Statistical Modeling and Analysis of Chip-Level Leakage Power by Spectral Stochastic Method *

Ruijing Shen †, Ning Mi†, Sheldon X.-D. Tan†, Yici Cai‡ and Xianlong Hong‡
†Department of Electrical Engineering, University of California, Riverside, CA 92521
‡Department of Computer Science and Technology, Tsinghua University, Beijing, China, 100081

ABSTRACT

In this paper, we present a novel statistical full-chip leakage power analysis method. The new method can provide a general framework to derive the full-chip leakage current or power in a closed form in terms of the variational parameters, such as the channel length, the gate oxide thickness, etc. It can accommodate various spatial correlations. The new method employs the orthogonal polynomials to represent the variational gate leakages in a closed form first, which is generated by a fast multi-dimensional Gaussian quadrature method. The total leakage currents then are computed by simply summing up the resulting orthogonal polynomials (their coefficients). Unlike many existing approaches, no grid-based partitioning and approximation are required. Instead, the spatial correlations are naturally handled by orthogonal decompositions. The proposed method is very efficient and it becomes linear in the presence of strong spatial correlations. Experimental results show that the proposed method is about 10× faster than the recently proposed method [4] with constant better accuracy.

1. INTRODUCTION

Process-induced variability has huge impact on the circuit performance in the sub-90nm VLSI technologies [15]. This is the particular case for leakage power, which has increased dramatically with the technology scaling and is becoming the dominant chip power dissipation [12]. The dominant factors in the leakage currents are the subthreshold leakage current I_{sub} and gate oxide leakage I_{gate}. The subthreshold leakage current has a rapid increasing rate (about 5X-10X increase per technology generation [5]), and it is highly sensitive to threshold voltage V_{th} variations, owning to the exponential relationship between subthreshold current I_{sub} and threshold voltage V_{th}. On the other hand, as the gate oxide thickness, T_{ox}, is scaling down, I_{gate} is growing rapidly as I_{gate} has an exponential dependence on T_{ox}.

As a result, leakage variations become significant, and traditional worst case based approach will lead to extremely pessimistic and expensive design solutions. Statistical estimation and analysis of leakage powers considering the process variability are critical in various chip design steps to improve the design yield and robustness.

Many existing works have been proposed to statistically model and analyze the full-chip leakage currents and powers considering the process variations in the past [4]. Early work in [21] gives the analytic expressions of mean and variance of leakage currents of CMOS gates considering only subthreshold leakage. Method in [14] provides simple analytic expressions of leakage currents of the whole chip considering global variations only. In [13], reverse biased source/drain junction band-to-band-tunneling (BTBT) leakage current is considered, in addition to the subthreshold leakage currents, for estimating the mean values and variances of the leakage currents of gates only. In [20], the probability density function (PDF) of stacked CMOS gates and whole chip are derived considering both inter-die and intra-die variations.

Recently a full-chip leakage analysis method considering spatial correlations in the intra-die and inter-die variations was proposed [4]. But it requires $O(n^2)$ time complexity to compute the variance, where n is the number of gates in the design, which can be expensive for the large circuits. The method then introduced the grid-based partitioning of the circuits to reduce the number of variables at the loss of accuracy [4]. Work in [10] proposed a linear time complexity method to compute the mean and variance of full-chip leakage currents by exploiting the symmetric property of one existing exponential spatial correlation formula. The method only considers subthreshold leakage and it requires the chip cells/modules to be partitioned into a regular grid with similar uniform fitting functions, which typically is not the case practically.

In the paper, we propose a new general full-chip leakage modeling and analysis method. The new method starts with the process variational parameters such as the channel length, δL, gate oxide thickness, δT_{ox}, and it can derive the full-chip leakage I_{leak} in terms of those variables directly (or their corresponding transformed variables). Unlike existing grid-based methods, which trade the accuracy for speedups. The new method is gate-based method and uses principal component analysis (PCA) to reduce the number of variables with much less accuracy loss assuming that the geometrical variables are Gaussian. For non-Gaussian variables, independent component analysis (ICA) [2] can be used. The new method considers both inter-die and intra-die variations and it can work with various spatial correlations. The proposed method becomes linear under strong spatial correlations. Unlike the existing approaches [4, 10], the new method does not make any assumptions about the distributions of final total leakage currents for both gates and chips and does not require any grid-based partitioning of the chip.

In the new method, we first fit both the subthreshold and gate oxide leakage currents into analytic expressions in terms of parameter variables. We show that by using more terms in the gate level analytic models, we can achieve better accuracy than [4], Second, The new method employs the orthogonal polynomials, which gives the best representation for specific distributions [9] and is also called the *spectral stochastic* method, to represent the variational gate leakages in an analytic form in terms of the random variables. The step is achieved by using the numerical Gaussian quadrature method, which is much faster than the Monte Carlo method. The total leakage currents are finally computed by simply summing up the resulting analytical orthogonal polynomials of all gates (their coefficients). The spatial cor-

*This work is funded in part by NSF CAREER Award No. CCF-0448534, in part by NSF grant under No. OISE-0623038 and in part by National Natural Science Foundation of China (NSFC) grant under No. 60828008

978-1-4244-2749-9/09/$25.00 ©2009 IEEE

relations are taken care of by the PCA or ICA and at the same time, the number of random variables can also been substantially reduced in the presence of strong spatial correlations during the decomposition process. Experimental results on the PDWorkshop91 benchmarks on a 45nm technology show that the proposed method is about 10× faster than the recently proposed method [4] with constant better accuracy.

