Prediction of Leakage Power Under Process Uncertainties

HONGLIANG CHANG
Cadence Design Systems
and
SACHIN S. SAPATNEKAR
Department of Electrical and Computer Engineering, University of Minnesota

In this paper, we present a method to analyze the total leakage current of a circuit under process variations, considering inter-die and intra-die variations as well as the effect of the spatial correlations of intra-die variations. The approach considers both the subthreshold and gate-tunneling leakage power, as well as their interactions. With process variations, each leakage component is approximated by a lognormal distribution, and the total chip leakage is computed as a sum of the correlated lognormals. Since the lognormals to be summed are large in number and have complicated correlation structures due to spatial correlations and the correlation among different leakage mechanisms, we propose an efficient method to reduce the number of correlated lognormals for summation to a manageable number by identifying dominant states of leakage currents and taking advantage of the spatial correlation model and input states at the gates. An improved approach utilizing principal components computed from the spatially correlated process parameters is also proposed to further improve run-time efficiency. We show that the proposed methods are effective in predicting the probability distribution of total chip leakage, and that ignoring spatial correlations can underestimate the standard deviation of full-chip leakage power.

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

General Terms: Algorithms, Design, Performance

Additional Key Words and Phrases: Circuit, Leakage, Yield, Process Variation

Author's address: Hongliang Chang (hchang@cadence.com), Cadence Design Systems, 555 River Oaks Pkwy, Bldg. 2, San Jose, CA, 95134. Sachin S. Sapatnekar (sachin@umn.edu), Department of Electrical and Computer Engineering, 200 Union Street SE, University of Minnesota, MN, 55455

This work was supported in part by the NSF under award CCR-0205227 and by the SRC under contract 2003-TJ-1092. This paper is an extended version of an earlier conference paper [Chang and Sapatnekar 2005]. The enhancements to the conference version include a more complete description of the models and the algorithm, and the addition of an entirely new algorithm in section 5 with the corresponding experimental results in section 6.

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) [Semiconductor Industry Association 2005], 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 [Lee et al. 2003], 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 [Sirichotiyakul et al. 1999; Ketkar and Sapatnekar 2002]. 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, it is unclear when these will be introduced, and gate leakage is already seen to be very significant in 90nm, 65nm and 45nm technologies [Semiconductor Industry Association 2005], so that its analysis is of profound importance.

In the literature, several research works on the analysis and minimization of total circuit leakage including the effect of I_{gate} have been conducted [Lee et al. 2003]. 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 (PDF). 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.

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 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 have focused on statistical timing analysis, and only a few have investigated the variation of leakage power under the effect of process variations [Narendra et al. 2002; Srivastava et al. 2002; Mukhopadhyay and Roy 2003; Rao et al. 2003; Rao et al. 2004]. In [Srivastava et al. 2002; Narendra et al. 2002], analytical methods were proposed to estimate the mean and standard deviation of the total chip subthreshold leakage power under intra-die parameter variations. In [Mukhopadhyay and Roy 2003], 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 [Rao et al. 2003], the probability density function of the total chip subthreshold leakage was derived. The authors of [Rao et al. 2004] 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.

The remainder of the paper is organized as follows. Section 2 formulates the problem to solve in this work. Section 3 describes the models of process variation and spatial correlation. A first method for estimating the distribution of full-chip leakage power is given in section 4, and this is followed by an improved approach, presented in section 5. Finally, a list of experimental results are shown in section 6.

2. PROBLEM DESCRIPTION

The total leakage power consumption of a circuit is input-pattern-dependent, i.e., the value differs as the input signal to the circuit changes, because the leakage power consumption, due to subthreshold and gate tunneling leakage, of a gate depends on the input vector state at the gate. As illustrated in [Acar et al. 2003], the dependency of leakage on process variations is more significant than on input vector states. Therefore, it is sufficient to predict the effects of process variations on total circuit leakage by studying the variation of average leakage current for all possible input patterns to the circuit. However, it is impractical to estimate the average leakage by simulating the circuit at all input patterns, and thus an input pattern-independent approach is more desirable.

In switching power estimation, probabilistic approaches [Najm 1994] have been used for this purpose. The work of [Acar et al. 2003] proposed a similar approach that computes the average leakage current of each gate and estimates the total average circuit leakage as a sum of the average leakage currents of all gates:

$$I_{tot}^{avg} = \sum_{k=1}^{N_g} I_{leak,k}^{avg} = \sum_{k=1}^{N_g} \sum_{\forall vec_{i,k}} Prob(vec_{i,k}) \cdot I_{leak,k}(vec_{i,k}) \tag{1}$$

where N_g is the total number of gates in the circuit, $I_{leak,k}^{avg}$ is the average leakage current of the k^{th} gate, $vec_{i,k}$ is the i^{th} input vector at the k^{th} gate, $Prob(vec_{i,k})$

is the probability of occurrence of $vec_{i,k}$, and $I_{leak,k}(vec_{i,k})$ is the leakage current of the k^{th} gate when the gate input vector is $vec_{i,k}$.

In this work, we will solve the problem of computing the probability distribution of the average circuit leakage current I_{tot}^{avg}, formulated in Equation (1), under process variations, with both subthreshold and gate tunneling leakage currents taken into account in the computation.

3. MODELING PROCESS PARAMETER VARIATIONS

In general, a process parameter variation δ_{total} can be decomposed into:

$$\delta_{total} \;=\; \delta_{inter} + \delta_{intra}, \tag{2}$$

where δ_{inter} is the inter-die variation and δ_{intra} is the intra-die variation, where δ_{inter} and δ_{intra} can be both modeled as Gaussian random variables.

The inter-die variation is the variation of parameter values across identically manufactured chips. 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.

The intra-die variation is the difference of parameter values across identically manufactured transistors [wires] inside a chip. Intra-die variation can be divided into systematic and random variations. The systematic variations are those that may be modeled deterministically, and the random variations are the remaining unmodeled variations. For intra-die variation δ_{intra}, we use the same model as in the work of [Chang and Sapatnekar 2003], in which, under intra-die variation, the value of a parameter p located at (x, y) can be modeled as:

$$p \;=\; \bar{p} + \delta_x x + \delta_y y + \epsilon \tag{3}$$

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, corresponding to the slowly and smoothly varying global systematic trend spatially across the die. The term, ϵ, stands for the remaining uncertainties or unmodeled intra-die variation. The vector of all random components across the chip has a correlated multivariate normal distribution due to spatial correlation in the intra-die variation: $\vec{\epsilon} \sim N(0, \Sigma)$, where Σ is the covariance matrix of the spatially correlated parameters, as will be described in the remainder of this section.

The spatial correlation of intra-die variation identifies the extent of any remaining unexplained systematic variation in the residual term that is left after decomposition of process variations [Stine et al. 1997]. The result of spatial correlation is that devices [wires] located close to each other are more likely to have the similar characteristics than those placed far away. The spatial correlation can be modeled as a function of separation distance as in [Friedberg et al. 2005], or using a grid-based model as in [Agarwal et al. 2003; Chang and Sapatnekar 2003; Xiong et al. 2006]. In this work, we employ the model of [Chang and Sapatnekar 2003] which models intra-die spatial correlations of parameters by partitioning the die region into $nrow \times ncol = n$ grids. Since devices [wires] close to each other are more likely to have more similar characteristics than those placed far away, perfect correlations are assumed among the devices [wires] in the same grid, high correlations among

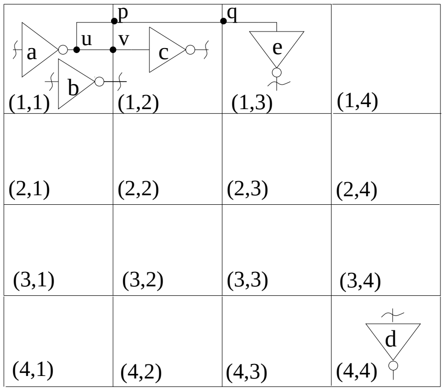


Fig. 1. Grid model for spatial correlations

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. Gates a and c lie in neighboring grids, and their parameter variations are not identical but highly correlated (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).

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 L_{eff} and gate oxide thickness T_{ox}, since I_{sub} and I_{gate} are most sensitive to these parameters [Taur and Ning 1998; Mukhopadhyay and Roy 2003]. 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. Note that 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 N_{sub}, etc.

4. COMPUTING THE DISTRIBUTION OF FULL-CHIP LEAKAGE CURRENT

We will now present the method used to estimate the distribution of average full-chip leakage current, I_{tot}^{avg}, under process variations. As implied by Equation (1), the distribution of I_{tot}^{avg} can be calculated in two steps: first, computing the distribution of each $I_{leak,k}(vec_{i,k})$, the leakage current of the k^{th} gate when the gate input vector is $vec_{i,k}$; and second, finding the distribution of the weighted sum of all $I_{leak,k}(vec_{i,k})$ terms. Since each $I_{leak,k}(vec_{i,k})$ can further be decomposed into $I_{sub,k}(vec_{i,k}) + I_{gate,k}(vec_{i,k})$, where $I_{sub,k}(vec_{i,k})$ and $I_{gate,k}(vec_{i,k})$ are the subthreshold and gate tunneling leakage currents, respectively, for the k^{th} gate with input state $vec_{i,k}$, I_{tot}^{avg} can be computed as:

$$I_{tot}^{avg} = \sum_{k=1}^{N_g} \sum_{\forall vec_{i,k}} Prob(vec_{i,k}) \cdot (I_{sub,k}(vec_{i,k}) + I_{gate,k}(vec_{i,k})) \tag{4}$$

In the discussion that follows, we will first present how the distributions of subthreshold leakage current, $I_{sub,k}(vec_{i,k})$, and gate tunneling leakage current, $I_{gate,k}(vec_{i,k})$, are estimated in section 4.1 and 4.2, respectively. The analytical approach to obtain the probability density function for the total weighted sums of all $I_{sub,k}(vec_{i,k})$ and $I_{gate,k}(vec_{i,k})$ terms will then be presented in section 4.3. As the same framework can be applied for computing the distribution of each $I_{sub,k}(vec_{i,k})$, for conciseness, we will use I_{sub} for $I_{sub,k}(vec_{i,k})$, and similarly, I_{gate} for $I_{gate,k}(vec_{i,k})$, in later sections.

