A Statistical Framework for Estimation of Full-Chip Leakage-Power Distribution Under Parameter Variations

Hamed F. Dadgour, *Student Member, IEEE*, Sheng-Chih Lin, *Student Member, IEEE*, and Kaustav Banerjee, *Senior Member, IEEE*

***Abstract*—This paper presents a novel framework for accurate estimation of key statistical parameters of the subthreshold- and gate-leakage distributions of a chip under parameter variations while considering both within-die and die-to-die variabilities in process (P), temperature (T), and supply voltage (V). For the first time, temperature variations and, more importantly, electrothermal couplings between junction (substrate or die) temperature and leakage power have been accounted for in a full-chip leakage estimation methodology. In the proposed framework, instead of exact leakage distribution profile, its statistically important parameters, such as nominal value and spread, are computed. Initially, at the transistor level, a quantitative analysis of the relative sensitivities of device leakage components to P–T–V variations is performed to extract a transistor-level variation model. It is shown that the proposed statistical model, as compared to others in the literature, shows better agreement with BSIM[1] model-based simulations. It is also demonstrated that failing to account for temperature variations and electrothermal couplings can result in significant inaccuracy in chip-level leakage estimation. Furthermore, the full-chip leakage-power distribution is used to estimate the leakage-constrained yield under the impact of variations. The calculations show that yield is significantly lowered due to the within-die and die-to-die process and temperature variations. Subsequently, the proposed framework is applied in the leakage estimation of complex logic circuits with a consideration of spatial correlations of process parameters and transistor stacking effects.**

***Index Terms*—CMOS devices, leakage currents, parameter variations, power dissipation, temperature variations, within-die variations, yield estimation.**

I. Introduction

A. Parameter Variations in Nanometer-Scale CMOS Designs

PARAMETER variation makes device and interconnect parameters appear as statistical variables instead of deterministic values. Parameter fluctuations can result from different types of sources such as process (P), supply voltage (V), and temperature (T) variations. Parameter variations are often categorized as within-die and die-to-die variations. Die-to-die variation refers to the situation where a design parameter varies from one die to another, while within-die variation occurs when a design parameter is not constant across a single die. While die-to-die variation was dominant in the past, for sub-100-nm technologies, the impact of within-die variation has become particularly important. This is primarily because of the increasing gap between the minimum feature size of CMOS devices and the lithography wavelength used for their fabrication, as shown in Fig. 1(a), leading to increased variations in lithography-dependent parameters such as channel length [1]. Moreover, it should be noted that due to threshold-voltage rolloff phenomenon, the threshold-voltage deviation $(\Delta V_{\rm th})$ due to a constant channel length variation (ΔL) is higher for smaller devices [Fig. 1(b)]. As a result, the impact of channel length variation becomes more severe in scaled technologies. In general, process variations can arise because of photography proximity effects, mask/lens/photo system deviations, plasma etch dependences, or even chemical–mechanical polishing used for planarization of various layers at the backend, including metal (Cu) layers.

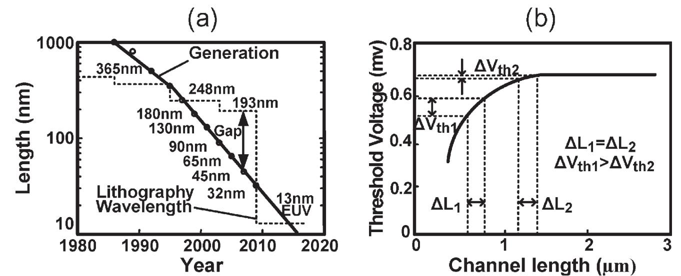

Fig. 1. (a) Difference between the minimum feature size and the lithography wavelength increases in future CMOS technology generations [1]. (b) Schematic showing the impact of threshold-voltage rolloff on $V_{\rm th}$ variation. With technology scaling, the same amount of channel length variation $(\Delta L_1 = \Delta L_2)$ results in greater variations in the threshold voltage $(\Delta V_{\rm th1} > \Delta V_{\rm th2})$.

Parameter variations can also be categorized as either correlated or uncorrelated (or random). Correlated variations refer to situations where parameters of two identical devices vary depending on their spatial proximity on the die, while uncorrelated variations imply that corresponding parameters for two identical devices are statistically independent, irrespective of their location on the die (such as $V_{\rm th}$ variation due to random dopant fluctuations or line-edge roughness). It should be noted that the impact of random variations is more dominant

Manuscript received March 12, 2007; revised July 16, 2007. This work was supported in part by a grant from Intel Corporation, by the University of California-MICRO Program (03-004), and by the National Science Foundation (Award CCF-0541465). The review of this paper was arranged by Editor C.-Y. Lu.

The authors are with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 USA (e-mail: hamed@ece.ucsb.edu; sclin@ece.ucsb.edu; kaustav@ece.ucsb.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TED.2007.906960

[1]Berkeley Short-channel IGFET Model.

0018-9383/$25.00 © 2007 IEEE

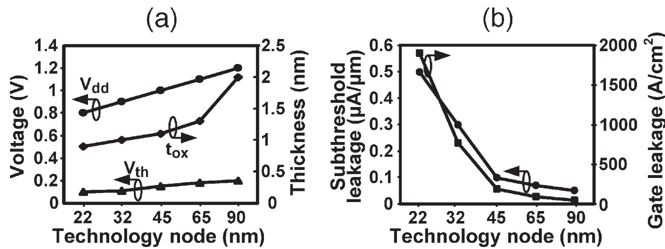

Fig. 2. (a) Supply voltage ($V_{\rm dd}$), threshold voltage ($V_{\rm th}$), and oxide thickness ($t_{\rm ox}$) scaling predicted by the ITRS [3]. (b) Subthreshold and gate leakages predicted by the ITRS for NMOS devices at different technology nodes [3].

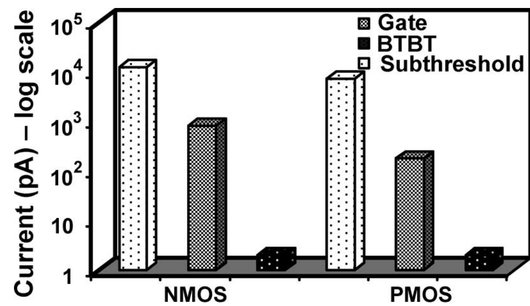

Fig. 3. Relative importance of various leakage mechanisms at the 90-nm node.

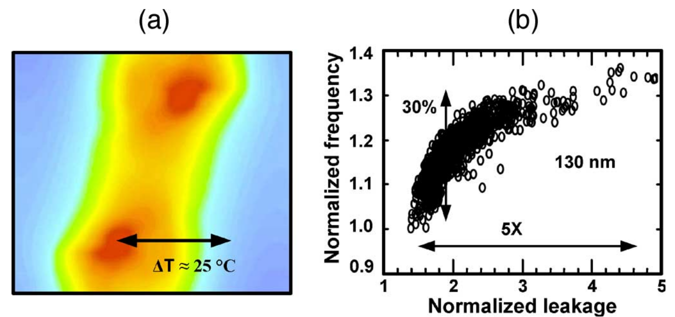

Fig. 4. (a) Thermal map of Montecito processor [4], [5]. Montecito is a 90-nm Itanium processor family microprocessor. It has two logic cores, 24 MB of L3 cache, 1.7-billion transistors, and a 100-W power envelope. (b) Variation in leakage power versus frequency for 130-nm technology (Courtesy: S. Borkar, Intel).

in minimum-sized devices. Larger W/L transistors (such as those used in this paper) are relatively insensitive to random variations. Therefore, only the correlated variations are considered in this paper.

Additionally, environmental variations such as variations in temperature or supply voltage can also be severe in nanoscaled high-performance ICs. Temperature variations arise mainly due to uneven levels of activity in different parts of a high-performance microprocessor chip, such as low-activity caches and high-speed and high-activity execution cores. In addition, sleep modes of functional blocks and low-power design techniques, such as dynamic power management and clock/power gating, can also result in temperature variations [2]. Furthermore, there could be supply voltage variations due to different sets of gate switching in different clock cycles resulting in time-varying currents on the power and ground lines, or simply due to resistive power supply lines, which give rise to Ldi/dt and IR drops, respectively. In this paper, all variations, including supply voltage, are assumed to be only due to static sources. Any dynamic or time-dependent variations are outside the scope of this paper.

B. Implications of Scaling and Parameter Variations for Leakage

Aggressive scaling trends in CMOS technology require lower threshold voltage and thinner gate oxide. Fig. 2(a) shows the ITRS-predicted scaling scenario for important circuit and device parameters such as supply voltage ($V_{\rm dd}$), threshold voltage ($V_{\rm th}$), and oxide thickness ($t_{\rm ox}$) [3]. To reduce switching (dynamic) power consumption and maintain gate-oxide reliability, supply voltage must be scaled down, and as a result, threshold voltage must also be lowered to maintain acceptable drive currents. On the other hand, thinner gate oxide is needed to maintain control of gate over the channel under short channel effects. This, however, increases direct-tunneling-induced leakage through the thin gate-oxide layer.

There are several different leakage mechanisms in CMOS circuits such as subthreshold, gate, and band-to-band-tunneling (BTBT); however, as the simulation results in Fig. 3 for NMOS and PMOS devices in a 90-nm technology suggest, subthreshold and gate leakages are the two most important components. In this figure, the Y-axis is plotted in a logarithmic scale; therefore, the BTBT leakage component is at least one order of magnitude lower than the other two mechanisms, and thus, it has been neglected in this paper.

Both the subthreshold and gate leakages dramatically increase according to the ITRS predictions, which are shown in Fig. 2(b). It should be noted that due to the exponential dependence of the subthreshold and gate leakages on threshold voltage and oxide thickness, respectively, the leakage sharply rises for sub-45-nm feature-size devices. Note that, in Fig. 2(b), the subthreshold and gate leakages are plotted using two separate Y-axes with different units; therefore, this figure does not show a comparison between the subthreshold- and gate-leakage mechanisms.

As will be discussed in the following sections, subthreshold leakage is a strong function of channel length, oxide thickness, supply voltage, and temperature. In addition, gate leakage depends on gate oxide thickness, channel length, and supply-voltage level. Therefore, variation in leakage can occur due to a fluctuation of each of these variables, which become more severe in sub-100-nm regimes. Moreover, temperature variation across microprocessor chips is an important concern even with the introduction of multicore processors. In Fig. 4(a), the temperature profile of a dual-core microprocessor called Montecito is shown [4], [5]. This microprocessor has been reported to have a temperature gradient of greater than 25 °C between (lightly shaded) cool and (dark) hot areas. Such temperature gradients lead to higher subthreshold leakage in those areas, which, in turn, increases within-die temperature variations.

Because of the combined impact of scaling and increasing parameter variations, leakage-power variation significantly increases and has been reported to have a 5× variation for a 130-nm CMOS technology, as shown in Fig. 4(b). Thus, designing with the worst case leakage values may result in excessive

guard banding, while underestimating the leakage might result in highly optimistic designs. Therefore, in the present scenario, a probabilistic modeling is more meaningful in comparison to a deterministic analysis. Additionally, due to a 5× increase of total leakage power in every generation [6], the design constraint due to the leakage power may soon limit the yield. Therefore, it is critical to develop a probabilistic framework to accurately estimate the full-chip subthreshold- and gate-leakage power distributions under the P–T–V variations, which can be subsequently used to accurately estimate the yield. Furthermore, a quantitative analysis of the relative sensitivities of leakage to the P–T–V variations is highly desirable so that relevant variations can be targeted to improve the yield.

C. Prior Work

Although existing work [7] has successfully quantified the impact of parameter variations on performance, there is no known work that has sufficiently described the combined effects of P–T–V variations on leakage power. Su *et al.* [8] estimated a full-chip leakage considering uneven voltage drop and uneven temperature, but it is not a probabilistic approach and, hence, cannot be used to estimate the yield. Rao *et al.* [9] studied the impact of channel length variations on subthreshold leakage, but their analysis is based on an empirical relationship between the leakage and the channel length. Moreover, their analysis cannot be easily extended to other variations such as oxide thickness or temperature since that would first require an empirical and invertible relationship between the leakage and the oxide thickness or the temperature. Although, the work presented in [10]–[12] developed statistical models to estimate leakage under variations, they do not account for the combined effects of the within-die and die-to-die P–T–V variations. For instance, Narendra *et al.* [11] analyzed the impact of only within-die $V_{\rm th}$ variation on subthreshold leakage, whereas Mukhopadhyay and Roy [10] analyzed the sensitivity of various leakage components to only within-die process and voltage variations. Srivastava *et al.* [12] analyzed just the impact of within-die process variations. In addition, due to the approximations involved, these analyses are inaccurate when compared to the simulations based on BSIM models, as will be discussed later in this paper. Moreover, these models cannot be used to estimate the yield since they do not provide the probability distribution function of leakage power. Recently, Agarwal *et al.* [13], Srivastava *et al.* [14], and Hongliang and Sapatnekar [15] have presented leakage estimation methodologies under variations; however, none of them has considered the impact of temperature variations on leakage or taken the strong coupling between the leakage and the temperature into consideration.