2. MODELING

In this section, we first present process variational models used in this work. Then we present the static analytic models used in our work for computing the full-chip leakage currents.

2.1 Process variational models

In this subsection, we present the process variations for computing variational leakage currents. Process variations include variations at different levels: wafer level, inter-die level, and intra-die level. They are caused by different sources such as lithograph, materials, aging, etc. Some of the variations are systematic, e.g., those caused by lithography process [7, 17]. Some are purely random, e.g., the doping density of impurities and edge roughness [3].

The main process parameter that has big impact on leakage current is the transistor threshold voltage V_{th}, and V_{th} is observed to be the most sensitive to effective gate length L and gate oxide thickness T_{ox}. The ITRS'06 [12] indicates that the gate length variation is a primary factor for device parameter variation, and the number of dopants in channel results in unacceptably large statistical variation of the threshold voltage.

Our circuit model is built using the PTM (predictive technology model) 45nm technology [19], and simulated in HSpice with supply voltage of 1V. We list the detailed parameters for gate length and gate oxide thickness variations in Table 1. As indicated in the second column, we decompose each category into "inter-die" and "intra-die" variations. For intra-die variation, we further decompose it into with and without spatial correlation. These variations are modeled in Normal distributions ([22, 6]). The total variance (σ^2) is computed by summing up each component variance because the sum of Normal distributions is still a Normal distribution, and their variances are thus additive.

Table 1: Process variation parameter breakdown for 45nm technology

	σ^2 Distribution		(σ)
Gate Length(L)	Inter-die	20%	4%*18nm
	Intra-die		
	∗ Sp Correlated	80%	
Gate Oxide Thickness(T_{ox})	Inter-die	20%	4%*1.8nm
	Intra-die		
	∗ Non-Correlated	80%	

Electrical measurements of a full wafer showed that the intra-die gate length variation has strong spatial correlations [7]. This implies that devices that are physically close to each other are more likely to be similar than those that are far apart. Therefore, in our model, the intra-die variation of gate lengths is modeled based on such kind of correlation. We use the empirical formulation such as the exponential model [25]

$$\gamma(r) = e^{-r^2/\eta^2} \tag{1}$$

where r is the distance between two panel centers and η is the correlation length. We notice that the strong spatial correlation suggested by (1) has been exploited by [4] to speed up the calculation, where the full chip is divided into N grids and the correlated random variables are perfectly correlated in a grid. In our approach, the strong spatial correlation is explored naturally by principal component analysis (PCA), which can transfer the correlated R.V.s into independent ones with reduced numbers. But for comparison purpose, we also implemented the grid-based variational model. We remark that non-Gaussian distributions, we can use independent component analysis (ICP) to perform the reduction [2]. For gate oxide thickness, T_{ox}, there is no such strong spatial correlation and its values are uncorrelated.

The last column of Table 1 shows the standard deviation (σ) of each variation. According to statistical theory, in normal distribution, 99% of the samples should fall in the range of $\pm 3\sigma$. According to ITRS'06 [12], the physical gate length for high performance logic in 45nm technology will be 18nm, and the physical variation should be controlled within +/-12%. Therefore, we let 3σ be 12%. While it is similar for T_{ox}.

For a gate/module in a chip with equivalent length (ΔL), using our model parameters in Table 1, we have:

$$\Delta L = \Delta L_{inter} + \Delta L_{intra\_corr} \tag{2}$$

ΔL_{inter} is constant for all cells in all grids because it is a global factor that applies to the whole chip. For one chip sample, we only need to generate it once. ΔL_{intra_corr} is different between each gate(our method) or each grid(method in [4]), and has spatial correlations. Therefore, we generate one value for each gate(our method)/grid(method in [4]), and the spatial correlation is modeled such that the correlation coefficient value diminishes equally with the distance between any two gates/grids.

2.2 Static leakage modeling for gates

Full-chip leakage current has two components, subthreshold and gate leakage currents. Here we describe the empirical models for them, based on which the leakage current under process variations is estimated under lognormal distributions.

The subthreshold leakage current, I_{sub} is exponentially dependent on the threshold voltage, V_{th}, and V_{th} is observed to be most sensitive to gate oxide thickness T_{ox} and effective gate channel length L due to short-channel effects. When the change in L or T_{ox} is small, the precise relationship shows an exponential dependent effect on I_{sub}, with the effect of T_{ox} relatively weak. For the gate oxide leakage current, both channel length and oxide thickness have strong impacts on the leakage currents, which are exponential functions of the two variables.