4.1 Distribution of Subthreshold Leakage Current

The commonly used model for subthreshold leakage current through a transistor expresses this current as [Taur and Ning 1998]:

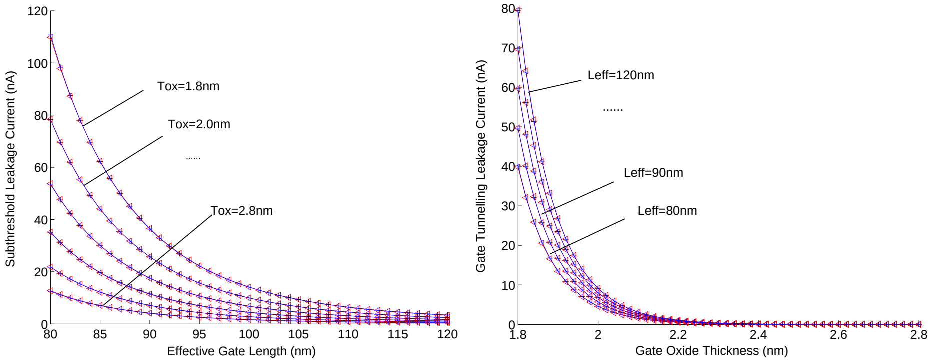
$$I_{sub} = I_0 e^{(V_{gs}-V_{th})/n_s V_T}(1 - e^{-V_{ds}/V_T}) \tag{5}$$

Here, $I_0 = \mu_0 C_{ox}(W_{eff}/L_{eff})V_T^2 e^{1.8}$, where μ_0 is zero bias electron mobility, C_{ox} is the gate oxide capacitance, W_{eff} and L_{eff} are the effective transistor width and length, respectively, V_{gs} and V_{ds} are gate-to-source voltage and drain-to-source voltage, respectively, n_s is the subthreshold slope coefficient, $V_T = kT/q$ is the thermal voltage, where k is Boltzman constant, T is the operating temperature in Kelvin (K), q is charge on an electron, and V_{th} is the subthreshold voltage.

It is observed that V_{th} is most sensitive to gate oxide thickness T_{ox} and effective transistor gate length L_{eff} due to short-channel effects [Taur and Ning 1998]. Due to the exponential dependency of I_{sub} on V_{th}, a small change on L_{eff} or T_{ox} will have a substantial effect on I_{sub}. From this intuition, we estimate the subthreshold leakage current per transistor width by developing an empirical model through curve-fitting, similarly to [Mukhopadhyay and Roy 2003; Rao et al. 2003]:

$$I_{sub} = c \times e^{a_1 + a_2 L_{eff} + a_3 L_{eff}^2 + a_4 T_{ox}^{-1} + a_5 T_{ox}} \tag{6}$$

where c and the a_i terms are the fitting coefficients. To quantify the empirical model, the values of I_{sub} achieved from expression (6) are compared with those through SPICE simulations over a ranged values of T_{ox} and L_{eff}. As an example, for the NMOS transistor of an inverter in a 100 nm technology [BPT], the curves of I_{sub} are plotted as a function L_{eff} under several fixed values of T_{ox} in Figure 2(a). It can be seen that the simulation results from the empirical model fits well with those from the SPICE model, and similar results are observed for other types of cells.



(a) Subthreshold leakage current

(b) Gate-tunneling leakage current

Fig. 2. Comparison of the curve fitted empirical model with qualified models (SPICE and [Bowman et al. 2001]) for the leakage current of the NMOS transistor of an inverter (100 nm technology). Points on the curve corresponding to the empirical model are marked with the symbol "+" and those from the SPICE simulation are marked with triangles. The two curves show a near-perfect match.

In this way, I_{sub} is modeled as an exponential function in the form of $c \times e^U$, where U is an explicit function of L_{eff} and T_{ox}. When L_{eff} and T_{ox} show process variations, the exponent U, and thus I_{sub}, become random variables. Since the magnitude of process variations is observed to be around $10-20\%$ in practice, I_{sub} can be well approximated by expanding its exponent U using a first-order Taylor expansion at the nominal values of the process parameters:

$$I_{sub} = c \times e^{U_0+\beta_1\cdot\Delta L_{eff}+\beta_2\cdot\Delta T_{ox}} \tag{7}$$

where U^0 is the nominal value of the exponent U, β_0 and β_1 are the derivatives of U to L_{eff} and T_{ox} evaluated at their nominal values, respectively, and ΔL_{eff} and ΔT_{ox} are random variables standing for the variations in the process parameters L_{eff} and T_{ox}, respectively.

Expression (7) for I_{sub} can also be written as $e^{ln(c)+U_0+\beta_1\cdot\Delta L_{eff}+\beta_2\cdot\Delta T_{ox}}$[1]. Since ΔL_{eff} and ΔT_{ox} are assumed to be Gaussian-distributed, I_{sub} is seen as an exponential function of a Gaussian random variable, with mean $ln(c)+U_0$ and standard deviation $\sqrt{\beta_1^2\sigma_{L_{eff}}^2+\beta_2^2\sigma_{T_{ox}}^2}$, where $\sigma_{L_{eff}}$ and $\sigma_{T_{ox}}$ are standard deviations of ΔL_{eff} and ΔT_{ox}, respectively.

[1] To consider the effect of varying N_{sub} on I_{sub}, the equation (7) can be adapted by adding an additional term for ΔN_{sub} in the exponent. As in the case of T_{ox}, the variation of N_{sub} does not show spatial correlation, and thus N_{sub} can be handled using a similar method as was used for T_{ox} in the framework.

In general, if x is a Gaussian random variable, then $z = e^x$ is a lognormal distributed random variable and the probability density function of z is given by [Papoulis and Pillai 2002]:

$$f(z) = \frac{1}{z\sqrt{2\pi}\sigma} e^{-(ln(z)-\mu)^2/(2\sigma^2)} \tag{8}$$

where μ and σ are the mean and standard deviation of the Gaussian random variable x, respectively. Therefore, it is obvious that I_{sub} can be approximated as a lognormally distributed random variable whose probability density function can be characterized using the values of c, U_0 and β_i's.

Since subthreshold leakage current has a well-known input state dependency due to the stack effect [Sirichotiyakul et al. 1999], the PDFs of subthreshold leakage currents must be characterized for all possible input states for each type of gate in the library, for which the same approach described in this section can be applied. Once the library is characterized, a simple look-up table (LUT) can then be used to retrieve the corresponding model characterized given the gate type and input vector state at a gate.

4.2 Distribution of Gate Tunneling Leakage Current

In [Bowman et al. 2001], an analytical model was proposed for the gate oxide tunneling current density J_{tunnel}.

$$J_{tunnel} = \frac{4\pi m^* q}{h^3}(kT)^2(1 + \frac{\gamma kT}{2\sqrt{E_B}})e^{\frac{E_{F0,Si/SiO_2}}{kT}} e^{-\gamma\sqrt{E_B}} \tag{9}$$

Here m^* is the transverse mass that equals $0.19m_0$ for electron tunneling and $0.55m_0$ for hole tunneling, where m_0 is the free electron rest mass, h is Planck's constant, γ is defined as $4\pi T_{ox}\sqrt{2m_{ox}}/h$, where m_{ox} is the effective electron [hole] mass in the oxide, E_B is the barrier height, $E_{F0,Si/SiO_2} = q\phi_S - q\phi_F - E_G/2$ is the Fermi level at the Si/SiO_2 interface, where ϕ_S is surface potential, ϕ_F is the Fermi energy level potential, either in the Si substrate for the gate tunneling current through the channel, or in the source/drain region for the gate tunneling current through the source/drain overlap, and E_G is the Si band gap energy.

In [Bowman et al. 2001], the gate-tunneling current of PMOS devices is neglected due to the larger effective mass and barrier height for holes compared to electrons at the $\mathrm{SiO_2}$/Si interface. Moreover, 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 [Sultania et al. 2004]. Therefore, in this work, the gate tunneling leakage current is taken into account only for NMOS transistors at logic "1"[2].

Although the formulation (9) possesses a high accuracy, it does not lend itself easily to the analysis of the effects of parameter variations. Therefore, we again use

[2] The proposed method can be easily adapted to include the effect of gate leakage of PMOS devices by appropriately modeling their leakage currents as for the NMOS devices, and the same framework in the paper can be used to find full-chip leakage.

an empirically characterized model to estimate I_{gate} per transistor width through curve-fitting:

$$I_{gate} = c' \times e^{b_1 + b_2 L_{eff} + b_3 L_{eff}^2 + b_4 T_{ox} + b_5 T_{ox}^2} \tag{10}$$

where c' and the b_i terms are the fitting coefficients. The empirical model is qualified by comparing it with that of formulation (9), over a range of values of T_{ox} and L_{eff}. In Figure 2(b), for the NMOS transistor of an inverter in 100 nm technology, the value of I_{gate} is plotted as a function of T_{ox} under several fixed values of L_{eff}. The curves for formulation (10) are marked with the symbol "+" and those for (9) by triangles. It can be observed that there are close matches between the model (10) and (9), and similar results are observed for all other types of cells.

Similar to the method for estimating the distribution of I_{sub}, under the variations of L_{eff} and T_{ox}, I_{gate} can be approximated by applying first-order Taylor expansion to the exponent U' of Equation (10):

$$I_{gate} = c' \times e^{U'_0 + \lambda_1 \cdot \Delta L_{eff} + \lambda_2 \cdot \Delta T_{ox}} \tag{11}$$

where U'_0 is the nominal value of the exponent U', and λ_0 and λ_1 are the derivatives of U' to L_{eff} and T_{ox} evaluated at their nominal values, respectively.