Temperature variation and electrothermal couplings between subthreshold leakage current and junction temperature were addressed by Zhang *et al.* [16]. It was shown that for accurate estimation of leakage, it is critical to consider the die-to-die temperature variations and electrothermal couplings, and if ignored, as done in previous works [10]–[15], it would result in a significant error. However, in [16], gate leakage, stacking effects, and spatial correlation of parameters are not taken into account.

TABLE I
SUMMARY OF LITERATURE SURVEY ON LEAKAGE ESTIMATION UNDER VARIATIONS

	[8]	[9]	[10]	[11]	[12]	[13]	[14]	[15]
Temperature	–	–	–	–	–	–	–	–
Supply voltage	–	–	✓	–	–	–	–	–
Process variation	✓	✓	✓	✓	✓	✓	✓	✓
Gate leakage	–	–	✓	–	✓	✓	✓	✓
Stacking effect	–	✓	–	–	–	✓	✓	✓
Spatial correlation	–	–	–	–	–	–	–	✓
Electrothermal coupling	✓	–	–	–	–	–	–	–

✓ Considered – Not considered

Table I summarizes the existing works in the literature in terms of different aspects that were included in each paper. As shown, there is no comprehensive work that considers all aspects of leakage estimation under variations. In this paper, all key aspects have been considered for accurate leakage-power estimation under parameter variations.

D. This Work

A probabilistic framework that is used to analyze the impact of both the within-die and die-to-die P–T–V variations on the subthreshold- and gate-leakage powers for sub-100-nm CMOS technologies has been introduced. For the first time, electrothermal couplings between the subthreshold leakage and the temperature have been correctly incorporated in a leakage estimation methodology. Previous works have ignored this important phenomenon, resulting in a significant estimation error.

This paper is focused on the subthreshold and gate leakages since they are the most dominant components of total leakage in high-performance ICs. The analytical models presented under this framework are compared to the detailed BSIM models and are found to be more accurate than the existing statistical models [10]–[12] (as discussed in Section III). Furthermore, the full-chip leakage-power distribution is subsequently used to estimate the leakage-constrained yield under the impact of these variations.

In this paper, there are certain assumptions that have been made to simplify the analysis. The foremost assumption is that some of the parameters that can vary, such as channel length, oxide thickness, supply voltage, and temperature, are independent. Any correlation that might exist among these parameters has been ignored. At the same time, the correlation between most of these parameters is expected to be negligible, since they are determined at different steps of a typical IC manufacturing process (for instance, oxide thickness and channel length). However, there might be strong couplings between temperature and other parameters (through leakage power), and that has been accounted for by self-consistently evaluating power and junction temperature (see Section V).

This paper is organized as follows. In Section II, a statistical framework that is used to analyze the impact of single- and

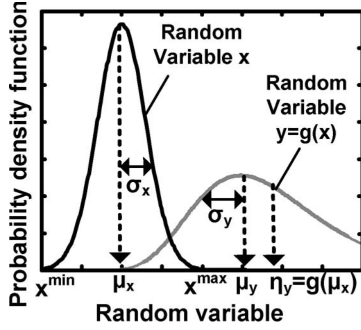

Fig. 5. Schematic indicating different statistical variables corresponding to probability density distributions of input (x) and output (y) random variables.

multiparameter variations is presented. This framework is then used to study the impact of within-die variations on the subthreshold and gate leakages in Sections III and IV, respectively. In Section V, the proposed methodology is used to estimate leakage under die-to-die variations. Section VI addresses the leakage-constrained yield estimation under the impact of variations. The new methodology is applied to estimate leakage of some complex benchmark circuits while accounting for within-die correlations and stacking effect in Section VII. Finally, concluding remarks are made in Section VIII.

II. Statistical Framework Formulation

The electrical performance of a circuit is a function of environmental factors, such as supply voltage (V_{dd}), temperature (T), etc., and physical parameters, such as channel length (L), oxide thickness (t_{ox}), etc. This dependence of electrical performance (y) can be represented by $y = g(x_1, x_2, \ldots, x_n)$, where $x_1, x_2, \ldots, x_n$ are random variables representing the aforementioned parameters.

Fig. 5 shows the relationship between different parameters of input and output probability distributions. Here, the single random variable x represents an independent input parameter (such as channel length), and $y = g(x)$ is the dependent output random variable (such as subthreshold leakage). It should be noted that in actual situations, the output variable is a function of multiple independent variables. The mean and standard deviation of random variable x are labeled as μ_x and σ_x, respectively. The nominal value (η_y), mean value (μ_y), and variance (σ_y) of y, which are also indicated in Fig. 5, can be expressed as follows:

$$\eta_y = g(\mu_{x_1}, \mu_{x_2}, \ldots, \mu_{x_n}) \tag{1}$$

$$\mu_y = \int_{x_n^{\min}}^{x_n^{\max}} \cdots \int_{x_1^{\min}}^{x_1^{\max}} g(x_1, x_2, \ldots, x_n) \times p(x_1, x_2, \ldots, x_n) dx_1 \ldots dx_n \tag{2}$$

$$\sigma_y^2 = \int_{x_n^{\min}}^{x_n^{\max}} \cdots \int_{x_1^{\min}}^{x_1^{\max}} (g(x_1, x_2, \ldots, x_n) - \mu_y)^2 \times p(x_1, x_2, \ldots, x_n) dx_1 \ldots dx_n. \tag{3}$$

TABLE II
Equations Used in Estimation of Delta for Different Functions

$Y=g(x)$	Delta (δ_y)
$y_1 = Ae^{\beta\frac{x}{\mu}}$	$\delta_{y_1} \approx e^{\frac{\beta^2 S^2}{2}} - 1$
$y_2 = \frac{A}{x} e^{-\beta\frac{x}{\mu}}$	$\delta_{y_2} \approx (1 + \beta S^2) e^{\frac{\beta^2 S^2}{2}} - 1$
$y_3 = \frac{A}{x^2} e^{-\kappa\frac{x}{\mu}}$	$\delta_{y_3} \approx (1 + 2\kappa S^2) e^{\frac{\kappa^2 S^2}{2}} - 1$
$y_4 = Ax^2 e^{-\kappa\frac{x}{\mu}}$	$\delta_{y_4} \approx (1 - 2\kappa S^2) e^{\frac{\kappa^2 S^2}{2}} - 1$

In the aforementioned equations, $\mu_{x1}, \mu_{x2}, \ldots, \mu_{xn}$ represent the mean, and $p(x_1, x_2, \ldots, x_n)$ represents the joint probability density function (pdf) of random variables $x_1, x_2, \ldots, x_n$. In addition to the aforementioned parameters, delta (δ_y) and spread (S_y) values of y are defined as

$$\delta_y = \frac{\mu_y - \eta_y}{\eta_y} \quad \text{and} \quad S_y^2 = \frac{\sigma_y^2}{\eta_y^2}. \tag{4}$$

The parameter delta (δ_y) indicates a normalized increase in y, while spread (S_y) indicates a normalized variance of y. The analysis presented in this paper assumes Gaussian distributions for $x_1, x_2, \ldots, x_n$ which is supported by experimental observations [17]. In the next section, first, a mathematical framework to model the fluctuations in y under Gaussian variation of a single parameter is presented. Then, four different functions (y_1, y_2, y_3, and y_4) relevant to this paper have been analyzed under this framework (Table II). Subsequently, a framework to extend this methodology for independent multivariant Gaussian parameter variations is presented.

A. Single Gaussian Parameter Variation

Given a function $y = g(x)$, the mean of y can be expressed as

$$\mu_y = \int_{-\infty}^{+\infty} g(x) f_{\mu_x,\sigma_x}(x) dx \tag{5}$$

where $f_{\mu_x,\sigma_x}(x)$, μ_x, and σ_x represent the pdf, the mean, and the variance of x, respectively. Applying Taylor's theorem, one arrives at

$$\mu_y = \sum_{n=0}^{+\infty} \frac{g^{(n)}(\mu_x)}{n!} \int_{-\infty}^{+\infty} (x - \mu_x)^n f_{\mu_x,\sigma_x}(x) dx. \tag{6}$$

Considering the fact that odd terms ($n = 2k + 1$) inside the integral vanish to zero and $\eta_y = g(\mu_x)$, (6) can be simplified as

$$\mu_y = \sum_{k=0}^{+\infty} \frac{g^{(2k)}(\mu_x)}{2^k k!} \sigma^{2k} = \eta_y + \sum_{k=1}^{+\infty} \frac{g^{(2k)}(\mu_x)}{2^k k!} \sigma^{2k}. \tag{7}$$

By neglecting higher order terms in (7) and preserving only first two terms, one arrives at

$$\delta_y = \frac{\mu_y - \eta_y}{\eta_y} = \frac{g''(\mu_x)}{g(\mu_x)} \frac{\mu_x^2 S_x^2}{2}. \tag{8}$$

From the aforementioned equation, it can be observed that the sign of δ_y depends on the second derivative of $g(x)$. A positive delta (δ_y) indicates an average increase in the parameter y. Table II lists delta (δ_y) for two different functions $y_1 = g_1(x)$ and $y_2 = g_2(x)$ that will be used in the estimation of subthreshold leakage. For these functions, delta (δ_y) was calculated using (9) and (10). Note that parameters μ and S without subscript are used for simplicity and refer to statistical variables of input, i.e., μ_x and S_x, respectively. Details of the derivation are shown in the Appendix.

$$\delta_{y_1} = e^{\frac{\beta^2 S^2}{2}} - 1 \tag{9}$$

$$\delta_{y_2} = (1 + \beta S^2) e^{\frac{\beta^2 S^2}{2}} - 1. \tag{10}$$

In a similar way, delta (δ_y) can be derived for two other functions $y_3 = g_3(x)$ and $y_4 = g_4(x)$, which will be used in gate leakage estimation in Section IV. Table II shows these two functions along with their corresponding delta approximations obtained from analytical calculations. Note that all these approximations are only valid within 3σ range around the mean value. In addition, it should be noted that integrating the functions $y_2(x)$ and $y_3(x)$ from $-\infty$ to $+\infty$ is not feasible due to the $1/x$ factor that goes to infinity when $x = 0$. However, since $x = 0$ is outside the region of interest (we are interested only in positive values of different parameters), this does not affect the mathematical framework.

B. Multiple Gaussian Parameter Variation

In a manufacturing process, process variations can be expressed in terms of variations in multiple parameters (e.g., channel length, gate oxide thickness, supply voltage, and temperature). If these parameters are determined at different steps of the manufacturing process, they can be assumed to be statistically independent [17]. In that case, it can be assumed that these Gaussian variables are perfectly uncorrelated. If the function $y = g(x_1, x_2, \ldots, x_n)$ can be expressed in a variable separable form, expression for δ_y can be easily evaluated. For instance, if y can be expressed as a product of individual functions $g_i(x_i)$, i.e.,

$$y = \prod_{i=1}^{n} g_i(x_i), \quad \text{then} \quad \delta_y = \frac{\mu_y - \eta_y}{\eta_y}. \tag{11}$$

Since $x_1, x_2, \ldots, x_n$ are independent, one arrives at

$$\delta_y = \frac{\mu_{g_1}\mu_{g_2}\ldots\mu_{g_n} - g_1(\mu_{x_1})g_2(\mu_{x_2})\ldots g_n(\mu_{x_n})}{g_1(\mu_{x_1})g_2(\mu_{x_2})\ldots g_n(\mu_{x_n})} \tag{12}$$

$$\Rightarrow \quad \delta_y = \left(\frac{\mu_{g_1}}{g_1(\mu_{x_1})}\right)\left(\frac{\mu_{g_2}}{g_2(\mu_{x_2})}\right)\cdots\left(\frac{\mu_{g_n}}{g_n(\mu_{x_n})}\right) - 1$$

$$\Rightarrow \quad \delta_y = [(1+\delta_{g_1})(1+\delta_{g_2})\cdots(1+\delta_{g_n})] - 1$$

$$\Rightarrow \quad \delta_y = \left[\prod_{i=1}^{n}(1+\delta_{g_i})\right] - 1 \tag{13}$$

where δ_{gi} represents the spread of leakage due to $g_i(x_i)$ only. μ_{gi} and μ_{xi} represent the mean values of $g_i(x_i)$ and x_i, respectively $(i = 1, 2, \ldots, n)$. Equation (13) will be used in Section III to calculate δ_y under multivariable parameter variations.