In our work, we also follow the analytical expressions given in [4], which estimate the subthreshold leakage currents and the gate oxide leakage currents as follows:

$$I_{sub} = e^{a_1 + a_2 L + a_3 L^2 + a_4 T_{ox}^{-1} + a_5 T_{ox}} \tag{3}$$

$$I_{gate} = e^{a_1 + a_2 L + a_3 L^2 + a_4 T_{ox} + a_5 T_{ox}^2} \tag{4}$$

where a_1 through a_5 are the fitting parameters for each unique input combination of a gate. Then we can use a look-up table (LUT) to store the fitting parameters, and for a k-input gate, the size of the LUT is 2^k as we have two equations for each input combination. However, we observed the I_{sub} based on the model in (3) can still have large errors compared to the simulation results. Table 2 shows that the errors compared with industry SPICE simulation results for a AND gate for I_{sub}. *Err.* is the error for one input combination and *Avg Err.* refers the average errors over all the input patterns. If we add more terms into (3) as shown in the table, we can reduce the errors from 8% to about 3%.

Table 2: Relative errors by using different fitting formulas for leakage currents of a AND gate

Fitting components	Err.	Avg Err.
Original: $L, L^2, T_{ox}^{-1}, T_{ox}$	14.7%	8.46%
Add T_{ox}^2	13.95%	8.26%
Add $T_{ox}^2, T_{ox}/L$	7.08%	5.95%
Add T_{ox}^2, $T_{ox}/L, L/T_{ox}$	7.14%	4.94%
Add $T_{ox}^2, T_{ox}/L$, $L/T_{ox}, T_{ox} * L$	3.67%	3.49%

After we obtain the analytic expression for each input combination, we take the average of the leakage currents of all the input combinations to arrive final analytic expression for each gate in lieu of the dominant states used in [4].

3. REVIEW OF THE ORTHOGONAL POLYNOMIAL METHOD

In the following, we briefly review the orthogonal polynomial based modeling approaches. Note that for the Gaussian and log-normal distributions, Hermite polynomial is the best choice as it leads to exponential convergence rate [9]. For non Gaussian and non log-normal distributions, there are other orthogonal polynomials such as Legendre for uniform distribution, Charlier for Poisson distribution and Krawtchouk for Binomial distribution, etc [8, 23]. The proposed method can be extended to other distributions with different orthogonal polynomials [26].

3.1 Concept of Hermite polynomial chaos

Hermite polynomial chaos (PC) utilizes a series of orthogonal polynomials (with respect to the Gaussian distribution) to facilitate stochastic modeling [26]. These polynomials are used as the orthogonal basis to decompose a random process in the similar way as sine and cosine functions are used to decompose a periodic signal in Fourier series expansion.

Homogeneous chaos expansion is guaranteed to converge for any Gaussian random process with finite second-order moments [9]. Moreover, the Askey principle [27] shows that the expansion based on Hermite polynomials has the optimal convergence rate for a Gaussian random process. Also, the advantage of homogeneous chaos expansion over the traditional Taylor expansion is clearly demonstrated in recent related works [24, 23].

For a random variable $v(t, \xi)$ with limited variance, where $\xi = [\xi_1, \xi_2, ...\xi_n]$ is a vector of independent orthonormal Gaussian random variables with zero mean. The random variable can be approximated by truncated Hermite PC expansion as follows [9]:

$$x(\xi) = \sum_{k=0}^{P} a_k H_k^n(\xi) \tag{5}$$

where $H_k^n(\xi)$ is nth order Hermite polynomials and a_k is the deterministic coefficient. The number of terms P is given by

$$P = \sum_{k=0}^{p} \frac{(n-1+k)!}{k!(n-1)!} \tag{6}$$

where p is the order of the Hermite PC. When one random variable is considered, one-dimensional Hermite polynomials are expressed as follows:

$$H_0^1(\xi) = 1, H_1^1(\xi) = \xi, H_2^1(\xi) = \xi^2 - 1, H_3^1(\xi) = \xi^3 - 3\xi, ... \tag{7}$$

Hermite polynomials are orthogonal with respect to Gaussian weighted expectation (the superscript n is dropped for simple notation):

$$< H_i(\xi), H_j(\xi) >=< H_i^2(\xi) > \delta_{ij} \tag{8}$$

where δ_{ij} is the Kronecker delta and <*,*> denotes an inner product defined as follows:

$$< f(\xi), g(\xi) >= \frac{1}{\sqrt{(2\pi)^n}} \int f(\xi) g(\xi) e^{-\frac{1}{2}\xi^T \xi} d\xi \tag{9}$$

Like Fourier series, the coefficient a_k can be found by a projection operation onto the Hermite PC basis:

$$a_k(t) = \frac{< x(\xi), H_k(\xi) >}{< H_k^2(\xi) >}, \ \forall k \in \{0, ..., P\}. \tag{10}$$

In our problem, $x(\xi)$ will be the leakage current for each gate and for the full chip eventually.