Under this approximation, I_{gate} becomes a lognormally distributed random variable, and its PDF can be characterized through the values of c', U'_0 and λ'_i. Since the gate tunneling leakage current is input state dependent, the PDFs of the I_{gate} variables are characterized for all possible input states for each type of gate in the library, and a simple look-up table (LUT) is used for model retrieval while evaluating a specific circuit.

4.3 Distribution of Full-Chip Leakage Current

Sections 4.1 and 4.2 show that each of $I_{sub,k}(vec_{i,k})$ or $I_{gate,k}(vec_{i,k})$, i.e., the subthreshold or gate tunneling leakage current, respectively, of the k^{th} gate when its input vector is $vec_{i,k}$, can be modeled as a lognormal random variable under process variations. In this section, we will present the approach to find the distribution of I_{tot}^{avg} as formulated in Equation (4), which is a weighted sum of all $I_{sub,k}(vec_{i,k})$ and $I_{gate,k}(vec_{i,k})$ variables, weighted by $Prob(vec_{i,k})$ terms, the probabilities of input vector $vec_{i,k}$ at the gate. Since the probability of each $vec_{i,k}$ can be computed by specifying signal probabilities at the circuit primary inputs and propagating the probabilities into all gates pins in the circuit, as in [Acar et al. 2003], in this section, we focus on the computation of the PDF of the weighted sum.

As each of $I_{sub,k}(vec_{i,k})$ or $I_{gate,k}(vec_{i,k})$ has a lognormal distribution, it can easily be seen that any multiplication by a constant maintains this property; specifically, $Prob(vec_{i,k}) \cdot I_{sub,k}(vec_{i,k})$ and $Prob(vec_{i,k}) \cdot I_{gate,k}(vec_{i,k})$ are both lognormally distributed. Therefore, the problem of calculating the distribution of I_{tot}^{avg} becomes that of computing the PDF of the sum of a set of lognormal random variables. Furthermore, the set of lognormal random variables in the summation could be correlated since:

—the leakage current random variables for any two gates may be correlated due to spatial correlations of intra-die variations of process parameters.

—within the same gate, the subthreshold and gate tunneling leakage currents are correlated, and the leakage currents under different input vectors are correlated, because they are sensitive to the same process parameters of the same gate, regardless of whether these are spatially correlated or not.

In this section, we will present an efficient approach to predict the probability density function of the full-chip leakage current, by computing the PDF of the sum of correlated lognormal random variables, so that the spatial correlations of process parameters, and correlations between different leakage components can be correctly taken into account. This section is organized as follows. We first describe Wilkinson's method [Abu-Dayya and Beaulieu 1994] for approximating a sum of correlated lognormal random variables. Next, a more efficient approach is proposed to reduce the computational complexity of this calculation. For clarity, the approach described first considers only intra-die variations of process parameters. The extension to handling inter-die variations is trivial, and will be shown briefly in the end of this section.

4.3.1 *Finding the Sum of Correlated Lognormals by Wilkinson's Method.* Theoretically, the sum of several lognormal distributed random variables is not known to have a closed form. However, it may be well approximated as a lognormal, as is done in Wilkinson's method [Abu-Dayya and Beaulieu 1994][3]. That is, the sum of m lognormals, $S = \sum_{i=1}^{m} e^{Y_i}$, where each Y_i is a normal random variable with mean m_{y_i} and standard deviation σ_{y_i}, and the Y_i variables can be correlated or uncorrelated, can be approximated as a lognormal e^Z, where Z is normally distributed, with mean m_z and standard deviation σ_z. In Wilkinson's approach, the values of m_z and σ_z are obtained by matching the first two moments, u_1 and u_2, of e^Z and S as follows:

$$u_1 = E(e^Z) = E(S) = \sum_{i=1}^{m} E(e^{Y_i}) \tag{12}$$

$$\begin{aligned} u_2 &= E(e^{2Z}) = E(S^2) = Var(S) + E^2(S) \qquad (13) \\ &= \sum_{i=1}^{m} Var(e^{Y_i}) + 2\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} cov(e^{Y_i}, e^{Y_j}) + E^2(S) \\ &= \sum_{i=1}^{m} Var(e^{Y_i}) + 2\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} \big(E(e^{Y_i}e^{Y_j}) - E(e^{Y_i})E(e^{Y_j})\big) + E^2(S) \end{aligned}$$

where $E(.)$ and $Var(.)$ are the symbols for the mean and variance values of a random variable, and $cov(.,.)$ represents the covariance between two random variables.

In general, the mean and variance of a lognormal random variable e^{X_i}, where X_i

[3] An approximation of the sum of correlated lognormal random variables by Monte Carlo simulations is computationally difficult for large-sized problems. As an alternative, three analytical approaches have been overviewed and compared in [Abu-Dayya and Beaulieu 1994]: Wilkinson's approach, Schwartz and Yeh's approach, and the cumulant-matching approach. Through numerical comparisons, [Abu-Dayya and Beaulieu 1994] concluded that Wilkinson's method is the best in terms of computational simplicity and accuracy, and this is why the Wilkinson's approach is selected in this paper for this approximation.

is normal distributed with mean m_{x_i} and standard deviation σ_{x_i}, is computed by:

$$\begin{aligned} E(e^{X_i}) &= e^{m_{x_i}+\sigma_{x_i}^2/2} && (14)\\ Var(e^{X_i}) &= e^{2m_{x_i}+2\sigma_{x_i}^2} - e^{2m_{x_i}+\sigma_{x_i}^2} && (15) \end{aligned}$$

The covariance between two lognormal random variables e^{X_i} and e^{X_j} can be computed by:

$$cov(e^{X_i}, e^{X_j}) = E(e^{X_i} \cdot e^{X_j}) - E(e^{X_i})E(e^{X_j}) \qquad (16)$$

Superposing Equations (14), (15) and (16) into Equations (12) and (13) results in:

$$\begin{aligned} u_1 &= E(e^Z) = e^{m_z+\sigma_z^2/2} = E(S) = \sum_{i=1}^{m} (e^{m_{y_i}+\sigma_{y_i}^2/2}) && (17)\\ u_2 &= E(e^{2Z}) = e^{2m_z+2\sigma_z^2} = E(S^2) && (18)\\ &= \sum_{i=1}^{m} (e^{2m_{y_i}+2\sigma_{y_i}^2} - e^{2m_{y_i}+\sigma_{y_i}^2}) + 2\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} (e^{m_{y_i}+m_{y_j}+(\sigma_{y_i}^2+\sigma_{y_j}^2+2r_{ij}\sigma_{y_i}\sigma_{y_j})/2}\\ &\quad - e^{m_{y_i}+\sigma_{y_i}^2/2} e^{m_{y_j}+\sigma_{y_j}^2/2}) + u_1^2 \end{aligned}$$

where r_{ij} is the correlation coefficient between Y_i and Y_j.

Solving (17) and (18) for m_z and σ_z yields:

$$\begin{aligned} m_z &= 2\ln u_1 - \frac{1}{2}\ln u_2 && (19)\\ \sigma_z^2 &= \ln u_2 - 2\ln u_1 && (20) \end{aligned}$$

The computational complexity of Wilkinson's approximation can be analyzed through the cost of computing m_z and σ_z. The computational complexities of m_z and σ_z are determined by those of u_1 and u_2, whose values can be obtained using the formulas in (17) and (18). It is clear that the computational complexity of u_1 is dominated by that of u_2, since the complexity of calculating u_1 is $O(m)$, while that of u_2 is $O(m \cdot N_{corr})$, where N_{corr} is the number of correlated pairs among all pairs of Y_i variables. The cost of computing u_2 can also be verified by examining the earlier expression of u_2 in (13), in which the second term in the summation, in fact, corresponds to the covariance of Y_i and Y_j, which becomes zero when Y_i and Y_j are uncorrelated. Therefore, if $r_{ij} \neq 0$ for all pairs of Y_i and Y_j, the complexity of calculating u_2 is $O(m^2)$; if $r_{ij} = 0$ for all pairs of i and j, the complexity is $O(m)$.

As explained earlier, for full-chip leakage analysis, the number of correlated lognormal distributed leakage components in the summation could be extremely large, which could lead to a prohibitive amount of computation. If Wilkinson's method is applied directly, when the total number of gates in the circuit is N_g, the complexity for computing the sum will be $O(N_g^2)$, which is impractical for large circuits. In the remainder of this section, we will propose to compute the summation in a more efficient way.

4.3.2 *Reducing the Number of Correlated Lognormals to be Summed.* Since Wilkinson's method has a quadratic complexity with respect to the number of correlated lognormals to be summed, we now introduce mechanisms to reduce the number of correlated lognormals in the summation, to improve the computational efficiency.

First, the number can be reduced by identifying dominant states for subthreshold and gate tunneling leakage currents for each type of gate in the circuit.

Due to state dependencies of subthreshold and gate tunneling leakage currents, the computation of full-chip leakage current must take into account all possible input patterns at all gates in the circuit. In general, for a gate with N_{inpin} input pins, the number of input states to be considered can be $2^{N_{inpin}}$. However, the leakage currents at some input states may not be as important as at others. It is sufficient to identify the important ones, corresponding to dominant states, and consider the leakage currents only at dominant states without losing much of accuracy.

When only subthreshold leakage current is considered, the dominant states for subthreshold leakage current in a transistor stack correspond to those with only one "off" transistor in the pull-up or pull-down chain [Sirichotiyakul et al. 1999; Ketkar and Sapatnekar 2002]. In this way, for a transistor stack of length q, the number of input states to consider is reduced to a much smaller size, q instead of 2^q. However, when gate tunneling leakage current is also considered, the dominant states must be characterized based on both leakage mechanisms and their interactions.

The interaction effects between the two mechanisms are analyzed in [Lee et al. 2003] by studying three scenarios for the middle transistor t_n in a NMOS transistor stack of length 3, as shown in Figure 3: in scenario (a) where t_n has a conducting path to supply and nonconducting path to gate output, I_{gate} does not interact with I_{sub} in the stack and the total leakage in stack is the sum of the two; in scenario (b) where t_n has a nonconducting path to supply and conducting path to gate output, I_{gate} is one order of magnitude smaller than that of case (a) and can be ignored safely; in scenario (c) where t_n has a nonconducting path to supply and gate output, due to the interaction between I_{sub} and I_{gate}, I_{sub} can be ignored safely. For details, the reader is referred to [Lee et al. 2003].