III. Analysis of Subthreshold Leakage Under Within-Die Parameter Variations

A. Calculation of Delta for Subthreshold Leakage

Subthreshold leakage current for a MOSFET can be modeled as [18]

$$\begin{aligned} I_{\text{OFF}} &= I_{\text{ds}} \\ &= \mu_{\text{eff}} C_{\text{ox}} \frac{W_{\text{eff}}}{L_{\text{eff}}}(m-1)V_T^2 e^{\frac{V_{\text{gs}}-V_{\text{th}}}{mV_T}}\left(1 - e^{-\frac{V_{\text{ds}}}{V_T}}\right) \\ &\approx I_{s0}\frac{W_{\text{eff}}}{L_{\text{eff}}} e^{\frac{V_{\text{gs}}-V_{\text{th}}}{mV_T}} \end{aligned} \tag{14}$$

where

$$m = 1 + \frac{\sqrt{\varepsilon_{\text{si}} q N_a / 4\Psi_B}}{C_{\text{ox}}}$$

and

$$I_{s0} = \mu_{\text{eff}} C_{\text{ox}}(m-1)V_T^2\left(1 - e^{-\frac{V_{\text{ds}}}{V_T}}\right).$$

Here, μ_{eff} is the effective mobility, C_{ox} is the gate-oxide capacitance, L_{eff} is the effective channel length, W_{eff} is the effective width, V_T is the thermal voltage, N_a is the channel doping concentration, q is the charge of electron, ε_{Si} is the permittivity of silicon, and ψ_B is the difference between Fermi potential and intrinsic potential.

Since μ_{eff} and V_T are proportional to $T^{-1.5}$ and T, respectively, I_{s0} in (14) can essentially assumed to be temperature independent [18]. In other words, it can be regarded as a technology parameter, which is independent of device parameters such as effective channel length, supply voltage, and temperature. The temperature dependence for I_{s0} and I_{OFF} was estimated using BSIM3 equations and is shown in Fig. 6. Here, the values are normalized to their respective values at 300 K. It can be clearly observed that I_{s0} is relatively independent of temperature and, hence, assumed to be constant in this paper. It should be noted that the power of the exponential term in (14) is a function of temperature both due to the temperature dependence of the threshold voltage V_{th} and the thermal voltage $V_T = kT/q$. However, since $V_{\text{gs}} - V_{\text{th}} \gg V_T$, the impact of V_T variation becomes negligible and I_{OFF} increases exponentially with temperature [19]. Nonetheless, to obtain more accurate results, it is possible to account for the impact of V_T variation

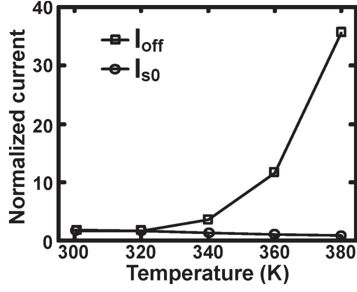

Fig. 6. Temperature dependence of I_{OFF} and I_{s0}.

by using a Taylor approximation for the $1/T$ term of V_T in a fashion similar to the approaches that will be explained in Section IV for gate leakage estimation under V_{dd} and t_{ox} variations.

Now, for small variations in channel length, threshold voltage (V_{th}) can be assumed to be linear around the nominal value of channel length $(\eta_{L\mathrm{eff}})$ and can be expressed as

$$V_{\mathrm{th}} = \mu_{V_{\mathrm{th}}} - \beta_{L_{\mathrm{eff}}} \frac{\Delta L_{\mathrm{eff}}}{\eta_{L_{\mathrm{eff}}}} \cdot mV_T \tag{15}$$

where $\mu_{V_{\mathrm{th}}}$ is the mean value of threshold voltage, $\Delta L_{\mathrm{eff}} = \eta_{L\mathrm{eff}} - L_{\mathrm{eff}}$, and $\beta_{L\mathrm{eff}}$ is a constant for a device (NMOS/PMOS) defined as

$$\beta_{L_{\mathrm{eff}}} = \frac{\eta_{L_{\mathrm{eff}}}}{mV_T} \frac{dV_{\mathrm{th}}}{dL_{\mathrm{eff}}}. \tag{16}$$

By using (14)–(16), one can write

$$I_{\mathrm{ds}} = \frac{A}{L_{\mathrm{eff}}} e^{-\beta_{L_{\mathrm{eff}}} \frac{L_{\mathrm{eff}}}{\eta_{L_{\mathrm{eff}}}}} \quad \text{where}$$
$$A = I_{s0} W_{\mathrm{eff}} \cdot e^{\frac{V_{\mathrm{gs}} - \mu_{V_{\mathrm{th}}}}{mV_T} + \beta_{L_{\mathrm{eff}}}}. \tag{17}$$

The result from Table II can now be applied to calculate delta (δI_{ds}) of leakage due to L_{eff} variation as

$$\delta_{I_{\mathrm{ds}}} \approx \left(1 + \beta_{L_{\mathrm{eff}}} \cdot S^2_{L_{\mathrm{eff}}}\right) e^{\frac{\beta^2_{L_{\mathrm{eff}}}}{2} S^2_{L_{\mathrm{eff}}}} - 1 \tag{18}$$

where $S_{L\mathrm{eff}}$ is the spread in channel length. Thus, using (18), increase in subthreshold leakage current due to the within-die channel length variations can be easily calculated. The only unknown in (18), $\beta_{L\mathrm{eff}}$, which is a critical parameter, will be estimated in the next section. The model presented above for calculating the subthreshold-leakage increase is simple but accurate since $\beta_{L\mathrm{eff}}$ is estimated from the BSIM models.

B. Calculation of β

Analytically, $\beta_{L\mathrm{eff}}$ can be calculated using (16). The calculation of β for other variables such as T, t_{ox}, and V_{dd} can also be done using equations similar to (16). The threshold voltage (V_{th}) in (16) can be expressed as [20]

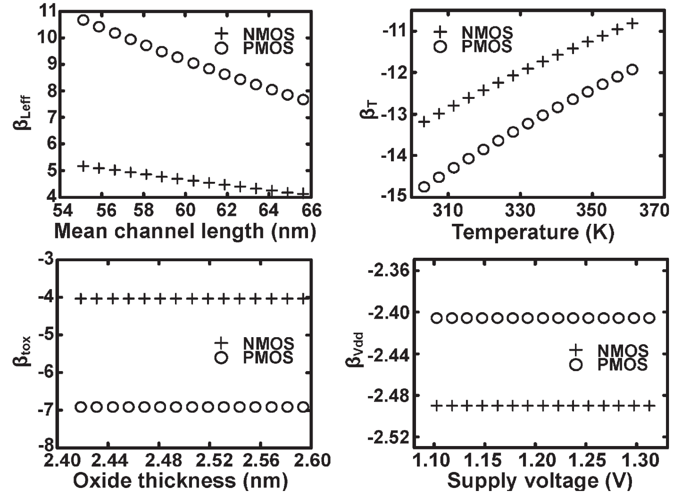

Fig. 7. Variation of β with L_{eff}, T, t_{ox}, and V_{dd}.

$$V_{\mathrm{th}} = V_{\mathrm{th0}} + \Delta V_{\mathrm{th}}(\mathrm{BODY}) + \Delta V_{\mathrm{th}}(\mathrm{HALO}) - \Delta V_{\mathrm{th}}(\mathrm{SCE, DIBL}). \tag{19}$$

In this formula, $\Delta V_{\mathrm{th}}(\mathrm{BODY})$, $\Delta V_{\mathrm{th}}(\mathrm{HALO})$, and $\Delta V_{\mathrm{th}}(\mathrm{SCE, DIBL})$ are changes in threshold voltage from its original value (V_{th0}) due to body bias, pocket (Halo) implant, and short channel effects, including drain-induced barrier lowering (DIBL), respectively. The fringe-field and narrow-width effects have been ignored in (19). Vertical nonuniform doping (retrograde doping) is used to increase the threshold voltage, and it is considered within V_{th0}. $\Delta V_{\mathrm{th}}(\mathrm{BODY})$ and $\Delta V_{\mathrm{th}}(\mathrm{HALO})$ are independent of effective channel length, whereas $\Delta V_{\mathrm{th}}(\mathrm{SCE, DIBL})$ is exponentially dependent on the channel length

$$\Delta V_{\mathrm{th}}(\mathrm{SCE, DIBL}) = [2(V_{\mathrm{bi}} - \Phi_s) + V_{\mathrm{DS}}] \cdot \left(e^{\frac{L_{\mathrm{eff}}}{2L_t}} + 2e^{\frac{L_{\mathrm{eff}}}{L_t}}\right) \tag{20}$$

where V_{bi} is the built-in voltage of the source/drain junctions, L_t represents the characteristic length of the device, and Φ_S is the Fermi potential. Because of an exponential dependence of $\delta_{I_{\mathrm{ds}}}$ on β, the value of β is critical, and hence, BSIM equations are used to obtain an accurate estimate of β. Berkeley Predictive Technology Model (BPTM) parameters [21] and BSIM3.2 modeling equations [20] are first used to calculate V_{th} as a function of L_{eff}, and then, (16) is used to evaluate $\beta_{L\mathrm{eff}}$. In a similar fashion, β values corresponding to threshold-voltage fluctuation due to temperature, oxide thickness, and supply voltage variations are calculated using (16) by replacing L_{eff} with one of the aforementioned parameters. Results are shown in Fig. 7, where $\beta_{L\mathrm{eff}}$, β_T, β_{OX}, and $\beta_{V_{\mathrm{dd}}}$ are different β values measured under variations of effective channel length, temperature, oxide thickness, and supply voltage, respectively. It can be observed that β values vary over a small range; hence, the BSIM equations are used to calculate $\delta_{I_{\mathrm{ds}}}$ as a function of effective channel length (or other parameters) spread, which is then curve-fitted with (18) to calculate an

average value of β. Curve fitting allows us to consider the fact that β is slightly higher when the effective channel length is shorter than the nominal value and that the transistors with shorter effective channel length contribute more to total leakage than the transistors with longer effective channel length. In the aforementioned analysis, the BPTM parameters for 100 nm ($V_{\rm ds} = 1.2$ V, $T_{\rm nom} = 300$ K, and $L_{\rm eff} = 60$ nm) have been used.

Since β is the slope of threshold voltage with respect to parameters such as channel length, oxide thickness, temperature, and supply voltage, it represents the sensitivity of threshold voltage to these parameters. From Fig. 7, it can be observed that β for channel length is positive, since larger channel length increases the threshold voltage. In addition, this parameter decreases with an increasing channel length, and therefore, threshold voltage becomes less and less sensitive to channel length for longer channel devices. β for temperature is negative since increasing temperature reduces the threshold voltage, and thus, a lower gate voltage is required to turn ON a transistor. The absolute value of β decreases with an increasing temperature, implying that the threshold voltage becomes less and less sensitive at higher temperatures. β for oxide thickness is negative because decreasing oxide thickness increases gate capacitance, and hence, lower gate voltage is required to invert the channel. β for supply voltage is negative since increasing voltage at the drain lowers the threshold voltage because of DIBL.

In summary, it is shown in this section that calculating β through the aforementioned methodology and inserting its value in (18) give results that are similar to those predicted by the BSIM models. Although it might be time consuming to extract the values of β, it is important to note that the values of β (for each parameter) need to be extracted only once for a particular technology node. Once evaluated, (18) provides an analytical equation to estimate the subthreshold leakage given the variation in channel length (or another parameter). Since the proposed approach is much faster than the traditional Monte Carlo analysis, it provides circuit designers with a simple and, yet, accurate statistical model for the estimation of subthreshold leakage.