4. THE PROPOSED METHOD

In this section, we will present the new full-chip statistical leakage analysis method. The algorithm flow is shown in Fig. 1.

Algorithms: NEW FULL-CHIP LEAKAGE CURRENT COMPUTATION ALGORITHM

Input: standard cell lib, netlist, placement information of design, σ of L and T_{ox}
Output: analytic expression of the full-chip leakage currents in terms of Hermite polynomials.

1. Generate fitting parameter matrices a_{sub} and a_{gate} of I_{sub} and I_{gate} in (3) and (4) for each type of gates (after SPICE run on each input pattern) (Section 2).
2. Perform PCA to transform and reduce the original parameter variables in $\mathbf{L}$ into independent random variables in $\mathbf{L_k}$. (Section 3).
3. Generate Smolyak spare grid point set Θ_n^2 with corresponding weights.
4. Calculate the coefficients of Hermite polynomial of $I_{sub,k}$ and $I_{gate,k}$ for the final leakage analytic expression for each gate using (26) and (27).
5. Calculate the analytic expression of the full-chip leakage current by simple polynomial additions and calculate $\mu_{leakage}$, $\sigma_{leakage}$, PDF and CDF of the leakage current if required.

Figure 1: The flow of proposed algorithm.

The new algorithm basically consists of three major parts. The first part (step 1) is pre-characterization, which builds the analytic leakage expressions (3) and (4) for each type of gates. The second part (step 2-5) generates a set of independent random variables and builds the gate-level analytic leakage current expressions and covariances. The final part (step 6) computes the final leakage expressions by simple polynomial additions and calculates other statistical information.

4.1 Gaussian quadrature technique

The Gaussian quadrature method is an efficient numerical method to compute the definite integral of a function [11]. We apply it to compute the coefficients $a_k(t)$ in (10). We review the method based on the Hermite polynomial below.

Our goal is to compute the integral equation $< x(\xi), H_j(\xi) >$ numerically. In this case, this problem boils down to the one-dimensional numerical quadrature problem based on the

Hermite polynomials [18]. Specifically, for Hermite polynomials, we have

$$< x(\xi), H_k(\xi) > = \frac{1}{\sqrt{(2\pi)}} \int x(\xi) H_k(\xi) e^{-\frac{1}{2}\xi^2} d\xi \quad (11)$$

$$\approx \sum_{i=0}^{P} x(\xi_i) H_i(\xi_i) w_i \quad (12)$$

Here, $\xi = \{\xi\}$, contains only one random variable. ξ_i and w_i are Gaussian Hermite quadrature abscissas (quadrature points) and weights.

The Quadrature rule basically says that if we select the roots of Pth Hermite polynomial as the quadrature points, the quadrature is exact for all polynomials of degree $2P-1$ or less for (11). This is called (P-1)-level accuracy of Gaussian-Hermite quadrature here.

For multiple random variables, which require multi-dimensional quadrature, traditional way to compute the multi-dimensional quadrature is to use a direct tensor product based on one dimensional Gaussian Hermite quadrature abscissas and weights [16]. With this method, the number of quadrature points needed for n dimension (variables) and Pth level is about $(P+1)^n$, which is well known as the curse-of-dimensionality.

4.2 Smolyak quadrature for multi-dimensional integration

Smolyak quadrature [16] is used as an efficient method to reduce the number of quadrature points (also called sparse grid quadrature). Let's define one-dimensional sparse grid quadrature point set $\Theta_1^P = \{\gamma_i, \gamma_2, ..., \gamma_P\}$, which uses $P+1$ points to achieve degree $2P+1$ of exactness. The level-P sparse grid for n-dimensional quadrature chooses points from the following set:

$$\Theta_n^P = \bigcup_{P+1 \le |\vec{i}| \le P+n} (\Theta_1^{i_1} \times ... \times \Theta_1^{i_n}) \quad (13)$$

where $|\vec{i}| = \sum_{j=1}^{n} i_j$. And the corresponding weight is:

$$w_{j_{i_1}...j_{i_n}}^{i_1...i_n} = (-1)^{P+n-|\vec{i}|} \begin{pmatrix} n-1 \\ n+P-|\vec{i}| \end{pmatrix} \prod_m w_{j_{i_m}}^{i_m} \quad (14)$$

where $\begin{pmatrix} n-1 \\ n+P-|\vec{i}| \end{pmatrix}$ is a combination number and w is the weight for the corresponding quadrature points. It was shown that interpolation on a Smolyak grid ensures an error bound for the mean-square error [16]

$$|E_P| = O(N_P^r (log N_P)^{(r+1)(n-1)}),$$

where N_P is the number of quadrature points and k is the order of the maximum derivative that exist for the delay function. The number of quadrature points increases as $O(\frac{n^P}{(P)!})$.

It can be shown that sparse grid of at least level P is required for an order P representation. The reason is that the approximation contains order P polynomials for both $x(\xi)$ and $H_j(\xi)$ for some j. So there exists $x(\xi)H_j(\xi)$ with order $2P$, which requires sparse grid of at least level P with degree $2P+1$ of exactness.