The analysis shows that a dominant state for subthreshold leakage current may not be one for gate tunneling leakage current, e.g., scenario (b) is a dominant state for I_{sub}, but not I_{gate}, and scenario (c) is a dominant state for I_{gate}, but not I_{sub}. Therefore, one way of identifying the dominant states for leakage current of a gate is to separately determine the set of dominant states for the subthreshold and gate tunneling leakage currents. From the analysis above, the dominant states for subthreshold and gate tunneling leakage currents can be identified by the following rules. For a transistor stack, the set of dominant states for subthreshold leakage current remains being the one with only one "off" transistor in the pull-up or pull-down chain, since the value of I_{sub} is strongly reduced only when there is more than one "off" transistor in the pull-up or pull-down chain. The determination of dominant states for gate tunneling leakage current is based on the following rule: in a transistor stack, the gate tunneling leakage current of a transistor is negligible if there is a conducting path to the gate output from this transistor.

To show the accuracy of leakage current estimation considering only dominant states under process variations, for each type of gate in library, we compare, by Monte Carlo simulation, the distribution of the average subthreshold leakage current, $\sum_{\forall vec_{i,k}} Prob(vec_{i,k}) \cdot I_{sub,k}(vec_{i,k})$, and the distribution of average gate tun-

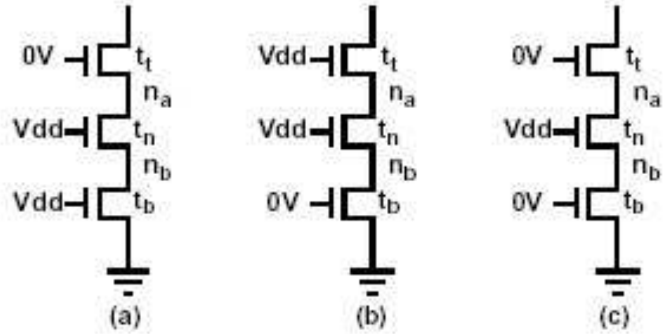
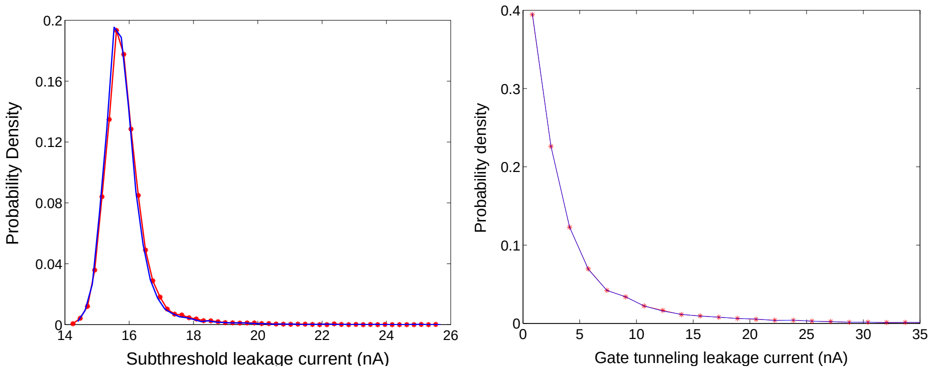


Fig. 3. Three scenarios of combined I_{sub} and I_{gate} for a three-input NMOS transistor stack [Lee et al. 2003].

neling leakage current, $\sum_{\forall vec_{i,k}} Prob(vec_{i,k}) \cdot I_{gate,k}(vec_{i,k})$, using only dominant states with that using a full set of input vectors, assuming all input vectors having equal probabilities of occurrence. As an example, Figure 4(a) shows the PDF curves for simulations, with dominant states only and with the full set of states, for the average subthreshold leakage current of a 3-input NAND gate when the 3σ values of L_{eff} and T_{ox} are 20% of the mean. A close match is observed between these two PDF curves, and the same observation can be made when we compare the PDF curves of gate leakage for a 3-input NAND gate, using full-simulation and dominant states, as shown in Figure 4(b). For all types of gates in our library, the error can be controlled within 2%.



(a) Average subthreshold leakage current

(b) Average gate-tunneling leakage current

Fig. 4. Comparison of PDFs of average leakage currents using dominant states with that of full input vector states for a 3-input NAND gate, by Monte Carlo simulation with 3σ variations of L_{eff} and T_{ox} 20%. The solid curve shows the result when only dominant states are used, and the starred curve corresponds to simulation with all input vector states.

Second, instead of directly computing the sum of random variables of all leakage current terms, by grouping leakage current terms by model and grid location, and calculating the sum in each group separately first, the computational complexity in the computation of full-chip leakage reduces to quadratic in the number of groups.

This is because, as will be explained in this section, the summation in each group can be computed in linear time with respect to the number of leakage terms in each group. The results of the sums in all groups are then approximated as correlated lognormal random variables that can be then computed directly using Wilkinson's method. Since the number of groups is relatively small, a calculation that is quadratic in the number of groups is practically very economical.

Consider any dominant state for subthreshold leakage current that has only one "off" transistor in the transistor stack. It is observed that the values of subthreshold leakage currents *per unit width*, and thus their probabilistic distributions under process variations, 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 gate. For example, it is observed that the subthreshold leakage current per unit transistor width is the same for the pull-down of a NAND4 in state 0111, a NAND3 in state 011, a NAND2 in state 01, and an INV in state 0. Therefore, this equivalence can be used to compactly store the PDF of the subthreshold leakage current per unit width in an LUT, and different types of gates, with different stack lengths, can be characterized by the same LUT entry. If q is the length of the longest stack in the library, the number of different models is $2q$ in the LUT of I_{sub} (q each for I_{sub} for the PMOS and the NMOS).

For a dominant state of the gate tunneling leakage current, it is observed that if a transistor shows gate tunneling leakage, the value and probability distribution of I_{gate} can be determined by the number of "off" transistors between the leaking transistor and its supply in the transistor stack. In this way, the number of distinct models that store the gate tunneling leakage current per unit width is limited. Specifically, the total number of different models used in the LUT is only $q-2$, if the length of the longest stack in the library has length q.

Therefore, the total number of distinct models used in the LUT for the PDFs of the subthreshold and gate tunneling leakage currents is reduced to $2q+q-2$, where q is the length of the longest stack in the library. Next, we will show that if the leakage current terms to be summed in Equation (4) are grouped by the LUT model that they correspond to and their grid location, then the sum in each group can be computed in linear time with respect to the number of leakage terms in the group.

For illustration purposes, we only describe the computation of grouped sum for subthreshold leakage current term; the computation of gate leakage current proceeds along similar lines. The subthreshold leakage current term here refers to the term $Prob(vec_{i,k}) \cdot I_{sub,k}(vec_{i,k})$ in I_{tot}^{avg} in Equation (4). If $I_{sub,k}(vec_{i,k})$ corresponds to the p^{th} model in the LUT for PDF of subthreshold leakage current and it is located in the l^{th} grid, then $Prob(vec_{i,k}) \cdot I_{sub,k}(vec_{i,k})$ can be written as $\alpha e^{U_{0,p}+\beta_{1,p}\cdot\Delta L_{eff}^{l}+\beta_{2,p}\cdot\Delta T_{ox,k}}$, where the values of $U_{0,p}$, $\beta_{1,p}$ and $\beta_{2,p}$ come from the p^{th} model in the LUT; the coefficient α is $Prob(vec_{i,k}) \cdot W_{eff,k} \cdot c_p$, where c_p is the coefficient from the p^{th} model; ΔL_{eff}^{l} represents the variation of L_{eff} in the l^{th}

grid in the spatial correlation model, and $\Delta T_{ox,k}$ the variation of T_{ox} at this gate.

As we write the summation over all these lognormals, we observe that several different gates within the circuit may use the same LUT model: in fact, in general, the number of models is dramatically smaller than the total number of gates, and in practice, can be upper-bounded by a constant. Let $I_{sub,p,l} = \{I^1_{sub,p,l}, \cdots, I^s_{sub,p,l}\}$, where s is the size of the set, be the group of all subthreshold leakage current terms that use the p^{th} model in the LUT and lie in the l^{th} grid. Obviously, any $I^j_{sub,p,l}$ can be expressed in the form of:

$$I^j_{sub,p,l} = \alpha_j e^{U_{0,p}+\beta_{1,p}\cdot\Delta L^l_{eff}+\beta_{2,p}\cdot\Delta T_{ox,j}} \tag{21}$$

Note that each $I^j_{sub,p,l}$ has the same values of $U_{0,p}$, $\beta_{1,p}$ and $\beta_{2,p}$ from the p^{th} model, but the values of α_j may be different for different $I^j_{sub,p,l}$ terms, corresponding to different probabilities of occurrence, or different transistor widths. All $I^j_{sub,p,l}$ terms share the same variable ΔL^l_{eff} since they are in the same l^{th} grid, but each $I^j_{sub,p,l}$ has a different $\Delta T_{ox,j}$ variable, with all such $\Delta T_{ox,j}$ variables being independent of each other (since the values of gate oxide thickness are uncorrelated from gate to gate).

Then, the sum of all terms in $I_{sub,p,l}$ can be written as:

$$I^{sum}_{sub,p,l} = \sum_{j=1}^{s} I^j_{sub,p,l} = e^{U_{0,p}+\beta_{1,p}\cdot\Delta L^l_{eff}} \cdot \sum_{j=1}^{s} \alpha_j \cdot e^{\beta_{2,p}\cdot\Delta T_{ox,j}} \tag{22}$$

Due to the independence of the $T_{ox,j}$ variables, the sum $\sum_{j=1}^{s} \alpha_j \cdot e^{\beta_{2,p}\cdot\Delta T_{ox,j}}$ is in fact a sum of independent lognormal random variables. As explained earlier in the description of Wilkinson's method, the sum of independent lognormal random variables can be approximated by a lognormal random variable with computational complexity linear to the number of independent lognormals. Therefore, the product of the term, $e^{U_{0,p}+\beta_{1,p}\cdot\Delta L^l_{eff}}$, with the lognormal approximation of $\sum_{j=1}^{s} \alpha_j \cdot e^{\beta_{2,p}\cdot\Delta T_{ox,j}}$ is also approximated as a lognormal, and the computational complexity of performing this calculation is $O(s)$.