C. Comparison With Existing Statistical Models

In this section, the results obtained under the proposed framework are compared with those obtained from the existing statistical models [11], [12] and the BSIM models. Equations (21) and (22) are used to calculate the subthreshold-leakage increase based on the BSIM model, which assumes Gaussian variation for channel length [1]

$$\delta_{I_{\rm ds}} = \frac{\frac{1}{a} \cdot \int\limits_{\mu-3\sigma}^{\mu+3\sigma} I_{\rm SUB}(L) \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(L-\mu)^2}{2\sigma^2}} dL}{I_{\rm SUB}(\mu)} - 1 \qquad (21)$$

where

$$a = \frac{1}{\sqrt{2\pi}\sigma} \int\limits_{\mu-3\sigma}^{\mu+3\sigma} e^{\frac{-(L-\mu)^2}{2\sigma^2}} dL. \qquad (22)$$

TABLE III
COMPARISON WITH THE EXISTING STATISTICAL MODELS

Model	Approximation for $\delta_{I_{ds}}$
This work	$\delta_{I_{ds}} \approx (1+\beta_{L_{eff}} . S_{L_{eff}}^2) e^{\frac{\beta_{L_{eff}}^2}{2} S_{L_{eff}}^2} - 1$
Narendra *et al.*, [11]	$\delta_{I_{ds}} = e^{\frac{\beta_{L_{eff}}^2}{2} S_{L_{eff}}^2} - 1$
Srivastava *et al.*, [12]	$\delta_{I_{ds}} = 2S^2 + 2\beta_{L_{eff}} S^2 + \beta_{L_{eff}}^2 S^2$

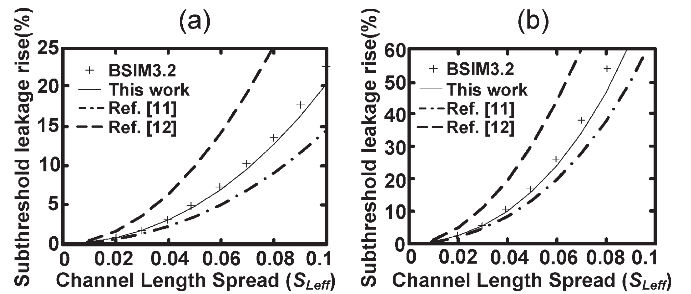

Fig. 8. Percentage increase in leakage plotted for different values of $S_{L\rm eff}$, as predicted by different models (Narendra *et al.* [11] and Srivastava *et al.* [12]) for the (a) NMOS and (b) PMOS devices.

In the aforementioned equation, $I_{\rm SUB}(L)$ is the subthreshold leakage current estimated using the BSIM equations. μ and σ are the mean and variance of channel length. In this paper, BSIM3.2 equations are exported into MATLAB software so that different mathematical expressions (such as for the β parameters) can be easily calculated based on BSIM.

Table III summarizes the increase in subthreshold leakage predicted by both the proposed and existing models [11], [12]. Fig. 8(a) and (b) plots the percentage increase in subthreshold leakage as a function of spread in channel length for NMOS and PMOS, respectively. In addition, the increase in leakage estimated by BSIM models is plotted. It can be observed that the proposed model compares well with the results from BSIM model in the region of interest. On the other hand, the model in the paper by Narendra *et al.* [11] underestimates the increase in leakage since it neglects the $1/L_{\rm eff}$ prefactor in (17). The model in the paper by Srivastava *et al.* [12] overestimates the leakage since it uses a Taylor series approximation for the calculation of the mean. It should be noted that since curve-fitted values of β are used while comparing [11], results obtained under [11] will appear even worse if nominal values of β are used. This is because, the curve fitting of β allows us to consider the variations in β which have been ignored in [11] and [12]. The model presented in [10] involves a large number of parameters, and unlike (18), the expressions derived in [10] do not give a clear insight on the impact of variations on leakage. For instance, (18) clearly shows that variations result in an increase in leakage, which cannot be easily discerned in [10].

The subthreshold leakage for PMOS transistors has stronger dependence on the channel length variation as compared to the NMOS transistors, as shown in Fig. 8. This is because of the fact that NMOS has a β of 5.2 as compared to a β of ten for

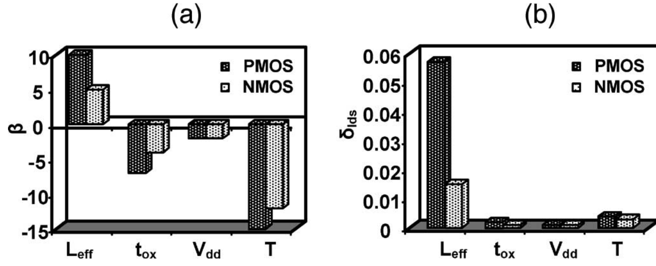

Fig. 9. (a) β values for different within-die variations for NMOS and PMOS for 90-nm devices at 300 K. (b) $\delta_{I_{\rm ds}}$ for leakage contributed by different within-die variations for NMOS and PMOS at 300 K.

PMOS at the 100-nm technology node. This is due to a steeper $V_{\rm th}$ rolloff for PMOS than for NMOS [20].

D. Considering All Within-Die Variations

In addition to the within-die channel length variation, now, assuming to be statistically independent, within-die gate oxide thickness $(t_{\rm ox})$, supply voltage $(V_{\rm dd})$, and temperature (T) variations are taken into account. Considering these variations, (14) can be rewritten as follows:

$$\begin{aligned} I_{\rm ds} &= I_{s0}\frac{W_{\rm eff}}{L_{\rm eff}} e^{\frac{V_{\rm gs}-\overline{V_{\rm th}}}{mV_T} - \beta_{L_{\rm eff}}\frac{\Delta L_{\rm eff}}{L_{\rm eff}} - \sum\limits_{X:r.\nu.} \beta_X \frac{\Delta X}{X}} \\ &= I_{s0} W_{\rm eff} e^{-\frac{V_{\rm gs}-\overline{V_{\rm th}}}{mV_T}} \left(\frac{1}{L_{\rm eff}} e^{-\beta_{L_{\rm eff}}\frac{\Delta L_{\rm eff}}{L_{\rm eff}}}\right) \prod_{X:r.\nu.} (e^{-\beta_X \frac{\Delta X}{\overline{X}}}). \end{aligned} \tag{23}$$

Here, $X : r.\nu.$ represents the random variable for different parameters (such as $t_{\rm ox}$, T, and $V_{\rm dd}$), and $\overline{X}$ stands for the nominal value of the corresponding X. In addition, ΔX denotes the difference between X and $\overline{X}$, and β_X is a constant defined in (24). Here, again, threshold voltage $(V_{\rm th})$ is assumed to vary linearly around the nominal value of X

$$\beta_X = \frac{\overline{X}}{mV_T}\frac{dV_{\rm th}}{dX}. \tag{24}$$

By using (13), one can calculate the delta (δ) of the leakage current as

$$\begin{aligned} \delta I_{\rm ds} &\approx \left[\left(1+\beta_{L_{\rm eff}}\cdot S_{L_{\rm eff}}^2\right) e^{\frac{\beta_{L_{\rm eff}}^2}{2} S_{L_{\rm eff}}^2}\right] \prod_{X:r.\nu.} (e^{\frac{\beta_X^2}{2} S_X^2}) - 1 \\ &\approx \left(1+\beta_{L_{\rm eff}}\cdot S_{L_{\rm eff}}^2\right) e^{\frac{\beta_{L_{\rm eff}}^2}{2} S_{L_{\rm eff}}^2} - 1 + \sum_{X:r.\nu.} (e^{\frac{\beta_X^2}{2} S_X^2} - 1). \end{aligned} \tag{25}$$

According to the ITRS [3], the effective channel length spread and the gate oxide thickness spread are projected to stay at $3\sigma = 10\%$ and 3%, respectively. From [8], one can obtain an estimation of within-die supply voltage and within-die temperature variations, and choose supply-voltage spread to be $3\sigma = 5\%$ and temperature spread to be $3\sigma = 3\%$ (3σ corresponds to 12 K variation at a nominal value of 300 K). Assuming these variations, Fig. 9(a) and (b) plots the β and $\delta_{I_{\rm ds}}$ values for NMOS as well as PMOS devices. β values are evaluated as discussed in Section III-B.

Note that threshold voltage is dependent on the supply voltage variations through body effect and DIBL. It can be observed that the impact of channel length and temperature variations is much greater than other variations. However, it is apparent from (25) that all of those variations serve to increase the total subthreshold leakage. In this analysis, it is assumed (for simplicity) that within-die temperature variation is independent of other within-die variations.

IV. Analysis of Gate Leakage Under Within-Die Parameter Variations

A. Calculation of Delta for Gate Leakage

Gate leakage current per unit area for a MOSFET device can be modeled as [18]

$$I_{\rm gate} = WL_{\rm eff}A_g\left(\frac{V_{\rm dd}}{t_{\rm ox}}\right)^2 e^{\left[\frac{-B_g\left(1-(1-V_{\rm dd}/\phi_{\rm ox})^{3/2}\right)}{V_{\rm dd}} t_{\rm ox}\right]} \tag{26}$$

where $\phi_{\rm ox}$ is the barrier height of tunneling electron, W is the width of the device, $L_{\rm eff}$ is the effective channel length, and A_g and B_g are the physical parameters. Due to their wider range of variation, $L_{\rm eff}$, $t_{\rm ox}$, and $V_{\rm dd}$ are the three dominant sources of variation for gate leakage current. Assuming all parameters to be constant other than $L_{\rm eff}$, gate leakage becomes linearly proportional to $L_{\rm eff}$. On the other hand, from (4), one can easily derive that for any linear function, spread of output variable (S_y) is equal to the spread of input variable (S_x). Therefore, one can conclude that a particular $L_{\rm eff}$ variation (for example, $S_x = 10\%$) results in the same amount of fluctuation in gate leakage $(S_y = 10\%)$ because of the linear relationship between the gate leakage and the channel length. Therefore, it is easy to calculate the impact of channel length variation on gate leakage, and there is no need for analytical approximations. However, the dependence of gate leakage on $t_{\rm ox}$ and $V_{\rm dd}$ is more complicated. To model the impact of oxide thickness and power supply voltage fluctuations, one can notice that assuming $L_{\rm eff}$ and $V_{\rm dd}$ to be constant and $x = t_{\rm ox}$, (26) can be written in the general form of $y_3(x)$ presented in Table II. Therefore, gate leakage deviation due to $t_{\rm ox}$ fluctuation can be easily calculated from the formula given for δy_3, where κ denotes the sensitivity of the exponential term in (26) to the oxide thickness variations. This exponential term (neglecting its negative sign) is represented by α in

$$\alpha = \frac{B_g\left(1-(1-V_{\rm dd}/\phi_{\rm ox})^{3/2}\right)}{V_{\rm dd}} t_{\rm ox}. \tag{27}$$

For small range of $t_{\rm ox}$ variation $(\Delta t_{\rm ox})$, α can be rewritten as (28), where $\overline{\alpha}$ is the nominal value of α at the nominal oxide thickness $(\overline{t_{\rm ox}})$, and $\kappa_{t{\rm ox}}$ denotes the sensitivity of α to the oxide thickness variations

$$\alpha = \overline{\alpha} + \kappa_{t_{\rm ox}} \cdot \frac{\Delta t_{\rm ox}}{\overline{t_{\rm ox}}}. \tag{28}$$

By using (26), (27), and (28), one arrives at

$$I_{\rm gate} = C\left(\frac{1}{t_{\rm ox}}\right)^2 e^{-\kappa_{t_{\rm ox}}\cdot\frac{\Delta t_{\rm ox}}{t_{\rm ox}}} \quad \text{where}$$
$$C = WL_{\rm eff}A_g(V_{\rm dd})^2 e^{-\overline{\alpha}}. \tag{29}$$

Therefore, assuming that spread of $t_{\rm ox}$ is given by $S_{t{\rm ox}}$, the gate leakage variation ($\delta_{I_{\rm gate}}$) due to oxide thickness fluctuations can be modeled as

$$\delta_{I_{\rm gate}} \approx \left(1 + 2\kappa_{t_{\rm ox}}\cdot S^2_{t_{\rm ox}}\right) e^{\frac{\kappa^2_{t_{\rm ox}}}{2}\cdot S^2_{t_{\rm ox}}} - 1 \tag{30}$$

where $\kappa_{t{\rm ox}}$ denotes the sensitivity of the exponential term in (26) to the oxide thickness variations. Procedure to model gate leakage variation due to $V_{\rm dd}$ fluctuations is almost similar. The only difference is that the power of the exponential function in (26) is not a linear function of $V_{\rm dd}$; therefore, in order to be able to use the formulas presented in Table II, Taylor approximation is required to convert it to an appropriate linear form. To do so, it is assumed that for small variations in supply voltage ($\Delta V_{\rm dd}$), (27) can be estimated by the first-order Taylor approximation around its nominal value ($\overline{\alpha}$) and can be expressed as

$$\alpha = \overline{\alpha} + \kappa_{V_{\rm dd}}\cdot\frac{\Delta V_{\rm dd}}{\overline{V_{\rm dd}}} \tag{31}$$

where $\overline{\alpha}$ is the nominal value of α at the nominal supply voltage ($\overline{V_{\rm dd}}$), and $\Delta V_{\rm dd}$ denotes the variations in supply voltage. By using (26), (27), and (31), one arrives at

$$I_{\rm gate} = DV^2_{\rm dd} e^{-\kappa_{V_{\rm dd}}\cdot\frac{\Delta V_{\rm dd}}{\overline{V_{\rm dd}}}}, \quad \text{where} \quad D = \frac{WL_{\rm eff}}{t^2_{ox}}A_g e^{-\overline{\alpha}}. \tag{32}$$

It is now possible to apply the result from Table II and calculate the deviation of gate leakage ($\delta_{I_{\rm gate}}$) due to the $V_{\rm dd}$ variation. By assuming that the spread of $V_{\rm dd}$ is given by $S_{V_{\rm dd}}$, the gate leakage variation can be modeled as

$$\delta_{I_{\rm gate}} \approx \left(1 - 2\kappa_{V_{\rm dd}}\cdot S^2_{V_{\rm dd}}\right) e^{\frac{\kappa^2_{V_{\rm dd}}}{2}\cdot S^2_{V_{\rm dd}}} - 1 \tag{33}$$

where $\kappa_{V_{\rm dd}}$ denotes the sensitivity of the exponential term in (26) to power supply variations.