Therefore, level 2 and level 1 sparse grid are required for quadratic and linear model, respectively. The number of quadrature points is about $2n$ and $2n^2$ for the linear and the quadratic model respectively. The time cost is about the same as the Taylor-conversion method, while keeping the accuracy of homogenous chaos expansion.

In addition to the sparse grid technique, we also employ several accelerating techniques summarized as follows:

- When n is too small, the number of quadrature points for sparse grid may be larger than that of direct tensor product of Gaussian quadrature. For example, if there are only 2 variables, the number is 5 and 14 for level 1 and 2 sparse grid, compared to 4 and 9 for direct tensor product. In this case, the sparse grid will not be used.
- The set of quadrature points (13) may contain the same points with different weights. For example, the level 2 sparse grid for 3 variables contain 4 instances of the point (0,0,0). Combining these points by summing the weights reduces 3 times of computation of $x(\vec{\gamma}_i)$.

4.3 Random variables transformation and reduction

In our gate-based approach, instead of using grid-based partitioning, as in [4], to reduce the number of channel length variables in presence of the strong spatial correlation, we applied the principal component analysis (PCA) to reduce the number of random variables. Our method starts with the following random variable vectors

$$\mathbf{L} = [L_1, L_2, ...L_n] + \delta L_{inter} \quad (15)$$

$$\mathbf{T_{ox}} = [T_{ox1}, T_{ox2}, ..., T_{oxn}] + \delta T_{ox,inter} \quad (16)$$

where δL_{inter} and $\delta T_{ox,inter}$ represent the inter-die (global) variations. In total, we have $2n+2$ random variables. There exist correlations between L among different gates, represented by the covariance matrix $cov(L_i, L_j)$ computed by (1).

The first step is to perform PCA on L to get a set of independent random variables $\mathbf{L'} = [L'_1, L'_2, ...L'_n]$, where $\mathbf{L} = P\mathbf{L'}$, and $P = \{p_{ij}\}$ is the n by n principal component coefficient matrix.

We make sure that the elements in $\mathbf{L'}$ are arranged in a decreasing weight order. Then the number of elements in $\mathbf{L'}$ can be reduced by only considering the dominant part of $\mathbf{L'}$ as $[L'_1, L'_2, ...L'_k]$ (for instance, the weight should be bigger than 1%), where k is the number of reduced random variables. Every element L'_i in $\mathbf{L'}$ can be then represented by orthogonal Gaussian random variable ξ_i with normal distribution.

$$L'_i = \mu_i + \sigma_i \xi_i. \quad (17)$$

where μ_i and σ_i are the mean value and standard deviation of L'_i. And $\mathbf{L}$ can be represented as

$$\mathbf{L} = \begin{pmatrix} \mu_{L1} \\ \mu_{L2} \\ \vdots \\ \mu_{Ln} \end{pmatrix} + \begin{pmatrix} p_{11} & \dots & p_{1k} \\ p_{21} & \dots & p_{2k} \\ \vdots & \vdots & \vdots \\ p_{n1} & \dots & p_{nk} \end{pmatrix} \begin{pmatrix} \sigma_1\xi_1 \\ \sigma_2\xi_2 \\ \vdots \\ \sigma_k\xi_k \end{pmatrix} + \delta L_{inter} \quad (18)$$

For $[T_{ox1}, T_{ox2}, ..., T_{oxn}]$, δL_{inter} and $\delta T_{ox,inter}$, we can also represent them using the standard Gaussian variables as

$$T_{ox,j} = \mu_{ox,j} + \sigma_{ox,j}\xi_{ox,j} \quad (19)$$

$$\delta L_{inter} = \sigma_{L,inter}\xi_{L,inter} \quad (20)$$

$$\delta T_{ox,inter} = \sigma_{ox,inter}\xi_{ox,inter} \quad (21)$$

where $\xi_{ox,j}$, $\xi_{L,inter}$ and $\xi_{ox,inter}$ are independent orthonormal Gaussian random variables. As a result, we can present $\mathbf{L}$ and $\mathbf{T_{ox}}$ by $k+n+2$ independent orthonormal Gaussian random variables.

$$\xi = [\xi_1, \xi_2, ..., \xi_{k+n+2}] \quad (22)$$

Then the $I_{sub}(\mathbf{L}, \mathbf{T_{ox}})$ and $I_{gate}(\mathbf{L}, \mathbf{T_{ox}})$ can be modeled as $I_{sub}(\xi)$ and $I_{gate}(\xi)$, respectively.

