and 99% lines. However, in Figure 4, where spatial correlations are ignored, the spread is much tighter in general: the average value of 90% point of full-chip leakage, with spatial correlation considered, is 1.5 times larger than that without for $\Delta L^{inter} \leq -1\sigma$; the same ratio is 1.1 times larger otherwise. Looking at the same numbers in a different way, in Figure 4, all points are contained between the 30% and 80% contours if $\Delta L^{inter} \leq -1\sigma$. In this range, I_{sub} is greater than I_{gate} by one order of magnitude on average, and thus the variation of L can have a large effect on the total leakage as I_{sub} is exponentially dependent on L. Consequently, ignoring spatial correlation results in a substantial under-estimation of the standard deviation, and thus the worst-case full-chip leakage. For $\Delta L^{inter} > -1\sigma$, I_{sub} decreases to a value comparable to I_{gate} and L has a relatively weak effect on the variation of total leakage. In this range, the number of points of larger leakage values is similar to that when spatial correlation is considered. However, the large number of remaining points show smaller variations and are within the 20% and 90% contours, due to the same reasoning given above for $\Delta L^{inter} \leq -1\sigma$.

We also study the components of the variation of full-chip leakage, subthreshold and gate-tunneling leakage due to the variations of L and T_{ox} alone. In Table 2, the results with varying L and T_{ox} at the nominal value are provided in columns 2 to 7, and the last 6 columns show the reverse. As seen in the table, the variations of L and T_{ox} can each individually lead to substantial variations in the full-chip leakage. When only L varies, I_{sub} varies substantially (the average ratio of the mean to the standard deviation is 40.2%) and I_{gate} trivially (the corresponding ratio is 5.5%), since I_{sub} is more sensitive to the variation of L than T_{ox}, and I_{gate} is a strong exponential function of T_{ox} over L. In this case, I_{sub} dominates I_{gate} by 4 to 5 times and the variation of full-chip leakage is mainly due to I_{sub}. In contrast, when only T_{ox} varies, the mean of I_{gate} doubles and standard deviation increases by 40 times, while standard deviation of I_{sub} is about 3 times smaller compared to the former case. In this case, although the mean of I_{gate} is about two times smaller than that of I_{sub}, the standard deviation is 3 times larger than that of I_{sub}. Therefore, although I_{sub} and I_{gate} are both major contributors to the full-chip leakage, the leakage variations are mainly due to I_{gate}.

7. CONCLUSION

We have presented a method for analyzing the leakage current distribution of circuit under process parameter variations consid-

Table 1: Comparison of the proposed method with Monte Carlo simulation

Circuit	Gate	Grid	Monte Carlo (MC)		Our Method		Error%		MCNoCorr		Error%	
Name	Number	Number	mean(nA)	std(nA)	mean(nA)	std(nA)	mean	std	mean(nA)	std(nA)	mean	std
c7552	5528	64	327.9	106.1	324.3	101.0	-1.1%	-4.9%	327.8	90.7	0.0%	-14.5%
c5315	3887	64	239.0	78.4	235.7	74.3	-1.4%	-5.2%	239.5	67.2	0.2%	-14.3%
c6288	2672	16	229.6	77.3	227.7	78.0	-0.8%	0.8%	229.7	71.8	0.0%	-7.1%
c3540	2606	16	158.9	53.4	156.8	50.9	-1.3%	-4.7%	158.3	44.1	-0.4%	-17.4%
c2670	1925	16	113.7	37.8	112.6	36.6	-1.0%	-3.3%	113.9	31.7	0.2%	-16.3%
c1908	1261	16	73.5	24.9	72.3	23.5	-1.6%	-5.6%	73.2	20.1	-0.4%	-19.1%
c880	594	4	37.4	13.3	36.9	12.7	-1.3%	-4.6%	37.3	10.5	-0.3%	-21.4%
c432	294	4	18.3	6.5	17.9	6.2	-1.8%	-5.0%	18.2	5.1	-0.4%	-21.5%

Table 2: Comparison of leakage by varying L and T_{ox} independently

Circuit Name	Leakage by varying effective gate length only (nA)						Leakage by varying gate oxide thickness only (nA)					
	I_{total}		I_{sub}		I_{gate}		I_{total}		I_{sub}		I_{gate}	
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
c7552	268.2	81.3	216.2	83.8	52.0	2.7	298.9	63.1	195.1	34.0	103.8	88.2
c5315	194.3	60.6	155.3	62.5	39.0	2.0	217.4	47.6	139.5	24.4	77.9	65.8
c6288	178.5	46.7	131.2	49.1	47.4	2.6	215.0	63.8	120.4	19.6	94.6	79.2
c3540	129.4	42.2	103.3	43.6	26.1	1.5	144.4	31.7	92.9	15.9	51.5	43.7
c2670	92.9	29.9	74.6	30.8	18.3	1.0	103.4	21.9	67.2	11.5	36.2	30.4
c1908	60.4	20.5	49.2	21.1	11.2	0.6	66.5	13.1	44.0	7.6	22.5	18.8
c880	30.6	10.9	24.5	11.2	6.1	0.4	34.1	7.5	22.0	3.8	12.1	10.4
c432	15.1	5.6	12.5	5.8	2.6	0.2	16.4	3.1	11.2	2.0	5.3	4.5
Avg	121.2	37.2	95.9	38.5	25.3	1.4	137.0	31.5	86.5	14.9	50.5	42.6