B. Calculation of κ

Unlike the calculation of β in case of subthreshold leakage, due to the complexity of equations, it is difficult to calculate κ analytically. Therefore, to find out accurate values for $\kappa_{V_{\rm dd}}$ and $\kappa_{t{\rm ox}}$, which are the sensitivity parameters of gate leakage with respect to supply voltage and oxide thickness fluctuations, curve fitting has been used. Simulation using the BSIM models has been employed as the reference model in the curve fitting. It should be noted that once $\kappa_{V_{\rm dd}}$ and $\kappa_{t{\rm ox}}$ are obtained for a particular technology node, they can be used for the rest of the modeling and analysis procedure without the need for further curve fitting. In Fig. 10(a), sensitivity parameters are presented for the gate leakage estimation of NMOS and PMOS devices in a 90-nm technology node. From Fig. 10(a), it can be observed

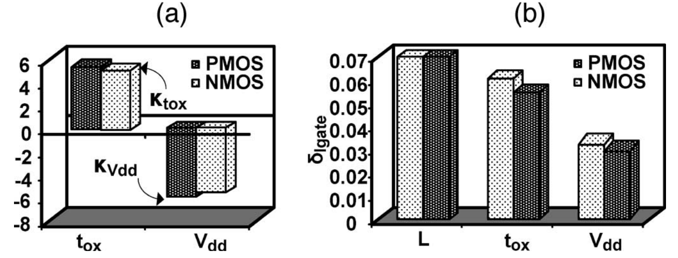

Fig. 10. (a) κ values for different within-die variations for NMOS and PMOS for 90-nm device at 300 K. The κ value corresponding to the channel length variation is equal to one (as explained in the text) and is not shown here. (b) $\delta_{I{\rm gate}}$ for leakage contributed by different within-die variations for NMOS and PMOS at 300 K.

that $\kappa_{V_{\rm dd}}$ is negative since higher supply voltage increases gate leakage and that $\kappa_{t{\rm ox}}$ is positive since thinner oxide results in higher leakage. Note that the behavior of κ is opposite to that of β (in Section III-A) due to the different factors (-1 or $+1$) that have been chosen in (15) and (28) or (31), respectively. Fig. 10(b) shows that channel length and oxide thickness are the dominant parameters affecting the delta of gate leakage.

V. Analysis of Leakage Under Die-to-Die Parameter Variations

Die-to-die variations include die-to-die channel length, temperature and voltage variations. For the purpose of yield estimation, ICs are generally screened, assuming a worst case supply voltage. Moreover, nowadays, state-of-the-art voltage regulators are relatively insensitive to the drawn current. Therefore, one can neglect the die-to-die voltage variations in the proposed analysis. On the other hand, die-to-die average temperature variations are a function of total chip power and correlate strongly to the within-die process variations such as within-die channel length variations. Therefore, the total chip power and the die temperature are self-consistently calculated, as introduced in [22], to study the impact of die-to-die temperature variations. Die-to-die channel length variations are taken into account by varying the mean value of channel length. To illustrate how die-to-die and within-die variations impact the subthreshold leakage distribution, the leakage distribution using both the BSIM3.2 and the analytical models is calculated for three cases. In case 1, only the most significant die-to-die process variation, i.e., effective channel length variation ($3\sigma = 5\%$), is considered. Equations (34) and (35) are used to calculate the leakage using the BSIM models and the analytical calculations

$$\mu_{L_{\rm eff}} - 3\sigma_{\rm die-to-die, \mathit{L}_{\rm eff}} \le L_{\rm eff} \le \mu_{L_{\rm eff}} + 3\sigma_{\rm die-to-die, \mathit{L}_{\rm eff}}$$
$$I_{\rm 1Leak,BSIM}(L_{\rm eff}) = W_n\cdot I_{{\rm SUB},n}(L_{\rm eff}) + W_p\cdot I_{{\rm SUB},p}(L_{\rm eff}) \tag{34}$$
$$I_{\rm 1Leak,Ana}(L_{\rm eff}) = I_{{\rm 1Leak,Ana},n}(L_{\rm eff}) + I_{{\rm 1Leak,Ana},p}(L_{\rm eff}) = W_n\cdot\frac{A_n}{L_{\rm eff}}e^{-\beta_{L_{{\rm eff},n}}\frac{L_{\rm eff}}{\overline{L_{\rm eff}}}} + W_p\cdot\frac{A_p}{L_{\rm eff}}e^{-\beta_{L_{{\rm eff},p}}\frac{L_{\rm eff}}{\overline{L_{\rm eff}}}}. \tag{35}$$

In the aforementioned equations $I_{\mathrm{SUB},n}$ and $I_{\mathrm{SUB},p}$ are the subthreshold currents calculated through the BSIM. Equation (35) is based on (23). Moreover, $I_{1\mathrm{Leak,BSIM}}$ and $I_{1\mathrm{Leak,Ana}}$ refer to the leakage currents calculated for case 1 using the BSIM and the proposed model, respectively. In case 2, apart from die-to-die channel length variations, within-die variations are also considered. Among the within-die variations, only the channel length and temperature variations are considered since they are more significant [Fig. 9(a)]. Ten percent of within-die variations on top of die-to-die variations are considered. Equations (36) and (37) are used for the BSIM and analytical calculations

$$\begin{aligned} I_{2\mathrm{Leak,BSIM}}(L_{\mathrm{eff}}) &= W_n \cdot I_n(L_{\mathrm{eff}}) + W_p \cdot I_p(L_{\mathrm{eff}}) \quad (36)\\ I_{2\mathrm{Leak,Ana},n} &= I_{1\mathrm{Leak,Ana},n}\left(1+\beta_{L_{\mathrm{eff}},n}\cdot S^2_{\mathrm{within-die},L_{\mathrm{eff}}}\right)\\ &\quad \times \left(e^{\frac{\beta^2_{L_{\mathrm{eff}},n}}{2}S^2_{\mathrm{within-die},L_{\mathrm{eff}}}}\cdot e^{\frac{\beta^2_T}{2}S^2_{\mathrm{within-die},T}}\right)\\ I_{2\mathrm{Leak,Ana}}(L_{\mathrm{eff}}) &= I_{2\mathrm{Leak,Ana},n}(L_{\mathrm{eff}}) + I_{2\mathrm{Leak,Ana},p}(L_{\mathrm{eff}}). \quad (37)\end{aligned}$$

In the aforementioned equations, $I_n(L_{\mathrm{eff}})$ and $I_p(L_{\mathrm{eff}})$ are the subthreshold currents for NMOS and PMOS devices which are calculated using the BSIM models assuming Gaussian distribution for channel length and temperature. $I_{2\mathrm{Leak,Ana},p}$, which represents analytically estimated leakage current for PMOS, is calculated in a similar way as $I_{2\mathrm{Leak,Ana},n}$. In case 3, die-to-die temperature variations, apart from other variations, are also included. Die-to-die temperature variations arise because of the couplings involved between power dissipation and die temperature. These couplings are taken into account by self-consistently evaluating the temperature as in [22]

$$\begin{aligned} P_{\mathrm{Total}}(L_{\mathrm{eff}},T_j) &= I_{2\mathrm{Leak,Ana}}(L_{\mathrm{eff}},T_j)\cdot V_{\mathrm{dd}}\\ &\quad + P_{\mathrm{Active}}(L_{\mathrm{eff}},T_j) + P_{\mathrm{Gate}}(L_{\mathrm{eff}}) \quad (38)\\ T_{j+1} &= T_{\mathrm{amb}} + \theta_{\mathrm{ja}}\cdot P_{\mathrm{Total}}(L_{\mathrm{eff}},T_j). \quad (39)\end{aligned}$$

In the aforementioned equations, P_{Total} is the total power dissipation $P_S = I_{2\mathrm{Leak}}(L_{\mathrm{eff}})$, V_{dd} is the total subthreshold leakage power dissipation under the within-die channel length variation, P_{Active} is the active power dissipation, $P_{\mathrm{Gate}}(L_{\mathrm{eff}})$ is the gate leakage, T_j is the average junction temperature, and θ_{ja} is the thermal resistance from silicon junction to ambient. A brief flowchart to explain the self-consistent methodology is shown in Fig. 11.

For initial average junction temperature T_0 (ambient temperature is used as an initial value), active (switching) power and total leakage (subthreshold and gate leakages) of the chip are first estimated. Total chip power is then used to calculate new junction temperature (T_{j+1}) using compact thermal models for a specific IC packaging and cooling technology [22]. Estimated junction temperature (T_{j+1}) is then compared with the value of T_j to check for convergence. The process continues until a convergence is achieved.

In this calculation, a 90-nm node microprocessor consisting of 96 million gates is employed (which are all assumed to be inverters), and average W/L per NMOS–PMOS pair is taken to be 15. BPTM specified parameters are used in the analysis. The results for the aforementioned three cases at 320 K are shown in Fig. 12(a). The X-axis plots the channel length variation (due to the die-to-die channel length variations), and the Y-axis plots the leakage power. Leakage estimations using analytical model are validated against the BSIM-based simulation results for all three cases. It can be observed that considering within-die parameter variations (case 2) results in higher leakage current. In addition, accurate leakage estimation with electrothermal couplings (case 3) predicts even higher leakage. Failing to consider this important concept causes severe estimation error.

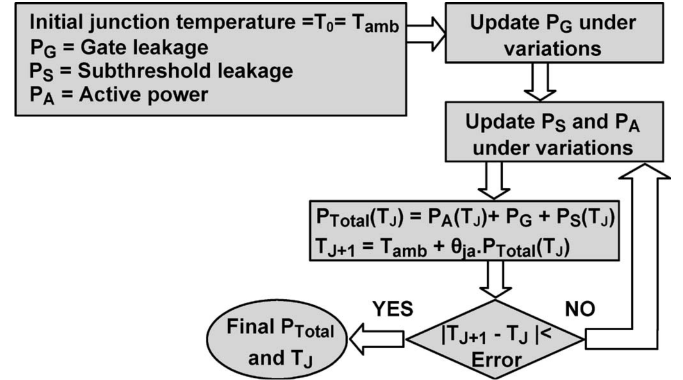

Fig. 11. Self-consistent evaluation of junction temperature and power dissipation under P–V–T variations [22]. Error is defined as a fraction of junction temperature (T_j); for example, algorithm can be executed until $|T_{j+1} - T_j| < 0.01T_j$.

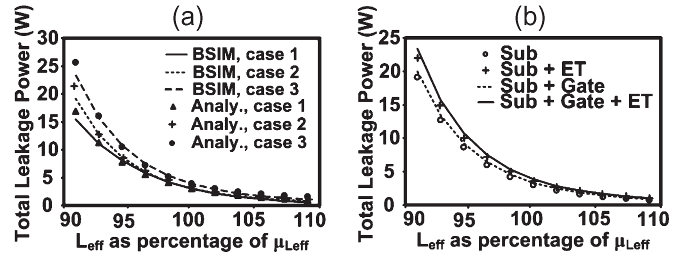

Fig. 12. (a) Subthreshold leakage power versus die-channel length variation at die temperature of 320 K. (b) Total leakage-power estimation with/without gate leakage and electrothermal couplings.