But among the $k+n+2$ variables, only $k+2$ variables related to the channel lengths are correlated. In other words, the n variables $T_{ox,i}$ of each gate are independent. As a result, for the jth gate, we only have $k+3$ independent

variables, the corresponding variable vector, $\xi_g = \{\xi_{g,j}\}$, is defined as

$$\xi_{g,j} = [\xi_1, ..., \xi_k, \xi_{ox,j}, \xi_{L,inter}, \xi_{ox,inter}] \quad (23)$$

4.4 Computation of full-chip leakage currents

For each gate, we need to present the leakage currents in order-2 Hermite polynomials first as shown below for both subthreshold and gate leakage currents – $I_{sub}(\xi_{\mathbf{g,j}})$ and $I_{gate}(\xi_{\mathbf{g,j}})$:

$$I_{sub}(\xi_{\mathbf{g,j}}) = \sum_{i=0}^{P} I_{sub,i,j} H_i^2(\xi_{\mathbf{g,j}}) \quad (24)$$

$$I_{gate}(\xi_{\mathbf{g,j}}) = \sum_{i=0}^{P} I_{gate,i,j} H_i^2(\xi_{\mathbf{g,j}}) \quad (25)$$

where $H_i^2(\xi_{\mathbf{g,j}})$s are order-2 Hermite polynomials. $I_{sub,i,j}$ and $I_{gate,i,j}$ are then computed by the numerical Gaussian quadrature method discussed in Subsections 4.1 and 4.2. Let S be the size of Z-dimensional second order (level-2) quadrature point set Θ_Z^2 and $Z = k + 3$. Then $I_{sub,i}$ and $I_{gate,i}$ can be computed as the following:

$$I_{sub,i,j} = \sum_{l=1}^{S} I_{sub}(\vec{\gamma_l}) H_i^2(\vec{\gamma_l}) w_l / < H_i^2(\xi_{\mathbf{g,j}}) > \quad (26)$$

$$I_{gate,i,j} = \sum_{l=1}^{S} I_{gate}(\vec{\gamma_l}) H_i^2(\vec{\gamma_l}) w_l / < H_i^2(\xi_{\mathbf{g,j}}) > \quad (27)$$

where $I_{sub}(\vec{\gamma_l})$ and $I_{gate}(\vec{\gamma_l})$ are computed using (3) and (4).

As a result, their coefficients for ith Hermite polynomial at jth gate can be added directly as

$$I_{leakage,i,j} = \sum I_{sub,i,j} + \sum I_{gate,i,j} \quad (28)$$

After the leakage currents are calculated for each gate, we can proceed to compute the leakage current for the whole chip as the follows:

$$I_{leakage}(\xi) = \sum_{j=1}^{n} (I_{sub}(\xi_{\mathbf{g,j}}) + I_{gate}(\xi_{\mathbf{g,j}})) \quad (29)$$

The summation is done for each coefficient of Hermite polynomials. Then we obtain the analytic expression of the final leakage currents in terms of the ξ.

We can then obtain the mean, variance of the leakage currents, PDF and CDF very easily. For instance, the mean value and variance for the full-chip leakage current are

$$\mu_{leakage} = I_{leakage,0th} \quad (30)$$

$$\sigma_{leakage}^2 = \sum I_{leakage,1st}^2 + 2 \sum I_{leakage,2nd,type1}^2 + \sum I_{leakage,2nd,type2}^2 \quad (31)$$

where $I_{leakage,ith}$ is the leakage coefficient for ith Hermite polynomial of second order defined as follows.

$$H_{0th}(\xi) = 1, H_{1st}(\xi) = \xi_i, H_{2nd,type1}(\xi) = \xi_i^2 - 1,$$
$$H_{2nd,type2}(\xi) = \xi_i \xi_j, i \neq j \quad (32)$$

4.5 Time complexity analysis

To analyze the time complexity, one typically do not count the pre-characterization cost of step 1 in Fig. 1. For PCA step (step 2), which essentially uses singular value decomposition (SVD) on the covariance matrix, its computation cost is $O(nk^2)$, if we only interested in the first k dominant singular values. This is the case for strong spatial correlation.

In step 3 and 4, we need to call (3) and (4) S times for each gate. In each call, we need to compute $k + 3$ variables in the Hermite polynomials. The computing cost for the two steps is $O(n(k + 3) * S)$, where n is number of gates. After the leakage currents are computed for each gate, it takes $O(n(k + 3))$ to compute the full-chip leakage current.

The total computing cost is $O(nk^2 + n(k+3)S + n(k+3))$. For second order Hermite polynomials, $S \propto k^2$, so the time complexity becomes $O(nk^3)$. If $k \ll n$ (for strong spatial correlation), we end up with a linear time complexity O(n).

5. EXPERIMENTAL RESULTS

The proposed method has been implemented in Matlab 7.0. For comparison purpose, we also implemented the grid-based method in [4] and the pure Monte-Carlo method. All the experimental results are carried out in a Linux system with quad Intel Xeon CPUs with 2.99Ghz and 16GB memory.

The methods for full-chip statistical leakage estimation were tested on circuits in the PDWorkshop91 benchmark set. The circuits were synthesized with Nangate Open Cell Library and the placement is from MCNC [1]. The technology parameters come from the 45nm FreePDK Base Kit and PTM models [19].