Now that each $I^{sum}_{sub,p,l}$ is approximated as a lognormal random variable, the full-chip leakage can be calculated as the sum

$$\sum_{p=1}^{N_{models}} \sum_{l=1}^{n} I^{sum}_{sub,p,l}, \tag{23}$$

where N_{models} is the total number of models in the library, and n is the number of grid partitions in the spatial correlation model. Note that any two $I^{sum}_{sub,p,l}$ terms may be correlated due to spatial correlations of the process parameter L_{eff}, and thus the computational complexity of the sum is $O(N^2_{models} \cdot n^2)$. Since the number of different models of a library is upper-bounded by a constant, and the number of grids is substantially smaller than the number of gates in the circuit, the computational complexity for estimating the distribution of full-chip leakage current is reduced from $O(N_g^2)$ to a substantially smaller number $O(N^2_{models} \cdot n^2)$.

4.3.3 *Handling Correlations Between Leakage Currents in Different Groups.* As described in the previous subsection, in order to reduce the number of correlated lognormals to sum, the leakage current terms are summed in groups, where each group is a set of terms that correspond to the same grid and the same model from the LUT. Let $I_{p1,l}^{sum}$ and $I_{p2,l}^{sum}$ be the results of two grouped sums that are both in the same l^{th} grid, and utilizing models $p1$ and $p2$ from the LUT, respectively. According to Equation (22), they can be computed as:

$$I_{p1,l}^{sum} = e^{U_{0,p1}+\beta_{1,p1}\cdot\Delta L_{eff}^{l}} \cdot \sum_{j=1}^{s1} \alpha_{j,p1} \cdot e^{\beta_{2,p1}\cdot\Delta T_{ox,j}} = e^{U_{0,p1}+\beta_{1,p1}\cdot\Delta L_{eff}^{l}} \cdot e^{\xi} \quad (24)$$

$$I_{p2,l}^{sum} = e^{U_{0,p2}+\beta_{1,p2}\cdot\Delta L_{eff}^{l}} \cdot \sum_{j=1}^{s2} \alpha_{j,p2} \cdot e^{\beta_{2,p2}\cdot\Delta T_{ox,j}} = e^{U_{0,p2}+\beta_{1,p2}\cdot\Delta L_{eff}^{l}} \cdot e^{\gamma} \quad (25)$$

where $s1$ and $s2$ are the number of terms in $I_{p1,l}^{sum}$ and $I_{p2,l}^{sum}$, respectively. The term e^{ξ} is the random variable approximating $\sum_{j=1}^{s1} \alpha_{j,p1} \cdot e^{\beta_{2,p1}\cdot\Delta T_{ox,j}}$, and e^{γ} for $\sum_{j=1}^{s2} \alpha_{j,p2} \cdot e^{\beta_{2,p2}\cdot\Delta T_{ox,j}}$, as described in the previous subsection.

It should be noted that $\sum_{j=1}^{s1} \alpha_{j,p1} \cdot e^{\beta_{2,p1}\cdot\Delta T_{ox,j}}$ and $\sum_{j=1}^{s2} \alpha_{j,p2} \cdot e^{\beta_{2,p2}\cdot\Delta T_{ox,j}}$ may be correlated. This is because although $I_{p1,l}^{sum}$ and $I_{p2,l}^{sum}$ correspond to different models in the LUT, they may include leakage currents of the same gate, and obviously leakage currents associated with the same transistors are correlated. Therefore, e^{ξ} and e^{γ} are correlated, and the correlation between ξ and γ must be considered while adding up the sums of all groups for full-chip leakage current calculation.

The correlation between e^{ξ} and e^{γ} can be computed by:

$$\begin{aligned} cov(e^{\xi}, e^{\gamma}) &= E(e^{\xi+\gamma}) - E(e^{\xi})E(e^{\gamma}) \\ &= e^{\mu_{\xi}+\mu_{\gamma}+(\sigma_{\xi}^2+\sigma_{\gamma}^2)/2}(e^{cov(\xi,\gamma)/2} - 1) \end{aligned} \quad (26)$$

where μ_{ξ} $[\mu_{\gamma}]$ and σ_{γ} $[\sigma_{\gamma}]$ are the mean and standard deviation of ξ $[\gamma]$, respectively.

Thus, the covariance between ξ and γ can be obtained by solving Equation (26) for $cov(\xi, \gamma)$:

$$cov(\xi, \gamma) = 2\log\left(1 + \frac{cov(e^{\xi}, e^{\gamma})}{e^{\mu_{\xi}+\mu_{\gamma}+(\sigma_{\xi}^2+\sigma_{\gamma}^2)/2}}\right) \quad (27)$$

In Equation (27), the mean and standard deviation of ξ and γ are known values. Since e^{ξ} and e^{γ} are approximations of $\sum_{j=1}^{s1} \alpha_{j,p1} \cdot e^{\beta_{2,p1}\cdot\Delta T_{ox,j}}$ and $\sum_{j=1}^{s2} \alpha_{j,p2} \cdot e^{\beta_{2,p2}\cdot\Delta T_{ox,j}}$, respectively, the value of $cov(e^{\xi}, e^{\gamma})$ can be obtained as:

$$cov(e^{\xi}, e^{\gamma}) = cov\left(\sum_{j=1}^{s1} \alpha_{j,p1} \cdot e^{\beta_{2,p1}\cdot\Delta T_{ox,j}}, \sum_{j=1}^{s2} \alpha_{j,p2} \cdot e^{\beta_{2,p2}\cdot\Delta T_{ox,j}}\right) \quad (28)$$

Note that any two $\Delta T_{ox,j}$ variables are independent, and thus the value of the above right hand side can easily be computed as:

$$\sum_{j} \alpha_{j,p1} \cdot \alpha_{j,p2} \cdot e^{(\beta_{2,p1}^2+\beta_{2,p2}^2)\sigma_{T_{ox,j}}^2/2} \cdot (e^{\beta_{2,p1}\cdot\beta_{2,p2}\cdot\sigma_{T_{ox,j}}^2} - 1) \quad (29)$$

where $\sigma_{T_{ox,j}}$ is the standard deviation of $\Delta T_{ox,j}$.

4.3.4 *Handling Inter-die Variations.* The described framework for statistical computation of full-chip leakage considering spatial correlations in intra-die variations of process parameters can easily be extended to handle inter-die variations. To include the effects of inter-die variations, for each type of process parameter, a global random variable can be applied to all gates in the circuit to model this effect. For spatially correlated process parameters, this is reflected as an update of the covariance matrix by adding to all entries the variance of the global random variable. For spatially uncorrelated process parameters, it introduces a correlation term between the leakage currents of different gates. However, the same framework of estimating the distribution of full-chip leakage current for handling intra-die variations proposed in section 4 can be applied.

5. AN IMPROVED ALGORITHM, HYBRIDIZED WITH THE PCA-BASED APPROACH

In previous sections, we have proposed techniques to improve the computational complexity by reducing the number of correlated lognormals to sum. Another possible approach is to modify the structure of each lognormal random variable so that the summation can be computed efficiently, as was done using a PCA-based (Principal Component Analysis) method in the work of [Srivastava et al. 2005]. In this section, we will first present the method proposed in [Srivastava et al. 2005], and an improved method hybridized with the PCA-based approach will be proposed in the following section.

5.1 PCA-based Method

The work of [Srivastava et al. 2005] proposes a PCA-based method to compute the full-chip leakage considering the effect of spatial correlations of L_{eff}. The principle of the method is very similar to the PCA-based statistical timing analysis proposed in [Chang and Sapatnekar 2003]. In this method, the same spatial correlation model introduced in section 3 is used. The leakage current of each gate is approximated by a lognormal random variable in a form similar to expression (7) or (11)[4], call it the "original lognormal form" for later reference, and then the expression is rewritten in a "PCA form" by expanding the variable ΔL_{eff} as a linear function of principal components. For example, let I_{sub}^{i} be the subthreshold leakage current of the i^{th} gate originally written in a form similar to Equation (7) as:

$$I_{sub}^{i} = e^{U_{0,i}+\beta_{1,i}\cdot\Delta L_{eff}^{l}+\beta_{2,i}\cdot\Delta T_{ox,i}} \tag{30}$$

Here, ΔL_{eff}^{l} is the random variable for the variation of L_{eff} in the l^{th} grid, and $\Delta T_{ox,i}$ is the variation of T_{ox} at the i^{th} gate. Note that for any $i \neq j$, $\Delta T_{ox,i}$ and $\Delta T_{ox,j}$ are independent since T_{ox} is spatially uncorrelated.

If principal component analysis is performed on the set of correlated variables $\Delta L_{eff}^{1}, \cdots, \Delta L_{eff}^{n}$, as in [Chang and Sapatnekar 2003], then ΔL_{eff}^{l} can be expressed as a linear function of the set of principal components:

$$\Delta L_{eff}^{l} = a_{l1} \times L_{eff}^{'1} + \cdots + a_{lN_p} \times L_{eff}^{'N_p} \tag{31}$$

[4] In [Srivastava et al. 2005], only process parameter L_{eff} is considered and an independent uncertainty term is introduced for ΔL_{eff}. For convenience, we do not distinguish such differences, since these factors can easily be considered and incorporated in any framework.

where the L'^{j}_{eff} variables are the mutually independent principal components computed from the covariance matrix of $\Delta L^1_{eff}, \cdots, \Delta L^n_{eff}$, the coefficients a_{lj} of each L'^{j}_{eff} are computed from principal component analysis, and N_p is the number of principal components.