In Fig. 12(b), in order to demonstrate relative importance of the gate and subthreshold leakages, leakage power is plotted for four different scenarios. In this figure, subthreshold leakage estimation without considering electrothermal coupling is shown with the curve labeled as "Sub," and results with electrothermally aware simulation are labeled as "Sub+ET," where for both cases, gate leakage is ignored. On the other hand, curves labeled "Sub+Gate" and "Sub+Gate+ET" show the simulation results adding gate leakage to simulation results of "Sub" and "Sub+ET." Therefore, differences between curves labeled with Sub and Sub+Gate or between Sub+Gate and Sub+Gate+ET exhibit the contribution of gate leakage to the total leakage. Moreover, it can be observed that all the curves merge together as L_{eff} increases to ~110% of $\mu_{L\mathrm{eff}}$. This is due to the fact that I_{sub} (subthreshold leakage) becomes

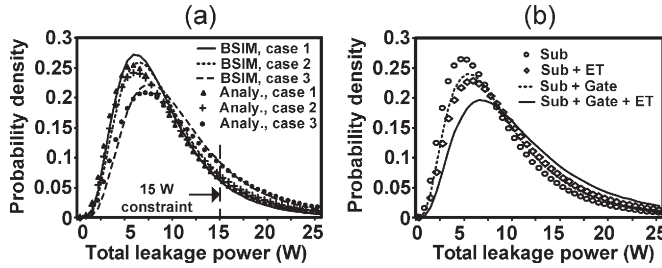

Fig. 13. (a) Leakage probability density versus total leakage power for three cases defined in Fig. 12(a). (b) Leakage probability density versus total leakage power with and without considering electrothermal couplings and gate leakage.

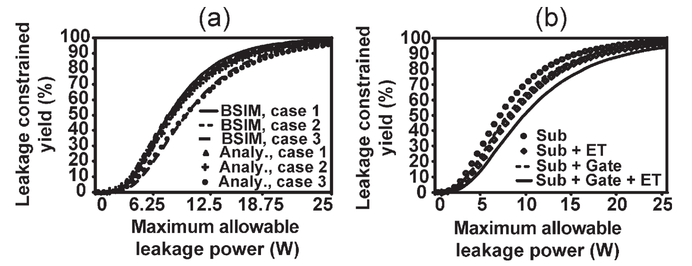

Fig. 14. (a) Leakage-constrained yield versus maximum allowable leakage calculated for Fig. 13(a). (b) Leakage-constrained yield versus maximum allowable leakage calculated for Fig. 13(b).

negligible for such channel lengths, thereby making the impact of electrothermal couplings also negligible.

VI. Leakage-Constrained Yield Estimation

It is now possible to estimate the distribution of leakage power across different dies which can be further used to calculate the yield. The Gaussian pdf $f_{P_{\rm leak}}(P_{\rm leak})$ can be calculated using

$$f_{P_{\rm Leak}}(P_{\rm Leak}) = f_{L_{\rm eff}}(L_{\rm eff}) \left(-\frac{dP_{\rm leak}(L_{\rm eff})}{dL_{\rm eff}}\right)^{-1}\Bigg|_{P(L_{\rm eff})=P_{\rm Leak}}. \quad (40)$$

Here, $f_{L{\rm eff}}(L_{\rm eff})$ represents the pdf for $L_{\rm eff}$, while $P_{\rm Leak}(L_{\rm eff})$ denotes leakage power at each particular value of the effective channel length. As more variations are taken into account, spread of total leakage-power distribution increases, which implies that larger number of dies have higher leakage power. Since active power is relatively insensitive to variations and can be assumed constant, the yield can be defined by the amount of maximum allowable leakage power consumption of the die.

Fig. 13(a) plots the pdf of leakage power versus the maximum allowable leakage power. Therefore, if leakage constraint is set to be 15 W, area under this pdf up to this point (as defined by the vertical broken line) indicates the percentage of chips that have leakage power lower than the constraint of 15 W. Therefore, leakage-constrained yield is defined as the percentage of chips which have lower leakage than this maximum allowable limit. In this figure (similar to Fig. 12), case 1 only considers the die-to-die variation, whereas in case 2, within-die variations are also taken into account. Finally, in case 3, self-consistent simulation is used to consider electrothermal couplings between subthreshold leakage and temperature. Fig. 13(a) shows the curves obtained from HSPICE simulation (using BSIM models) and analytical modeling for these three cases. It can be clearly observed that curves obtained from modeling are in good agreement with the simulation results. As one goes from case 1 to 3, the probability that chips consume higher leakage current increases, which corresponds to a shift in the distribution functions to the right. Therefore, within-die variations and electrothermally aware simulations cannot be ignored in the full-chip leakage and leakage-constrained-yield estimations.

In Fig. 13(b), to demonstrate the contribution of gate leakage to total leakage, four different probability functions have been plotted against the maximum allowable leakage power for four scenarios introduced in Fig. 12(b). Both die-to-die and within-die variations are considered. As can be observed, considering the gate leakage can further shift the pdfs toward higher leakage region. In other words, as expected, higher numbers of chips will fall in the high leakage regions if gate leakage is taken into account.

Fig. 14(a) plots the leakage-constrained yield as a function of maximum allowable leakage for the three different cases [considered in Fig. 13(a)] obtained from simulations and analytical modeling. This figure is generated using the pdf of Fig. 13(a). It can be clearly observed that variations always result in lower yield. In particular, die-to-die temperature variations due to electrothermal couplings between power and temperature significantly lower the yield. For instance, if 15 W of leakage power can be allowed in this particular design, then yield will reduce from 92% in case 1 to 88% in case 2 and to 82% in case 3. Fig. 14(b) plots the yield estimation for the four different scenarios analyzed in Fig. 13(b). Based on Fig. 14(b), considering the gate leakage results in up to 12% lower yield.

VII. Application of the Proposed Methodology for Leakage Estimation of Complex Circuits

In the previous sections, it was assumed that the entire chip is composed of only simple inverter gates. In this section, the proposed methodology is employed to estimate the subthreshold- and gate-leakage currents of more complex circuits. ISCAS85 circuits [23], which include variety of logic gates, have been chosen as benchmark circuits. A simple mathematical approach is proposed to estimate the nominal value (η) and the spread (S) of leakage for different logic gates. Then, these data are used to evaluate the nominal value (η) and the spread (S) of leakage for each benchmark circuit. Furthermore, impact of spatial correlation and stacking effect has been incorporated in the proposed methodology, as explained next.

A. Incorporating Spatial Correlations for Within-Die Variations

Within-die process variations are known to be spatially correlated due to lithography proximity effects. As a result, transistors (or digital gates) which are located close to each other are more likely to be affected in a similar way by the process variations. Spatial correlation has been studied in [24], and

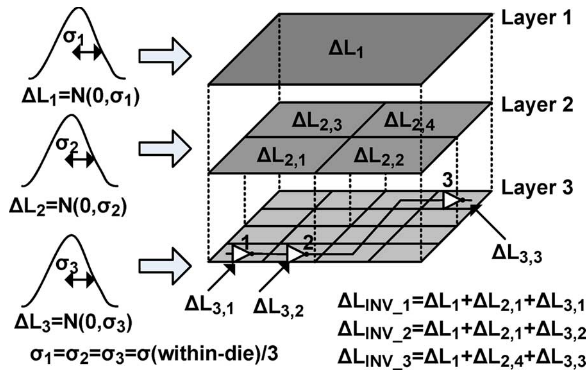

Fig. 15. Schematic illustrating the modeling of spatial correlation of within-die variations. For simplicity and due to lack of any experimental results, it is assumed that the variances (σ_1, σ_2, and σ_3) are identical.

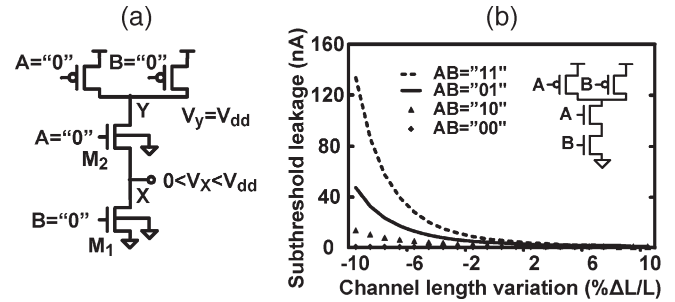

Fig. 16. Stacking effect for a two-input NAND gate in a 90-nm technology: (a) Illustration of stacking effect and (b) impact of input vectors on stacking effect and leakage. Note that we used similar size PMOS and NMOS devices ($W/L = 1$ in 90 nm). The reason for much higher leakage current in case of AB = "11" is the steeper $V_{\rm th}$ rolloff for PMOS transistors. This results in much higher subthreshold leakage for PMOS transistors of shorter channel length compared to the same-sized NMOS devices.

a model was presented, which has been incorporated in the proposed leakage estimation methodology. In this model, the variation of each parameter is split into several parts. For example, channel length of each transistor ($L_{\rm eff}$) is assumed to be equal to $L_0 + \Delta L_1 + \Delta L_2 + \cdots + \Delta L_n$, where L_0 is the nominal value of channel length, and its variation is modeled with n independent random variables ($\Delta L_1, \Delta L_2, \ldots,$ and ΔL_n). To account for spatial correlation, values of $\Delta L_1, \Delta L_2, \ldots,$ and ΔL_n must be assigned in a way such that distant devices do not share most of the $\Delta L_1, \Delta L_2, \ldots,$ and ΔL_n values, while the closer devices share some of the terms (the closer they are, the more equal terms they have), and devices that are located next to each other have identical $\Delta L_1, \Delta L_2, \ldots$, and ΔL_n values. In fact, in this model, random variable ΔL_1 accounts for the very low-frequency (the coarsest) part of the variation, whereas ΔL_n is used to model very high-frequency (the finest) portion of the variation.

This is graphically demonstrated by assigning each random variable to a certain "layer" in a vertical stack of layers, as shown in Fig. 15. To assign the ΔL_n values (for any layer n), the chip area, at each layer, is divided into a $2^{n-1} \times 2^{n-1}$ array, and different ΔL_n values are then assigned to the cells of layer n. It should be noted that since the lowest layer (with the highest value of n) is used to model high-frequency variations, therefore the grid size must be selected small enough so that variation can be assumed to be negligible inside each cell. Then, in the next layer, 2×2 arrays of cells from layer n are merged to form a single cell in layer $n-1$. In a similar fashion, ΔL_{n-1} values are assigned to each cell on layer $n-1$. This procedure continues until one reaches a single cell grid on layer 1, which corresponds to ΔL_1.

The procedure shown in Fig. 15 corresponds to a simple case where the within-die variation is modeled only with three ($n = 3$) random variables ΔL_1, ΔL_2, and ΔL_3. As a result, a three-level approach is used, one for each of ΔL_1, ΔL_2, and ΔL_3. Layer 3 is divided into a 4×4 ($2^{n-1} \times 2^{n-1}$) array, where different values of ΔL_3 are required for the 16 cells on this layer (such as, $\Delta L_{3,1}$ and $\Delta L_{3,2}$). In the next step, grid for layer 2 is built by merging each of the four 2×2 adjoining arrays of cells of layer 3, giving rise to four values of ΔL_2 (for example, $\Delta L_{2,1}$, $\Delta L_{2,2}$, etc.), and finally, a single value (obtained by merging the single 2×2 cell on layer 2) is used to model ΔL_1 in layer 1. The channel length variation for each gate (such as inverter 1) can be obtained by adding the partial values of variations (ΔL's) from different cells consisting of the one that contains the gate at layer 3 and the cells that are immediately above it at layers 1 and 2 (for example, variation of inverter 1 would be $\Delta L_1 + \Delta L_{2,1} + \Delta L_{3,1}$). As shown in Fig. 15, inverters 1 and 2, which are located in close proximity, share two components (ΔL_1 and $\Delta L_{2,1}$), whereas inverter 3, which is placed further away, only shares one parameter (ΔL_1) with the other two gates.

To incorporate this approach in the leakage estimation of ISCAS85 circuits, a three-layer grid approach (similar to Fig. 15) has been employed for all within-die parameters. In this fashion, one can ensure that gates located in nearby regions have similar parameter fluctuations.

B. Incorporating Stacking Effects

Stacking effect reduces the subthreshold leakage current of a chain of devices, where two or more transistors are connected in series. Subthreshold leakage decreases due to both higher threshold voltage ($V_{\rm th}$) and lower drain–source voltage difference ($V_{\rm ds}$) of some of the transistors (note that the subthreshold leakage is a function of $V_{\rm th}$ as well as $V_{\rm ds}$). In Fig. 16(a), the stacking effect is shown for a two-input NAND gate. In this figure, both pull-down transistors (M_1 and M_2) operate in the subthreshold region, and voltage of the internal node V_X is floating somewhere between the supply voltage and the ground. Therefore, the difference between source and body voltages of M_2 is not zero (body effect), and hence, M_2 exhibits higher $V_{\rm th}$. Moreover, the threshold voltages of M_1 and M_2 increase due to the reduced impact of the DIBL effect, as $V_{\rm ds}$ of both transistors are lower compared to a single transistor pull-down network where $V_{\rm ds}$ could be equal to the supply voltage. Furthermore, since $V_{\rm ds}$'s of both M_1 and M_2 are less than the supply voltage, the subthreshold leakage decreases even further.