The 3σ values of parameter variations for L and T_{ox} were set to 12% of the nominal parameter values, of which inter-die variations constitute 20% and intra-die variations, 80%. The parameter L is modeled as sum of correlated sources of variations, and the gate oxide thickness T_{ox} is modeled as an independent source of variation. The same framework can be easily extended to include other parameters of variations. Both L and T_{ox} in each gate are modeled as Gaussian parameters. For the correlated L, the spatial correlation was modeled based on the exponential special correlation in (1). For [4], we still partition the chip into a number of regular grids and the numbers of grid partitions of spatial correlation model used for the benchmarks are given in Table 1. For comparison purposes, we performed Monte Carlo (MC)

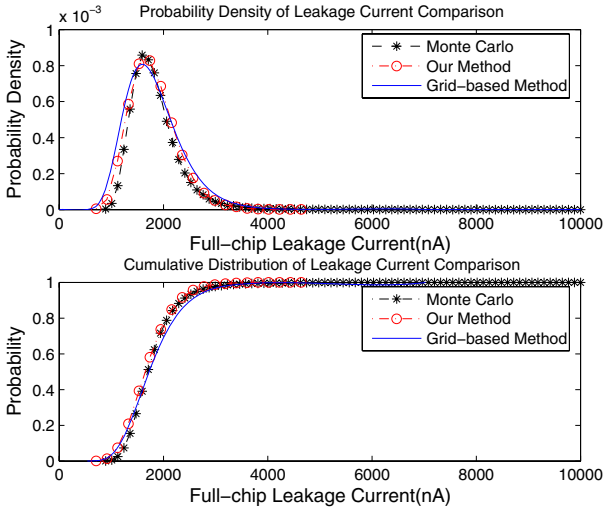

Figure 2: Distribution of the total leakage currents of the proposed method and the grid-based method and the MC method for circuit SC0

simulations with 50,000 runs, the grid-based method in [4], and the new method on the benchmarks. The large number of MC runs are due to the fact that proposed method is quite accurate. Fig. 2 shows the full-chip leakage current distribution (PDF and CDF) of circuit SC0 with 125 gates, considering variation in gate length and gate oxide thickness as in Table 1, and the spatial correlation of gate length. It shows that our method fits very well with the MC results, and is more accurate than [4].

Table 3: Comparison of the mean values of full-chip leakage currents among three methods.

Circuit	*Gate*	*Grid*	*Mean Value (μA)*			*Errors (%)*	
Name	*#*	*#*	*MC*	[4]	*New*	[4]	*New*
SC0	125	4	1.84	1.75	1.82	-4.71	-1.11
SC2	1888	16	29.98	28.89	29.71	-3.7	-0.91
SC5	6417	64	107.88	103.6	107.18	-3.9	-0.65

Table 4: Comparison standard deviations of full-chip leakage currents among three methods

Circuit	*Standard Deviation(μA)*			*Errors (%)*	
Name	*MC*	[4]	*New*	[4]	*New*
SC0	0.562	0.668	0.524	34.9	-5.77
SC2	8.606	10.86	7.332	26.2	-1.25
SC5	26.19	41.36	25.11	57.9	-4.12

The results of the comparison of mean value and standard deviations of full-chip leakage current are shown in Table 3 and Table 4. The average errors for mean and standard variance values of the new gate-based method are 0.54% and 4.10%, respectively. While for the grid-based method in [4], the average errors for mean and sigma value are 3.94% and 39.7%, respectively.

And Table 5 also compares the CPU times of the three methods, which shows that our method is much faster than the method in [4] and the Monte Carlo method. On average, the proposed method has about 25X speedup over the grid based method in [4]. We notice that method in [4] will become faster with smaller number of grids used. But this can lead to large errors even with strong spatial correlations.

Table 5: CPU time comparison among three methods.

Circuit	*Cost time(s)*			*Speedup (%)*	
Name	*MC*	[4]	*New*	[4]	*New*
SC0	757.2	76.11	1.52	9.9	498.2
SC2	5171.7	168.51	18.79	30.6	275.2
SC5	8.09×10^4	767.06	121.2	105.5	667.5

6. CONCLUSION

In this paper, we have presented a novel method for analyzing the full-chip leakage current distribution of digital circuit. The new method considers both intra-die and inter-die variations with spatial correlations. The new method employs the orthogonal polynomials and multi-dimensional Gaussian quadrature method to represent and compute variational leakage at the gate level and use the orthogonal decomposition to reduce the number of random variables exploiting the strong spatial correlations of intra-die variations. The resulting algorithm compares very favorable with the existing grid-based method in terms of both CPU time and accuracy. The new method has about 10X speedup over [4] with constant better accuracy.

7. REFERENCES

[1] "MCNC benchmark circuit placements," http://vlsicad.ucsd.edu/GSRC/bookshelf/Slots/Placement/.

[2] A. H. ane J. Karhunen and E. Oja, *Independent Component Analysis*. Wiley, 2001.