Then, the PCA form of I^i_{sub} is:

$$I^i_{sub} \;=\; e^{U_{0,i}+\sum_{t=1}^{N_p} k^i_t \cdot L'^{t}_{eff}+\beta_{2,i}\cdot \Delta T_{ox,i}} \tag{32}$$

where each $k^i_t = a_{l1}\cdot\beta_{1,i}$ can be computed by comparing this equation with Equation (31).

In [Srivastava et al. 2005], the sum $I^i_{sub} + I^j_{sub}$ is re-approximated again by a lognormal random variable I^h_{sub} in PCA form:

$$I^h_{sub} = e^{U_{0,h}+\sum_{t=1}^{N_p} k^h_t \cdot L'^{t}_{eff}+\beta^h_r \cdot r} \tag{33}$$

where r is a normalized Gaussian random variable generated by merging the two terms ΔT^i_{ox} and ΔT^j_{ox}, and β^h_r is the coefficient of r.

In Equation (33), the value of $U^{0,h}$ can be directly computed using Wilkinson's formula (19). The other coefficients can be obtained using the following expressions:

$$\begin{aligned} k^h_t \;&=\; log\frac{E(I^i_{sub}\cdot e^{L'^{t}_{eff}}) + E(I^j_{sub}\cdot e^{L'^{t}_{eff}})}{[E(I^i_{sub}) + E(I^j_{sub})]E(e^{L'^{t}_{eff}})} \\ \beta^h_r \;&=\; \left[log\left(1+\frac{Var(I^i_{sub})+Var(I^j_{sub})+2cov(I^i_{sub},I^j_{sub})}{(I^i_{sub}+I^j_{sub})^2}\right)-\sum_{t=1}^{N_p}(k^h_t)^2\right]^{0.5} \end{aligned} \tag{34}$$

Here, $E(.)$, $Var(.)$ and $cov(I^i_{sub}, I^j_{sub})$ can be computed using Equations (14), (15), and (16). Note that all terms in Equation (34) are in PCA form. The benefit of using a PCA form is that the mean and variance of a lognormal random variable can be computed in $O(N_p)$, as can the covariance of two lognormal random variables in PCA form. Therefore, the computation of all values and coefficients in I^h_{sub}, and thus the sum of two lognormals in PCA form, can be computed in $O(N_p)$. As mentioned in the description of Wilkinson's method, the computation of full-chip leakage current distribution requires a summation of N_g correlated lognormals. Thus, the PCA-based method has an overall computational complexity of $O(N_p \cdot N_g)$.

5.2 Hybridization with the PCA-based Approach

In this section, we will present an improved algorithm by hybridizing the basic approach proposed in section 4 with the PCA-based method in [Srivastava et al. 2005].

We summarize the similarities and differences between the basic approach and the PCA-based method as follows. Both methods use Wilkinson's method to approximate sum of lognormal random variables. The basic approach in section 4 improves run-time by reducing the number of correlated lognormals to sum, by first calculating the sum of leakage currents by groups, where each group contains leakage currents in the same grid and using the same LUT model, and then computing full-chip leakage by summing up leakage currents in all groups. The computational complexity of this approach is $O(n^2 \cdot N_{models}^2)$, where n is the number of grids partitioned in the spatial correlation model and N_{model} is the number of models in the LUT. The PCA-based method re-expresses each lognormal random variable in PCA form, and then directly computes the summation of all correlated lognormals using Wilkinson's method in $O(N_g \cdot N_p)$, where N_g is the total number of gates in the circuit and N_p is the number of principal components.

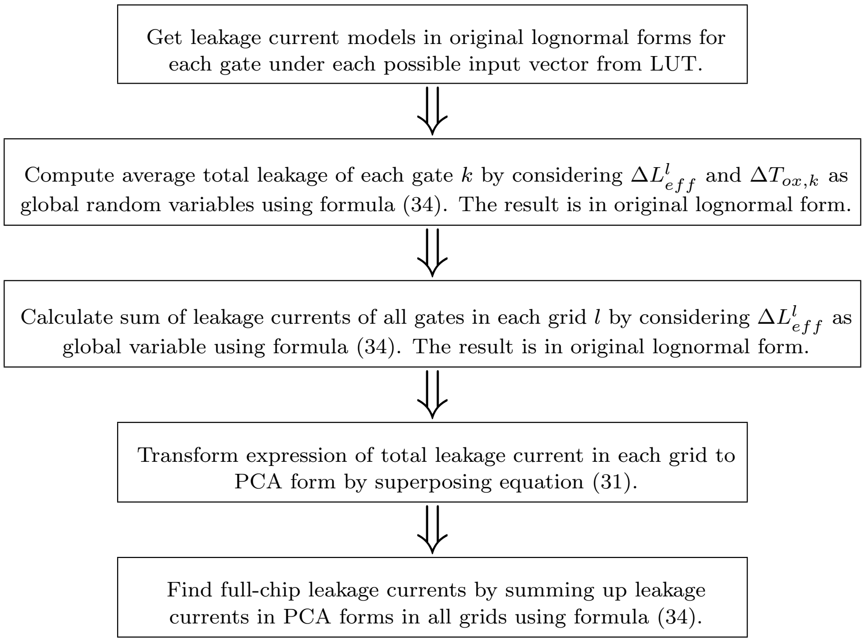
Similar to the basic approach, the improved algorithm proposed will compute the full-chip leakage current hierarchically in groups, and the sum of leakage current terms in each group will be computed in a more efficient way as in the PCA-based approach:

First, the average total leakage current of each gate can be computed as defined in Equation (1) and (4): $\sum_{\forall vec_{i,k}} Prob(vec_{i,k}) \cdot (I_{sub,k}(vec_{i,k}) + I_{gate,k}(vec_{i,k}))$. By using the models from the LUT, the average total leakage current becomes a weighted sum of several leakage current terms, and note that the number of the terms is no more than N_{models}. In general, if the gate is located in the l^{th} grid, then any leakage current term can be written in the form $e^{U_0+\beta_1 \cdot \Delta L_{eff}^l + \beta_2 \cdot \Delta T_{ox,k}}$. If we re-approximate the sum of any two leakage current terms in the same form, Equation (34) can be utilized to compute the desired values in the approximation. This is because the process parameters of all transistors in the same gate are fully correlated, so that ΔL_{eff}^l and $\Delta T_{ox,k}$ can be regarded as global random variables in the same gate. Thus, Equation (34) can easily be reused by first normalizing ΔL_{eff}^l and $\Delta T_{ox,k}$ to unit Gaussians in the original lognormal form, and then computing the sum using (34) by regarding the normalized ΔL_{eff}^l and $\Delta T_{ox,k}$ as principal components in the formula. Obviously, the complexity for summing any two leakage current terms in the same gate is $O(1)$, and thus the computation of the average total leakage current of a gate is $O(N_{models})$. If the total number of gates in the circuit is N_g, then the computational complexity of this step is $O(N_{models} \cdot N_g)$.

Next, the total leakage current in each grid is computed separately. Clearly, for all gates in the l^{th} grid, any average leakage current of a gate is expressed as an exponential function of the same random variable ΔL_{eff}^l, while the average leakage current terms for different gates correspond to different $\Delta T_{ox,k}$ variables (note that all $\Delta T_{ox,k}$ variables are independent). The sum of average leakage currents of any two gates can be approximated in a manner similar to that used in computing the average leakage current of a single gate, by first normalizing ΔL_{eff}^l to unit Gaussian in original lognormal forms of leakage currents, and then computing the sum using formula (34) by regarding normalized ΔL_{eff}^l as a principal component. Therefore, the sum has a computational complexity of $O(1)$. Since this step must compute the total leakage current of all gates in all grids, the computation complexity is $O(N_g)$.

Finally, the full-chip leakage is computed by adding up the total leakage currents computed in all grids. If the number of grids is n, then n correlated lognormals, with a complicated correlation structure, must be summed up. Therefore, we transform all lognormals in the summation into PCA forms, and the sum can be computed using the same method proposed in [Srivastava et al. 2005]. The computation complexity of this step is $O(N_p \cdot n)$.

From the analysis above, the total computational complexity of the improved



Get leakage current models in original lognormal forms for each gate under each possible input vector from LUT.

Compute average total leakage of each gate k by considering ΔL_{eff}^{l} and $\Delta T_{ox,k}$ as global random variables using formula (34). The result is in original lognormal form.

Calculate sum of leakage currents of all gates in each grid l by considering ΔL_{eff}^{l} as global variable using formula (34). The result is in original lognormal form.

Transform expression of total leakage current in each grid to PCA form by superposing equation (31).

Find full-chip leakage currents by summing up leakage currents in PCA forms in all grids using formula (34).

Fig. 5. Overall flow of the improved algorithm.

Table I. Comparison of the proposed basic method with Monte Carlo simulation.

Circuit	Gate	Grid	Total Circuit Leakage Current (μA)									
			MC		Basic Method		Error%		MCNoCorr		Error%	
Name	Number	Number	mean	std	mean	std	mean	std	mean	std	mean	std
s38584	20705	256	678.4	215.5	670.7	208.9	-1.1%	-3.1%	678.2	199.3	0.0%	-7.5%
s35932	17793	256	606.2	195.0	597.1	182.9	-1.5%	-6.2%	606.1	165.3	0.0%	-15.2%
s15850	10369	256	420.1	132.2	414.3	128.6	-1.4%	-2.7%	419.7	120.1	-0.1%	-9.2%
s13207	8620	256	352.6	110.0	349.9	108.4	-0.8%	-1.5%	350.9	97.5	-0.5%	-11.4%
s9234	5825	64	239.0	76.8	237.0	75.2	-0.8%	-2.1%	240.1	70.0	0.5%	-8.9%
s5378	2958	64	178.9	59.9	177.1	59.3	-1.0%	-1.0%	178.9	56.4	0.0%	-5.8%
c7552	5528	64	327.9	106.1	324.3	101.0	-1.1%	-4.8%	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%	0.9%	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%	-3.2%	113.9	31.7	0.2%	-16.1%
c1908	1261	16	73.5	24.9	72.3	23.5	-1.6%	-5.6%	73.2	20.1	-0.4%	-19.3%
c880	594	4	37.4	13.3	36.9	12.7	-1.3%	-4.5%	37.3	10.5	-0.3%	-21.1%
c432	294	4	18.3	6.5	17.9	6.2	-2.2%	-4.6%	18.2	5.1	-0.5%	-21.5%

algorithm is $O(N_p \cdot n + (N_{models} + 1) \cdot N_g) = O(N_p \cdot n + N_g)$. This is better than the complexity of $O(N_g \cdot N_p)$ for the PCA-based method, since the number of grids n is substantially smaller than the number of gates N_g in the circuit. If n is a small constant, the basic approach which has a computational complexity of $O(n^2 \cdot N_{models}^2)$ which may outperform the improved approach. However, as n grows to a relatively larger number, the basic approach grows quadratically with n, while improved approach grows linearly which results in a better run-time for the improved approach, as compared to the basic method. The flow of the improved algorithm is provided in Figure 5.