It should also be noted that the impact of stacking effect in complex CMOS gates depends on input patterns. The reason is that the input pattern alters bias condition of transistors and, hence, changes the impact of stacking effect. This is shown in Fig. 16(b) where the leakage current of a two-input NAND is plotted for four possible input combinations. In the

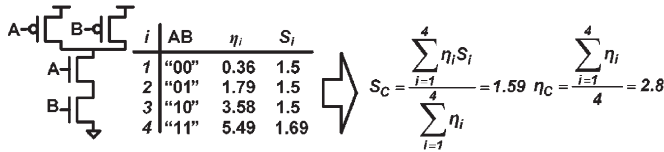

Fig. 17. Estimation of the average nominal value and spread for a NAND gate using (42).

case of $AB =$ "00," "01," and "10," the subthreshold leakage of the pull-down network is considered, whereas in the case of $AB =$ "11," the measured data correspond to the leakage of the pull-up circuit.

To explain input-pattern dependence of the stacking effect, it should be noted that current through the transistor chain is always limited to the current of transistor with the lowest subthreshold leakage. It is shown in Fig. 16(b) that leakage curve corresponding to the input pattern of $AB =$ "00" exhibits very low leakage compared to other cases due to higher $V_{\rm th}$ and lower $V_{\rm ds}$ of M_2. In the case of $AB =$ "10," M_2 is ON, and hence, voltage on the node X is one $V_{\rm th}$ lower than that of the node Y. Therefore, the subthreshold leakage of M_1 is slightly higher than that for the previous case due to lower $V_{\rm ds}$ only ($V_{\rm th}$ of M_1 is not increased due to body effect). In contrast, when $AB =$ "01," the subthreshold leakage is higher than both the previous cases because M_2 is neither affected by the body effect (source and body terminals at the same voltage) nor does it have lower $V_{\rm ds}$. Finally, in the case of $AB =$ "11," unlike the previous three cases, leakage current of the pull-up network is plotted. Since there are two parallel PMOS transistors, the subthreshold leakage in this case is the highest.

As previously illustrated, for complex gates, leakage current depends on the input pattern. As a result, the spread and nominal values of leakage are different for various input combinations. To simplify the leakage estimation process, different spread and nominal values are combined in order to obtain an average spread and nominal value for each gate. Therefore, first, the proposed analytical models are used (15)–(18) to obtain the spread and nominal values of leakage for each input pattern, and then, the results for different input patterns are merged, as shown next. Two simple formulas are proposed that enable us to calculate average nominal (η_A) and spread (S_A), given the nominal and spread values of two different input patterns (η_1, η_2, S_1, and S_2). Then, these formulas are generalized to the case of multiple input patterns so that one can merge the results for all complex gates. Considering (11), one can easily derive these equations for average nominal (η_A) and spread (S_A)

$$\eta_A = \frac{\eta_1 + \eta_2}{2} \quad \text{and} \quad S_A = \frac{S_1 \cdot \eta_1 + S_2 \cdot \eta_2}{\eta_1 + \eta_2}. \tag{41}$$

Equation (41) can be further generalized to n input patterns as

$$\eta_A = \frac{\sum_{i=1}^{n} \eta_i}{n} \quad \text{and} \quad S_A = \frac{\sum_{i=1}^{n} \eta_i \cdot s_i}{\sum_{i=1}^{n} \eta_i}. \tag{42}$$

In Fig. 17, using a NAND gate as an example, the formulas presented in (42) are used to predict the average nominal (η_A) and spread (S_A) values for this gate, given the nominal (η_i) and spread (S_i) values for each input pattern, which are already obtained from the proposed analytical models. Such information for different logic gates (INV, NAND2, etc.) has been collected which is deployed in the ISCAS85 circuits. This database is then used (as will be discussed in the next section) to evaluate the leakage current of ISCAS85 circuits.

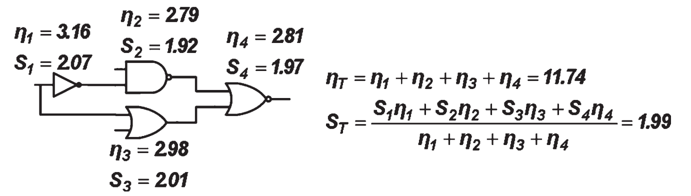

Fig. 18. Effective nominal value and spread for a simple combinational circuit using (43).

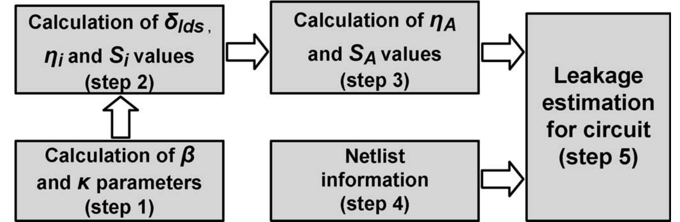

Fig. 19. Block diagram of leakage estimation of complex circuits using the proposed methodology.

C. Leakage Estimation for ISCAS85 Circuits

Assuming that one has calculated the average η and S values for all types of gates, the next step is to use these data to obtain the estimation for leakage current of each of the ISCAS85 circuits. Equation (43) can be used to calculate the overall leakage of these circuits. Here, η_A and S_A are the average nominal and spread of leakage values of different logic gates, respectively, and η_T and S_T are the nominal and spread values of the total leakage for an entire logic circuit, respectively. Equation (43) is similar to the previous equations; however, instead of average values for nominal leakage of each gate, summation of average nominal leakage values of all gates is used to predict the nominal value of total leakage. This procedure is shown for a simple logic circuit in Fig. 18 where the precalculated η_A and S_A values of logic gates are used to find η_T and S_T parameters of the circuit under study

$$\eta_T = \sum_{A=1}^{n} \eta_A \quad \text{and} \quad S_T = \frac{\sum_{A=1}^{n} \eta_A \cdot S_A}{\sum_{A=1}^{n} \eta_A}. \tag{43}$$

D. Simulation Results

The proposed methodology has been used to estimate the mean and spread values of leakage current distributions for the ISCAS85 benchmark circuits. Channel length variation of 10% at the 90-nm technology node has been considered. Fig. 19 shows a block diagram of steps which are involved in the simulation. The first step involves the calculation of β and κ

TABLE IV
ESTIMATION ERROR FOR VARIOUS ISCAS85 BENCHMARK CIRCUITS

	η (µA) HSPICE	η (µA) Estim.	η error	S (µA) HSPICE	S (µA) Estim.	S error
C499	3.79	3.64	3.9%	77.3	73	5.5%
C880	2.19	2.09	4.5%	73.4	69.5	3.9%
C1355	2.54	2.24	11.8%	62.9	58.5	6.9%
C1908	4.6	4.04	11.7%	76.4	69.5	9%
C2607	8.28	7.61	8.21%	62.7	56.9	9.3%

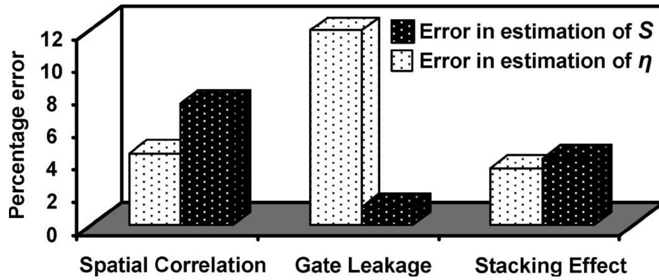

Fig. 20. Percentage error in nominal and spread of leakage current caused by ignoring spatial correlation, gate leakage, or stacking effect.

parameters for subthreshold- and gate-leakage estimations using (16), (28), and (31). Spatial correlation between parameters can be considered in this step by assigning different β and κ parameters to each portion of the chip or circuit according to their spatial placement and variation profile (as discussed in Section VII-A). In the second step, the nominal (η_i) and spread (S_i) values are evaluated for different combinations of inputs for each gate to incorporate the stacking effect (Section VII-B). In the following step, the average nominal (η_A) and spread (S_A) values are calculated for all logic gates using (42). In step 5, using netlist information (number of gates of each type from step 4), the nominal (η_T) and spread (S_T) values of total leakage are estimated for the circuits by using (43).

Using the proposed estimation methodology, the nominal and spread values of leakage current are calculated for the ISCAS85 circuits, and results are compared to those obtained from the HSPICE simulations. Assuming HSPICE results as reference values, the estimation of error for both the nominal (η) and spread (S) values has also been calculated and presented in Table IV. It can be observed that the estimation errors for both the mean and spread values are low, considering the simplicity of the method presented here.

To further investigate the relative importance of considering spatial correlation, gate leakage, and stacking effect in leakage estimation of circuits, the leakage current of ISCAS85 circuits is calculated, ignoring one of the aforementioned components at a time. For example, to evaluate the impact of gate leakage, it is assumed that gate leakage is zero, and then, the nominal (η) and spread (S) values are evaluated and compared to the corresponding values where all components were included. Results are shown in Fig. 20, which shows the relative error caused in the prediction of nominal (η) and spread (S) values due to ignoring the impact of spatial correlation, gate leakage, or stacking effects. It can be observed that dropping gate leakage results in significant error in the estimation of nominal value whereas neglecting the spatial correlation and the stacking effect deteriorates the accuracy of leakage spread estimation.

VIII. CONCLUSION

A novel framework for the accurate estimation of key statistical parameters (namely, nominal and spread values) of full-chip leakage distribution consisting of the subthreshold and gate leakages has been presented while considering both the within-die and die-to-die variations in process (P), temperature (T), and supply voltage (V). For the first time, temperature variations and, more importantly, electrothermal couplings between the substrate temperature and the leakage power have been accounted for in the leakage estimation methodology. A quantitative analysis of the relative sensitivities of device leakage components to the P–T–V variations has been performed to extract a transistor-level variation model. It was shown that the proposed statistical model, as compared to others in the literature, shows better agreement with the rigorous BSIM-model-based simulations. It is also demonstrated that failing to account for temperature variations and electrothermal couplings can result in significant inaccuracy in the leakage estimation. Furthermore, the full-chip leakage-power distribution was shown to be useful for estimating the leakage-constrained yield under the impact of variations. Calculations show that the yield is significantly lowered under the impact of within-die and die-to-die process and temperature variations. Finally, the proposed methodology was applied to estimate the leakage of complex logic circuits with a consideration of the spatial correlations of process parameters and transistor stacking effects. It was observed that neglecting the gate leakage results in significant error in the estimation of nominal value, whereas neglecting the spatial correlation and the stacking effect degrades the accuracy of leakage spread estimation.