[3] S. Borkar, T. Karnik, and V. De, "Design and reliability challenges in nanometer technologies," in *Proc. Design Automation Conf. (DAC)*. IEEE Press, 2004.

[4] H. Chang and S. S. Sapatnekar, "Full-chip analysis of leakage power under process variations, including spatial correlations," in *Proc. Design Automation Conf. (DAC)*. New York, NY, USA: ACM, 2005, pp. 523–528.

[5] V. De and S. Borkar, "Technology and design challenges for low power and high performance," in *Proc. Int. Symp. on Low Power Electronics and Design(ISLPED)*, Aug. 1999, pp. 163–168.

[6] S. G. Duvall, "Statistical circuit modeling and optimization," in *Intl. Workshop Statistical Metrology*, Jun 2000, pp. 56–63.

[7] P. Friedberg, "Modeling within-die spatial correlation effects for process design co-optimization," in *Proceedings of the 6th International Symposium on Quality of Electronic Design*, 2005, pp. 516–521.

[8] R. Ghanem, "The nonlinear Gaussian spectrum of log-normal stochastic processes and variables," *Journal of Applied Mechanics*, vol. 66, pp. 964–973, December 1999.

[9] R. G. Ghanem and P. D. Spanos, *Stochastic Finite Elements: A Spectral Approach*. Dover Publications, 2003.

[10] K. R. Heloue, N. Azizi, and F. N. Najm, "Modeling and estimation of full-chip leakage current considering within-die correlation," in *Proc. Design Automation Conf. (DAC)*. New York, NY, USA: ACM, 2007, pp. 93–98.

[11] A. Iserles, *A First Course in the Numerical Analysis of Differential Equations*, 3rd ed. Cambridge University, 1996.

[12] "Semiconductor industry association. international technology roadmap for semiconductors," 2006, http://www.itrs.net/Links/2006Update/2006UpdateFinal.htm.

[13] 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. on Low Power Electronics and Design(ISLPED)*. New York, NY, USA: ACM, 2003, pp. 172–175.

[14] S. Narendra, V. De, S. Borkar, D. A. Antoniadis, and A. P. Chandrakasan, "Full-chip subthreshold leakage power prediction and reduction techniques for sub-0.18-μm CMOS," *IEEE J. Solid-State Circuits*, vol. 39, no. 2, 2004.

[15] S. Nassif, "Delay variability: sources, impact and trends," in *Proc. IEEE Int. Solid-State Circuits Conf.*, San Francisco, CA, Feb 2000, pp. 368–369.

[16] E. Novak and K. Ritter, "Simple cubature formulas with high polynomial exactness," *Constructive Approximation*, vol. 15, no. 4, pp. 449–522, Dec 1999.

[17] M. Orshansky, L. Milor, and C. Hu, "Characterization of spatial intrafield gate cd variability, its impact on circuit performance, and spatial mask-level correction," in *IEEE Trans. on Semiconductor Devices*, vol. 17, Feb 2004, pp. 2–11.

[18] W. H. Press, S. A. Teukolsky, W. T. Veterling, and B. P. Flannery, *Numerical Recipes in C: The art of Scientific Computing*. Cambridge University Press, 1992.

[19] "Predictive technology model," http://www.eas.asu.edu/ ptm/.

[20] R. Rao, A. Srivastava, D. Blaauw, and D. Sylvester, "Statistical analysis of subthreshold leakage current for VLSI circuits," *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, vol. 1, no. 2, pp. 131–139, Feb 2004.

[21] A. Srivastava, R. Bai, D. Blaauw, and D. Sylvester, "Modeling and analysis of leakage power considering within-die process variations," in *Proc. Int. Symp. on Low Power Electronics and Design(ISLPED)*, Aug. 2002, pp. 64–67.

[22] R. Teodorescu, B. Greskamp, J. Nakano, S. R. Sarangi, A. Tiwari, and J. Torrellas, "A model of parameter variation and resulting timing errors for microarchitects," in *Workshop on Architectural Support for Gigascale Integration (ASGI)*, Jun 2007.

[23] S. Vrudhula, J. M. Wang, and P. Ghanta, "Hermite polynomial based interconnect analysis in the presence of process variations," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, vol. 25, no. 10, 2006.

[24] J. Wang, P. Ghanta, and S. Vrudhula, "Stochastic analysis of interconnect performance in the presence of process variations," in *Proc. Int. Conf. on Computer Aided Design (ICCAD)*, Nov 2004, pp. 880–886.

[25] J. Xiong, V. Zolotov, and L. He, "Robust extraction of spatial correlation," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, vol. 26, no. 4, 2007.

[26] D. Xiu and G.Karniadakis, "Modeling uncertainty in flow simulations via generalized polynomial chaos," *J. of Computational Physics*, no. 187, pp. 137–167, 2003.

[27] D. Xiu and G. Karniadakis, "The wiener-askey polynomial chaos for stochastic differential equations," *SIAM J.Sci.Comput*, vol. 24, no. 2, pp. 619–644, Oct 2002.