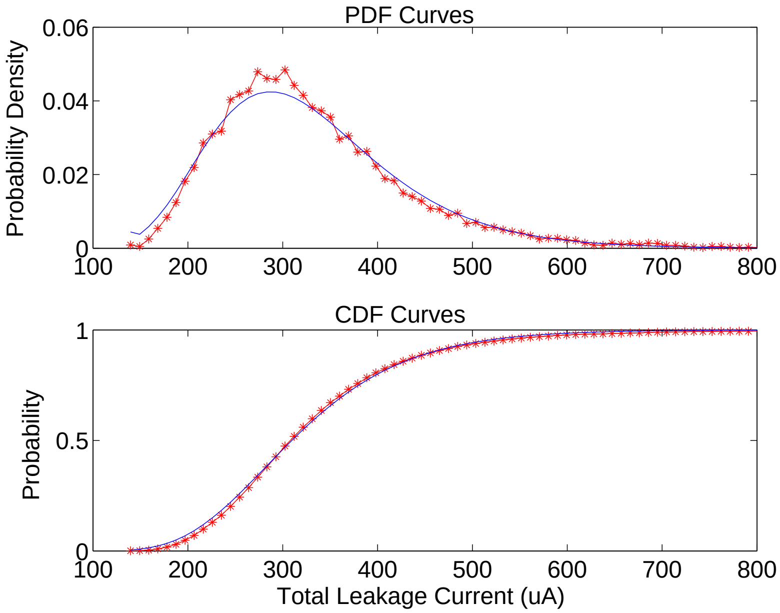


Fig. 6. Distributions of the total leakage using the proposed basic method against Monte Carlo simulation method for circuit c7552. The solid line illustrates the result of the proposed basic method, while the starred line shows the Monte Carlo simulation results.

6. EXPERIMENTAL RESULTS

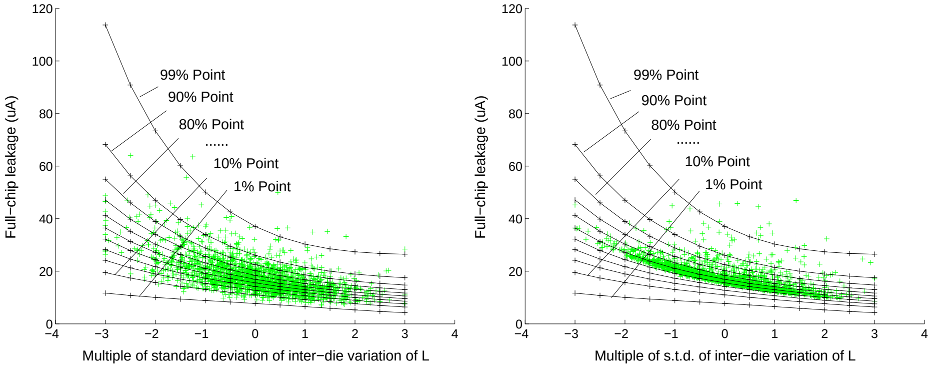
In this section, the experimental results for full-chip statistical leakage estimation will be presented. The results using the basic approach proposed in section 4 will be first provided, followed by those using the improved method of section 5. A study of effects of process parameter L_{eff} and T_{ox} on subthreshold and gate tunneling leakage currents is also provided at the end of this section.

Our experiments were performed on the set of circuits in the ISCAS85 and ISCAS89 benchmark set. The circuits were synthesized with SIS with a 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 [CAP]. The technology parameters that were used correspond to the 100 nm Berkeley Predictive Technology model [BPT], and the 3σ value of parameter variations for L_{eff} 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. The number of grid partitions in the spatial correlation model used for each benchmarks is listed in Table I, and depends on the size of the circuit.

6.1 Results of the Basic Method

First, we present the experimental results of the proposed basic method for full-chip leakage estimation introduced in section 4. For comparison purposes, we performed Monte Carlo simulations with 10,000 runs on the benchmarks. The results of the comparison of this method with the Monte Carlo (*MC*) approach are shown in Table I. The average errors for the mean and sigma values are 1.2% and 3.6%, respectively. In Figure 6, we show the distribution of total circuit leakage current achieved using the proposed basic method and using Monte Carlo simulation for circuit c7552: it is easy to see that the curve achieved by the basic method matches well with the Monte Carlo simulation result. For all testcases, the run-time of the basic method is less than one second, while the Monte Carlo simulation takes considerably longer: for the largest test case, c7552, this 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 intra-die variations of effective gate length L_{eff} of any two gates on the chip. The comparison data is also shown in Table I. It can be observed that although the mean values are close, on average, the variances of *MCNoCorr*, where spatial correlations are ignored, has a 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.



(a) Considering spatial correlation

(b) Ignoring spatial correlation

Fig. 7. Comparison of scatter plots of full-chip leakage of circuit c432 considering and ignoring spatial correlation.

To visualize the difference, in Figure 7, for circuit c432, we show the scatter plots for 2000 samples of full-chip leakage current generated by Monte Carlo simulations, with and without consideration of spatial correlations of L_{eff}. The x-axis marks the multiples of the standard deviation value of ΔL_{eff}^{inter}, inter-die variations of effective gate length, ranging from -3 to $+3$, since a Gaussian distribution is assumed. The y-axis are the values of total circuit leakage current. Therefore, at each specific value of ΔL_{eff}^{inter}, the scatter points list the various sampled values of total circuit leakage current due to variations in T_{ox} and intra-die variation of L_{eff}. The plots also show a set of contours lines that correspond to, with the effect of spatial correlation taken into account, a set of percentage points of the cumulative density function (CDF) of total circuit leakage current at different values of ΔL_{eff}^{inter}. In Figure 7(a), 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 7(b), 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 $\Delta L_{eff}^{inter} \leq -1\sigma$; the same ratio is 1.1 times larger otherwise. Looking at the same numbers in a different way, in Figure 7(b), all points are contained between the 30% and 80% contours if $\Delta L_{eff}^{inter} \leq -1\sigma$. In this range, I_{sub} is greater than I_{gate} by one order of magnitude on average, and thus the variation of L_{eff} can have a large effect on the total leakage as I_{sub} is exponentially dependent on L_{eff}. Consequently, ignoring spatial correlation results in a substantial underestimation of the standard deviation, and thus the worst-case full-chip leakage. For $\Delta L_{eff}^{inter} > -1\sigma$, I_{sub} decreases to a value comparable to I_{gate} and L_{eff} 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, a large number of remaining points show smaller variations and are within the 20% and 90% contours, due to the same reasoning given above for $\Delta L_{eff}^{inter} \leq -1\sigma$.

6.2 Improved Method

Table II. Comparisons of the basic, PCA and improved methods with Monte Carlo simulation.

Circuit	Total Circuit Leakage Current (μA)											
	MC		Basic		Error%		PCA		Error%		Improved		Error%	
Name	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
s38584	678.4	215.5	670.7	208.9	-1.1%	-3.2%	670.7	215.9	-1.1%	0.2%	672.0	209.5	-0.9%	-2.8%
s35932	606.2	195.0	597.1	182.9	-1.5%	-6.6%	597.1	183.6	-1.5%	-5.8%	597.6	195.0	-1.4%	0.0%
s15850	420.1	132.2	414.3	128.6	-1.4%	-2.8%	415.0	129.0	-1.2%	-2.4%	420.1	132.2	0.0%	0.0%
s13207	352.6	110.0	349.9	108.4	-0.8%	-1.5%	350.0	112.2	-0.7%	2.0%	350.5	108.8	-0.6%	-1.1%
s9234	239.0	76.8	237.0	75.2	-0.8%	-2.1%	237.0	77.7	-0.8%	1.2%	237.6	75.5	-0.6%	-1.7%
s5378	178.9	59.9	177.1	59.3	-1.0%	-1.0%	177.1	61.1	-1.0%	2.0%	179.1	59.8	0.1%	-0.2%
c7552	327.9	106.1	324.3	101.0	-1.1%	-5.0%	324.3	104.2	-1.1%	-1.8%	325.6	101.5	-0.7%	-4.3%
c5315	239.0	78.4	235.7	74.3	-1.4%	-5.5%	235.7	76.6	-1.4%	-2.3%	237.7	74.9	-0.5%	-4.5%
c6288	229.6	77.3	227.7	78.0	-0.8%	0.9%	227.7	80.3	-0.8%	3.9%	227.9	78.0	-0.7%	0.9%
c3540	158.9	53.4	156.8	50.9	-1.3%	-4.9%	156.8	52.5	-1.3%	-1.7%	157.8	51.4	-0.7%	-3.7%
c2670	113.7	37.8	112.6	36.6	-1.0%	-3.3%	112.6	37.7	-1.0%	-0.3%	117.0	37.6	2.9%	-0.5%
c1908	73.5	24.9	72.3	23.5	-1.6%	-6.0%	72.3	24.2	-1.6%	-2.8%	72.5	23.6	-1.4%	-5.2%
c880	37.4	13.3	36.9	12.7	-1.3%	-4.7%	36.9	13.1	-1.3%	-1.5%	37.0	12.8	-1.1%	-3.8%
c432	18.3	6.5	17.9	6.2	-2.2%	-4.8%	17.9	6.4	-2.2%	-1.5%	18.0	6.2	-1.6%	-4.6%

In this section, we present the experimental results using the improved algorithm, by comparing its accuracy and run-time efficiency with those of the basic and PCA methods.