APPENDIX

The details of the algebraic derivation for delta values shown in Table II are presented as follows.

$$\begin{aligned}
\delta_{y_1} &= \frac{1}{e^{\beta}}\left[\int_{-\infty}^{\infty} e^{\beta\frac{x}{\mu}}\cdot\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx\right]-1\\
&= \left[\int_{-\infty}^{\infty} e^{\beta\frac{x-\mu}{\mu}}\cdot\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx\right]-1\\
&= \left[\int_{-\infty}^{\infty}\left(1+\frac{\beta}{\mu}(x-\mu)+\frac{\beta^2}{\mu^2}\frac{(x-\mu)^2}{2!}+\dots\right)\right.\\
&\qquad\left.\cdot\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx\right]-1\\
&= \left[1+0+\frac{\beta^2}{\mu^2}\frac{\sigma^2}{2}+0+\frac{\beta^4}{\mu^4}\frac{3\sigma^4}{24}+\cdots\right]-1\\
&= e^{\frac{\beta^2}{\mu^2}\frac{\sigma^2}{2}}-1 = e^{\frac{\beta^2S^2}{2}}-1.
\end{aligned}$$

$$\delta_{y2} = \frac{\mu}{Ae^{-\beta}}\left[\int_{-\infty}^{\infty}\frac{A}{x}e^{-\beta\frac{x}{\mu}}\cdot\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx\right]-1$$
$$=\mu\left[\int_{-\infty}^{\infty}\frac{1}{x}e^{-\beta\frac{x-\mu}{\mu}}\cdot\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx\right]-1$$
$$\approx\mu\left[\int_{-\infty}^{\infty}\left(\frac{1}{\mu}-\frac{1}{\mu^2}(x-\mu)\right)\times e^{-\beta\left(\frac{x-\mu}{\mu}\right)}\cdot\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx\right]-1$$
$$=\int_{-\infty}^{\infty}e^{-\beta\left(\frac{x-\mu}{\mu}\right)}\cdot\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx + \left[\int_{-\infty}^{\infty}\left(-\frac{x-\mu}{\mu}\right)\times\left(1-\frac{\beta}{\mu}(x-\mu)+\frac{\beta^2}{\mu^2}\frac{(x-\mu)^2}{2!}-\frac{\beta^3}{\mu^3}\times\frac{(x-\mu)^3}{3!}\cdots\right)\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx\right]-1$$
$$=e^{\frac{\beta^2S^2}{2}}+\left[0+\frac{\beta}{\mu^2}\sigma^2+0+\frac{\beta^3}{\mu^4}\frac{3\sigma^4}{3!}+0+\frac{\beta^5}{\mu^6}\frac{15\sigma^6}{5!}+0+\cdots\right]-1$$
$$=e^{\frac{\beta^2S^2}{2}}+\beta S^2\left[1+\frac{\beta^2S^2}{2}+\frac{\beta^4S^4}{8}+\cdots\right]-1$$
$$=e^{\frac{\beta^2S^2}{2}}+\beta S^2e^{\frac{\beta^2S^2}{2}}-1=(1+\beta S^2)e^{\frac{\beta^2S^2}{2}}-1$$

$$\delta_{y3} = \frac{\mu^2}{Ae^{-\kappa}}\left[\int_{-\infty}^{\infty}\frac{A}{x^2}e^{-\kappa\frac{x}{\mu}}\cdot\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx\right]-1$$
$$=\mu^2\left[\int_{-\infty}^{\infty}\frac{1}{x^2}e^{-\kappa\frac{x-\mu}{\mu}}\cdot\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx\right]-1$$
$$\approx\mu^2\left[\int_{-\infty}^{\infty}\left(\frac{1}{\mu^2}-\frac{2}{\mu^3}(x-\mu)\right)e^{-\kappa\left(\frac{x-\mu}{\mu}\right)}\cdot\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx\right]-1,\ \text{and similar to } \delta_{y_2}$$
$$=\int_{-\infty}^{\infty}e^{-\kappa\left(\frac{x-\mu}{\mu}\right)}\cdot\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx + \left[\int_{-\infty}^{\infty}2\times\left(-\frac{x-\mu}{\mu}\right)\left(1-\frac{\kappa}{\mu}(x-\mu)+\cdots\right)\times\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx\right]-1$$
$$=e^{\frac{\kappa^2S^2}{2}}+2\times\kappa S^2e^{\frac{\kappa^2S^2}{2}}-1=(1+2\kappa S^2)e^{\frac{\kappa^2S^2}{2}}-1$$

$$\delta_{y_4} = \frac{1}{A\mu^2e^{-\kappa}}\left[\int_{-\infty}^{\infty}Ax^2e^{-\kappa\frac{x}{\mu}}\cdot\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx\right]-1$$
$$=\frac{1}{\mu^2}\left[\int_{-\infty}^{\infty}x^2e^{-\kappa\frac{x-\mu}{\mu}}\cdot\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx\right]-1$$
$$\approx\frac{1}{\mu^2}\left[\int_{-\infty}^{\infty}\left(\mu^2+2\mu(x-\mu)\right)e^{-\kappa\left(\frac{x-\mu}{\mu}\right)}\cdot\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx\right]-1,\ \text{and similar to } \delta_{y_2}$$
$$=\int_{-\infty}^{\infty}e^{-\kappa\left(\frac{x-\mu}{\mu}\right)}\cdot\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx + \left[\int_{-\infty}^{\infty}-2\times\left(-\frac{x-\mu}{\mu}\right)\left(1-\frac{\kappa}{\mu}(x-\mu)+\cdots\right)\times\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx\right]-1$$
$$=e^{\frac{\kappa^2S^2}{2}}-2\times\kappa S^2e^{\frac{\kappa^2S^2}{2}}-1=(1-2\kappa S^2)e^{\frac{\kappa^2S^2}{2}}-1.$$

REFERENCES

[1] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, "Parameter variations and impact on circuits and microarchitecture," in *Proc. Des. Autom. Conf.*, 2003, pp. 338–342.

[2] A. H. Ajami, K. Banerjee, and M. Pedram, "Modeling and analysis of non-uniform substrate temperature effects on global ULSI interconnects," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 24, no. 6, pp. 849–861, Jun. 2005.

[3] Semiconductor Industry Association. (2005 Edition). International Technology Roadmap for Semiconductors. [Online]. Available: http://www.itrs.net/Links/2005ITRS/Home2005.htm

[4] R. McGowen, "Adaptive designs for power and thermal optimization," in *Proc. Int. Conf. Comput.-Aided Des.*, 2005, pp. 118–121.

[5] C. Poirier, R. McGowen, C. Bostak, and S. Naffziger, "Power and temperature control on a 90 nm Itanium family processor," in *Proc. Int. Solid-State Circuits Conf.*, 2005, pp. 304–305.

[6] S. Borkar, "Design challenges of technology scaling," *IEEE Micro*, vol. 19, no. 4, pp. 23–29, Jul./Aug. 1999.

[7] K. A. Bowman, S. G. Duvall, and J. D. Meindl, "Impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution for gigascale integration," *IEEE J. Solid-State Circuits*, vol. 37, no. 2, pp. 183–190, Feb. 2002.

[8] H. Su, F. Liu, A. Devgan, E. Acar, and S. Nassif, "Full chip leakage estimation considering power supply and temperature variations," in *Proc. Int. Symp. Low Power Electron. Des.*, 2003, pp. 78–83.

[9] R. Rao, A. Srivastava, D. Blaauw, and D. Sylvester, "Statistical analysis of subthreshold leakage current for VLSI circuits," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 12, no. 2, pp. 131–139, Feb. 2004.

[10] S. Mukhopadhyay and K. Roy, "Modeling and estimation of total leakage current in nano-scaled CMOS devices considering the effect of parameter variation," in *Proc. Int. Symp. Low Power Electron. Des.*, 2003, pp. 172–175.

[11] S. Narendra, V. De, S. Borkar, D. Antoniadis, and A. Chandrakasan, "Full-chip sub-threshold leakage power prediction model for sub-0.18 μm CMOS," in *Proc. Int. Symp. Low Power Electron. Des.*, 2002, pp. 19–23.

[12] A. Srivastava, R. Bai, D. Blaauw, and D. Sylvester, "Modeling and analysis of leakage power considering within-die process variations," in *Proc. Int. Symp. Low Power Electron. Des.*, 2002, pp. 64–67.

[13] A. Agarwal, K. Kunhyuk, and K. Roy, "Accurate estimation and modeling of total chip leakage considering inter- & intra-die process variations," in *Proc. Int. Conf. Comput.-Aided Des.*, 2005, pp. 736–742.

[14] A. Srivastava, S. Shah, K. Agarwal, D. Sylvester, D. Blaauw, and S. Director, "Accurate and efficient gate-level parametric yield estimation considering correlated variations in leakage power and performance," in *Proc. Des. Autom. Conf.*, 2005, pp. 535–540.
[15] C. Hongliang and S. S. Sapatnekar, "Full-chip analysis of leakage power under process variations, including spatial correlations," in *Proc. Des. Autom. Conf.*, 2005, pp. 523–528.
[16] S. Zhang, V. Wason, and K. Banerjee, "A probabilistic framework to estimate full-chip subthreshold leakage power distribution considering within-die and die-to-die P–T–V variations," in *Proc. Int. Symp. Low Power Electron. Des.*, 2004, pp. 156–161.
[17] J. Zhang and M. Styblinski, *Yield and Variability Optimization of Integrated Circuits*. Boston, MA: Kluwer, 1995.
[18] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*. Cambridge, U.K.: Cambridge University Press, 1998.
[19] Y. S. Lin, C. C. Wu, C. S. Chang, R. P. Yang, W. M. Chen, J. J. Liaw, and C. H. Diaz, "Leakage scaling in deep submicron CMOS for SoC," *IEEE Trans. Electron Devices*, vol. 49, no. 6, pp. 1034–1041, Jun. 2002.
[20] BSIM3v3.2.2 MOSFET Model BSIM Group, Univ. of California Berkeley. [Online]. Available: http://www-device.eecs.berkeley.edu/~bsim3
[21] Berkeley Predictive Technology Model (BPTM): Device Group, Univ. of California at Berkeley. [Online]. Available: http://www.-device.eecs.berkeley.edu/~ptm/
[22] K. Banerjee, S.-C. Lin, A. Keshavarzi, S. Narendra, and V. De, "A self-consistent junction temperature estimation methodology for nanometer scale ICs with implications for performance and thermal management," in *IEDM Tech. Dig.*, 2003, pp. 887–890.
[23] F. Brglez and H. Fujiwara, "A neutral netlist of 10 combinational benchmark circuits and a target simulator in fortran," in *Proc. Int. Symp. Circuits Syst.*, 1985, pp. 695–698.
[24] A. Agarwal, D. Blaauw, V. Zolotov, S. Sundareswaran, M. Zhao, K. Gala, and R. Panda, "Path-based statistical timing analysis considering inter- and intra-die correlations," in *Proc. ACM/IEEE Int. Workshop Timing Issues*, 2002, pp. 16–21.

Hamed F. Dadgour (S'05) was born in Tabriz, Iran. He received the B.S. degree in electrical engineering from Sharif University of Technology, Tehran, Iran, in 1999, and the M.S. degree in electrical engineering from the University of Tehran, Tehran, in 2001. He is currently working toward the Ph.D. degree in the Department of Electrical and Computer Engineering, University of California, Santa Barbara.

He was with Sharif University of Technology as a Research Assistant until 2004. He joined Prof. Banerjee's Research Group at the University of California in Fall 2005. His current research is focused on parameter-variation issues in high-performance and low-power nanoscaled integrated circuits as well as design of low-power circuits.

Sheng-Chih Lin (S'03) received the B.S. degree in electrical engineering from the National Taiwan University, Taipei, Taiwan, in 1996. He is currently working toward the Ph.D. degree in the Department of Electrical and Computer Engineering, University of California, Santa Barbara.

From 1998 to 2002, he was with the Phoenixtec Electronics Company, Ltd., and the CHROMA ATE, Inc., respectively, in Taiwan. He joined Prof. Banerjee's Research Group at the University of California in Winter 2003. His research interests include electrothermal modeling and analysis of integrated circuits, variation-aware circuit design and optimization, and power/thermal management for nanoscale CMOS ICs. He has authored or coauthored over a dozen papers in journals and refereed international conferences.

Mr. Lin is a corecipient of the 2007 IEEE Micro Award.

Kaustav Banerjee (S'92–M'99–SM'03) received the Ph.D. degree in electrical engineering and computer sciences from the University of California, Berkeley, in 1999.

He was with Stanford University, Stanford, CA, from 1999 to 2001, as a Research Associate at the Center for Integrated Systems. From February to August 2002, he was a Visiting Faculty with the Circuit Research Laboratories, Intel, Hillsboro, OR. Since July 2002, he has been with the Faculty of the Department of Electrical and Computer Engineering, University of California, Santa Barbara, where he is currently a Professor. He has also held summer/visiting positions at Texas Instruments Incorporated, Dallas, TX, from 1993 to 1997, and the Swiss Federal Institute of Technology, Lausanne, Switzerland, in 2001. His research has been chronicled in over 135 journal and refereed international conference papers and in a book chapter. He has also coedited a book titled *Emerging Nanoelectronics: Life With and After CMOS* (Springer, 2004). His current research interests focus on nanometer-scale issues in high-performance/low-power very large scale integrated circuits as well as on circuit and system issues in emerging nanoelectronics.

Dr. Banerjee has served on the technical program committees of several leading IEEE and ACM conferences, including IEDM, DAC, ICCAD, and IRPS. He has also served on the organizing committee of ISQED at various positions, including Technical Program Chair in 2002 and General Chair in 2005. Currently, he serves as a member of the Nanotechnology Committee of the IEEE Electron Devices Society. He has received a number of awards in recognition of his work, including the ACM SIGDA Outstanding New Faculty Award in 2004, a Research Award from the Electrostatic Discharge Association in 2005, a Best Paper Award at the Design Automation Conference in 2001, an Outstanding Student Paper Award at the VLSI/ULSI Multilevel Interconnection Conference in 2005, and an IEEE Micro Award in 2007.