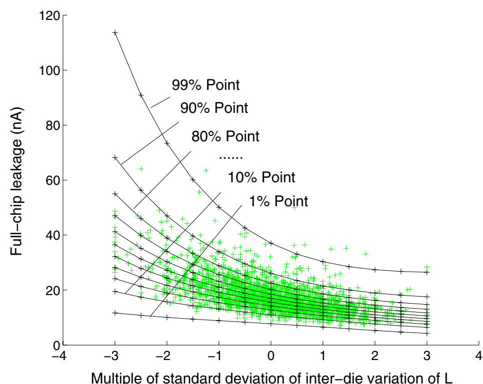

Figure 3: Scatter plot of full-chip leakage considering spatial correlation for circuit c432

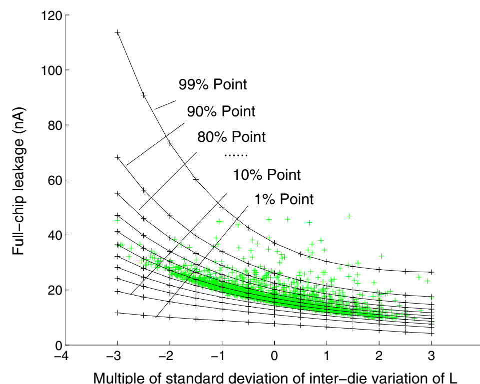

Figure 4: Scatter plot of full-chip leakage ignoring spatial correlation for circuit c432

ering the spatial correlations among parameters. The proposed method was shown to be effective in predicting the mean, standard deviation and the CDF/PDF of the total chip leakage. We have also shown that the spatial correlations of parameters must be considered appropriately in order to predict yield of chip correctly. We believe that this framework is general to predict the circuit leakage under other parameter variations. For example, leakage has a strong dependence on temperature and the variation of temperature is also highly spatially correlated. If the correlation statistics are available, this method can easily be extended to capture the effects of temperature variations.

8. REFERENCES

[1] A. A. Abu-Dayya and N. C. Beaulieu, "Comparison of methods of computing correlated lognormal sum distributions and outages for digital wireless applications," *IEEE 44th Vehicular Technology Conference*, vol. 1, pp. 175-179, 1994.

[2] K. A. Bowman, L. Wang, X. Tang and J. D. Meindl, "A Circuit Level Perspective of the Optimum Gate Oxide Thickness," *IEEE Transction on Electron Devices*, pp. 1800-1810, 2001.

[3] H. Chang and S. S. Sapatnekar, "Statistical Timing Analysis Considering Spatial Correlations Using a Single PERT-like Traversal," *International Conference on Computer Aided Design*, pp. 621–625, 2003.

[4] M. Ketkar and S. S. Sapatnekar", "Standby Power Optimization via Transistor Sizing and Dual Threshold Voltage Assignment," *International Conference on Computer-Aided Design*, pp. 375-378, 2002.

[5] D. Lee, W. Kwong, D. Blaauw and D. Sylvester, "Analysis and Minimization Techniques for Total Leakage Considering Gate Oxide Leakage," *Design Automation Conference*, pp. 175-180, 2003.

[6] S. Mukhopadhyay and K. Roy, "Modeling and Estimation of Total Leakage Current in Nano-scaled CMOS Devices Considering the Effect of Parameter Variation," *International Symposium on Low Power Electronics and Design*, pp. 172-175, 2003.

[7] S. Narendra, V. De, S. Borkar, D. Antoniadis and A. Chandrakasan, "Full-chip sub-threshold leakage power prediction model for sub-0.18um CMOS," *International Symposium on Low Power Electronics and Design*, pp. 19-23, 2002.

[8] S. Nassif, "Delay Variability: Sources, Impact and Trends," *IEEE International Solid-State Circuits Conference*, pp. 368-369, 2000.

[9] R. Rao, A. Devgan, D. Blaauw and D. Sylvester, "Parametric yield estimation considering leakage variability," *Design Automation Conference*, pp. 442-447, 2003

[10] R. Rao, A. Srivastava, D. Blaauw and D. Sylvester, "Statistical Estimation of Leakage Current Considering Inter- and Intra-Die Process Variation," *International Symposium on Low Power Electronics and Design*, pp. 19-23, 2003.

[11] S. Sirichotiyakul, T. Edwards, C. Oh, J. Zuo, A. Dharchoudhury, R. Panda and D. Blaauw, "Stand-by power minimization through simultaneous threshold voltage selection and circuit sizing," *Design Automation Conference*, pp. 436-441, 1999.

[12] A. Srivastava, R. Bai, D. Blaauw and D. Sylvester, "Modeling and analysis of leakage power considering within-die process variations," *International Symposium on Low Power Electronics and Design*, pp. 64-67, 2002.

[13] A. Sultania, D. Sylvester, and S. S. Sapatnekar, "Tradeoffs between Gate Oxide Leakage and Delay for Dual Tox Circuits," *Design Automation Conference*, pp. 761 - 766, 2004.

[14] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*, Cambridge Univerity Press, Cambridge, UK, 1998.

[15] "Capo: A large-scale fixed-die placer from UCLA," Available at: `http://vlsicad.ucsd.edu/GSRC/bookshelf/Slots/Placement`.

[16] Semiconductor Industry Association, "International Technology Roadmap for Semiconductors,"2004. Available at : `http://public.itrs.net`.