Table II lists the results generated using the basic, PCA and improved methods. As shown in the Table, the results of the three methods are not exactly the same, since the order in summing leakage terms are not all the same in these methods. However, as these approaches are all based on Wilkinson's approximation, similar accuracies for estimating total chip leakage are achieved: the average errors for the mean and sigma values are 1.2% and 3.6% respectively for the basic method, 1.2% and 2.1% the PCA method, and 1.0% and 2.4% the improved method.

However, the three methods differ in terms of run-time efficiencies. In Table III and IV, we show the run-times for different methods for ISCAS85 and ISCAS89 benchmark sets, respectively. In general, the proposed basic method is about 3 to

4 times faster than the PCA-based method. As expected, the proposed improved approach does not show any run-time advantage over the basic method for smaller grid sizes. However, run-time of both the proposed basic and the PCA-based methods grows much faster with the grid size than the improved method. In Table III and IV, when the number of grids grows to greater than 64, the improved approach is about 100 times faster than the other approaches. Therefore, the run-time can be significantly improved by combining the PCA-based with the proposed basic leakage estimation approach.

Table III. Run-time comparison of the proposed basic, PCA-based, and improved methods for the ISCAS85 benchmarks

Benchmark	c432	c880	c1908	c2670	c3540	c6288	c5315	c7552
Number of grids	4	4	16	16	16	16	64	64
Proposed basic method (s)	0.01	0.02	0.04	0.06	0.09	0.10	0.24	0.29
PCA-based method (s)	0.03	0.06	0.18	0.27	0.40	0.57	1.43	1.82
Proposed improved method (s)	0.01	0.03	0.06	0.09	0.12	0.14	0.19	0.25

Table IV. Run-time comparison of the proposed basic, PCA-based, and improved methods for the ISCAS89 benchmarks

Benchmark	s5378	s9234	s13207	s15850	s35932	s38584
Number of grids	64	64	256	256	256	256
Proposed basic method (s)	0.22	0.32	5.89	5.91	4.97	10.04
PCA-based method (s)	0.93	1.62	7.58	8.97	17.38	24.28
Proposed improved method (s)	0.16	0.30	0.47	0.56	1.03	1.34

6.3 Effects of L_{eff} and T_{ox} on Leakage Currents

We also study the effect by varying L_{eff} and T_{ox} separately on the variations of full-chip subthreshold and gate-tunneling leakage currents. As the purpose of this test is purely for studying the effects of process variations, Monte Carlo simulations are used in the tests. In Table V, the results by varying L_{eff} only keeping T_{ox} at its 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_{eff} and T_{ox} can each individually lead to substantial variations in the full-chip leakage. When only L_{eff} 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_{eff} than T_{ox}, and I_{gate} is a strong exponential function of T_{ox} over L_{eff}. 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}, its standard deviation is 3 times larger than that of I_{sub}. Therefore, in this case, although I_{sub} and I_{gate} are both major contributors to the full-chip leakage, the leakage variations are mainly due to I_{gate}.

Table V. Comparison of leakage by varying L_{eff} and T_{ox} independently

Circuit Name	Leakage by varying effective gate length only (μA)						Leakage by varying gate oxide thickness only (μA)					
	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

7. CONCLUSIONS

In this paper, we have proposed methods for computing the distribution of total circuit leakage power under process parameter variations considering the spatial correlations among parameters. Two approaches, the basic and the improved methods, have been described, with the latter as an extension of the basic approach hybridized with an idea in [Srivastava et al. 2005] that improved the computational complexity to a linear dependency on the number of grids in the intra-die spatial correlation model. The proposed methods have been shown to be effective in predicting the mean, standard deviation and the PDF of the total chip leakage. We have also shown that the spatial correlations of process parameters must be considered appropriately in order to predict chip yield correctly. We believe that the proposed frameworks are general enough to predict the total 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, the methods can easily be extended to capture the effects of temperature variations. The limitation of the proposed methods is the approximation of leakage current as lognormal distributions through first order expansions. Our experiments show the validity of this assumption for 20% variations, which is adequate for near-term technologies. However, as the process variations grow larger in the longer term, this approximation can introduce larger errors. Developing techniques for handling these larger process variations is a topic that we propose to explore in the future.

Acknowledgments

The authors would like to thank the anonymous reviewers, whose comments provided excellent feedback, resulting in an improvement in the quality of our paper.

REFERENCES

Berkeley predictive technology model (BPTM). Available at: http://www.eas.asu.edu/~ptm.

Capo: A large-scale fixed-die placer from UCLA. Available at: http://vlsicad.ucsd.edu/GSRC/bookshelf/Slots/Placement.

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

Acar, E., Devgan, A., Rao, R., Liu, Y., Su, H., Nassif, S., and Burns, J. 2003. Leakage and leakage sensitivity computation for combinational circuits. In *Proceedings of the International Symposium on Low Power Electronics and Design.* Seoul, Korea, 96 – 99.

Agarwal, A., Blaauw, D., Zolotov, V., Sundareswaran, S., Zhao, M., Gala, K., and Panda, R. 2003. Statistical delay computation considering spatial correlations. In *Proceedings of the Asia and South Pacific Design Automation Conference*. Kitakyushu, Japan, 271–276.

Bowman, K. A., Wang, L., Tang, X., and Meindl, J. D. 2001. A circuit level perspective of the optimum gate oxide thickness. *IEEE Transction on Electron Devices 48,* 8 (Aug.), 1800 – 1810.

Chang, H. and Sapatnekar, S. S. 2003. Statistical timing analysis considering spatial correlations using a single PERT-like traversal. In *Proceedings of the IEEE/ACM International Conference on Computer Aided Design*. San Jose, California, USA, 621–625.

Chang, H. and Sapatnekar, S. S. 2005. Full-chip analysis of leakage power under process variations, including spatial correlations. In *Proceedings of the ACM/IEEE Design Automation Conference*. Anaheim, California, USA, 523–528.

Friedberg, P., Cao, Y., Cain, J., Wang, R., Rabaey, J., and Spanos, C. 2005. Modeling within-die spatial correlation effects for process-design co-optimization. In *Proceedings of International Society for Quality Electronic Design*. San Jose, CA, USA, 516 – 521.

Ketkar, M. and Sapatnekar, S. S. 2002. Standby power optimization via transistor sizing and dual threshold voltage assignment. In *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*. San Jose, California, USA, 375 – 378.

Lee, D., Kwong, W., Blaauw, D., and Sylvester, D. 2003. Analysis and minimization techniques for total leakage considering gate oxide leakage. In *Proceedings of the ACM/IEEE Design Automation Conference*. Anaheim, California, USA, 175–180.

Mukhopadhyay, S. and Roy, K. 2003. Modeling and estimation of total leakage current in nano-scaled CMOS devices considering the effect of parameter variation. In *International Symposium on Low Power Electronics and Design*. Seoul, Korea, 172–175.

Najm, F. N. 1994. A survey of power estimation techniques in VLSI circuits. *IEEE Transactions on Very Large Scale Integration Systems 2,* 4 (Dec.), 446–455.

Narendra, S., De, V., Borkar, S., Antoniadis, D., and Chandrakasan, A. 2002. Full-chip sub-threshold leakage power prediction model for sub-0.18μm CMOS. In *Proceedings of the International Symposium on Low Power Electronics and Design*. Monterey, California, USA, 19–23.

Papoulis, A. and Pillai, S. U. 2002. *Probability, random variables, and stochastic processes.* McGraw-Hill, Boston, USA.

Rao, R., Devgan, A., Blaauw, D., and Sylvester, D. 2004. Parametric yield estimation considering leakage variability. In *Proceedings of Design Automation Conference*. San Diego, California, USA, 442 – 447.

Rao, R., Srivastava, A., Blaauw, D., and Sylvester, D. 2003. Statistical estimation of leakage current considering inter- and intra-die process variation. In *Proceedings of the International Symposium on Low Power Electronics and Design*. Seoul, Korea, 84–89.

Semiconductor Industry Association 1997-2005. *International Technology Roadmap for Semiconductors.* Semiconductor Industry Association.

Sirichotiyakul, S., Edwards, T., Oh, C., Zuo, J., Dharchoudhury, A., Panda, R., and Blaauw, D. 1999. Stand-by power minimization through simultaneous threshold voltage selection and circuit sizing. In *Proceedings of the IEEE/ACM Design Automation Conference*. New Orleans, Louisiana, USA, 436–441.

Srivastava, A., Bai, R., Blaauw, D., and Sylvester, D. 2002. Modeling and analysis of leakage power considering within-die process variations. In *Proceedings of the International Symposium on Low Power Electronics and Design*. Monterey, California, USA, 64–67.

Srivastava, A., Shah, S., Agarwal, K., Sylvester, D., Blaauw, D., and Director, S. W. 2005. Accurate and efficient gate-level parametric yield estimation considering correlated variations in leakage power and performance. In *Proceedings of Design Automation Conference*. Anaheim, California, USA, 535 – 540.

Stine, B. E., Boning, D. S., and Chung, J. E. 1997. Analysis and decomposition of spatial variation in integrated circuit processes and devices. *IEEE Transaction on Semiconductor Manufacturing 10,* 1 (Feb.), 24–41.

Sultania, A., Sylvester, D., and Sapatnekar, S. S. 2004. Tradeoffs between gate oxide leakage and delay for dual Tox circuits. In *Proceedings of Design Automation Conference*. San Diego, California, USA, 761 – 766.

Taur, Y. and Ning, T. H. 1998. *Fundamentals of Modern VLSI Devices*. Cambridge University Press.

Xiong, J., Zolotov, V., and He, L. 2006. Robust extraction of spatial correlation. In *Proceedings of ACM International Symposium on Physical Design*. San Jose, CA, USA, 2 – 9.

Received Feb. 2006; revised Aug. 2006; accepted Oct 2006.