# Modeling of Intra-die Process Variations for Accurate Analysis and Optimization of Nano-scale Circuits

Sarvesh Bhardwaj, Sarma Vrudhula, Praveen Ghanta, and Yu Cao
Arizona State University, Tempe, AZ
{sarvesh.bhardwaj,vrudhula,praveen.ghanta,yu.cao}@asu.edu

## ABSTRACT

This paper proposes the use of Karhunen-Loève Expansion (KLE) for accurate and efficient modeling of intra-die correlations in the semiconductor manufacturing process. We demonstrate that the KLE provides a significantly more accurate representation of the underlying stochastic process compared to the traditional approach of dividing the layout into grids and applying Principal Component Analysis (PCA). By comparing the results of leakage analysis using both KLE and the existing approaches, we show that using KLE can provide up to $4-5\times$ reduction in the variability space (number of random variables) while maintaining the same accuracy. We also propose an efficient leakage minimization algorithm that maximizes the leakage yield while satisfying probabilistic constraints on the delay.

**Categories and Subject Descriptors:** B.7.2 [Integrated Circuits]: Design Aids—Simulation, Verification

**General Terms:** Algorithms, Design, Performance

**Keywords:** Statistical, Process Variations, Leakage, Karhunen-Loeve, intra-die, correlations

## 1. INTRODUCTION

The lack of absolute control in the lithography as well as channel doping steps during the manufacturing of nano-scale circuits results in a significant amount of variability in the characteristics of manufactured devices [1–3]. This necessitates the need for techniques that model the parameters as stochastic quantities rather than existing methodologies that treat them as deterministic quantities. To this extent, a number of techniques [4–6] to name a few, have been proposed that model the parameters as random variables and estimate the *probability density functions* (PDFs) of circuit delay and leakage power in the presence of variations using PDFs of the parameters. In addition, a number of statistical optimization methods have also been proposed [7–10]. These methods span the domain of both continuous and discrete optimization techniques and either minimize the area [7, 10] or maximize the parametric yield of power [8] subject to timing yield constraints. In [9], the authors maximize the joint yield of leakage and delay using gate sizing.

In the presence of intra-die variations, the parameter of each device on a die has to be treated as a different random variable due to the presence of a random component (e.g. random dopant fluctuations in threshold voltage $V_t$). Since there can be potentially millions of transistors on a die, any analysis considering millions of correlated random variables will be computationally prohibitive. Hence, to account for intra-die variations, the conventional technique is to divide the die into a number of grids, say $r$, and assume that the parameter such as the gate length $L_e$ of the devices lying in the same grid are realizations of the *same* random variable [12], thus reducing the number of correlated random variables to $r$. This set of correlated random variables is transformed into a new set of uncorrelated random variables using Principal Component Analysis (PCA). Even though this approach reduces the dimensionality of the variability space, it poses an interesting question: What value of $r$ should one choose in order to accurately capture the correlations? In addition, for parameters such as $V_t$ that have a considerable amount of random variations, a large value of $r$ may be required to accurately model the correlations. Hence, different grid resolutions will be required to capture the correlations in different parameters ($L_e$ or $V_t$) with same accuracy.

The above mentioned drawbacks of existing *grid-based* approach can be addressed through the use of the Karhunen-Loève expansion (KLE) [11]. In this approach, the parameters of the devices are modeled as stochastic processes over the spatial domain of a die, thus making parameters of any two devices on the die, two different (correlated) random variables. Given a covariance function, the KLE provides a mechanism to write the stochastic process as a series expansion of some *uncorrelated* random variables whose coefficients can be obtained by solving an integral equation involving the covariance function. This obviates the need of a covariance matrix and the complexity associated with decorrelating it. In addition, as will be shown in the paper, the number of random variables required for representing the stochastic process with the same degree of accuracy is significantly smaller for the KLE than that required in the grid-based approach.

The organization of the rest of the paper is as follows: Section 2 formally describes the modeling of process variations, the KLE and an example are described in Sections 3. Section 4 describes the leakage minimization problem considering intra-die correlations using KLE. The Experimental


*DAC 2006,* July 24–28, 2006, San Francisco, California, USA.

results and conclusions are outlined in Section 5 and 6 respectively.

## 2. PROBLEM FORMULATION

Let $\Omega$ be the set of all possible manufacturing outcomes and $\boldsymbol{D} = D_X \times D_Y = [-a, a] \times [-b, b]$ be the the physical size of the die under consideration. In the presence of process variations, the *manufactured* values of various device parameters such as gate length $L_e$ and the threshold voltage $V_{th}$ are a function of the location $\boldsymbol{x} = (x, y) \in \boldsymbol{D}$ of the device on the die and the die on which the device lies $\theta \in \Omega$. Thus, each parameter $p(\boldsymbol{x}, \theta)$ can be modeled as a stochastic process. Without the loss of generality, we assume that $p(\boldsymbol{x}, \theta)$ is a zero-mean, unit-variance process. Given the spatial covariance function $\boldsymbol{C}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ of the parameter $p(\boldsymbol{x}, \theta)$, the common approach [4,12] is to model $p(\boldsymbol{x}, \theta)$ using a discrete set of random variables $\{p_k\}$, $k = 1, \ldots, r$. This is done by mapping each $\boldsymbol{x} \in \boldsymbol{D}$ to a unique $k$, with multiple $\boldsymbol{x}$'s mapped to the same $k$. Let $g : \boldsymbol{D} \rightarrow \{1, \ldots, r\}$ define this mapping. One such mapping is shown in Fig. 1, where $r = 4$ and $g(-2.2, -1.5) = 2$. Now, the covariance matrix $\boldsymbol{C}_d$ for $\{p_k\}$ needs to be obtained from the covariance function $\boldsymbol{C}(\boldsymbol{x}_1, \boldsymbol{x}_2)$. One way to do this is by assigning each element $c_{ij}$ of the covariance matrix $\boldsymbol{C}_d$ using $c_{ij} = \boldsymbol{C}(\boldsymbol{x}_{ci}, \boldsymbol{x}_{cj})$, where $\boldsymbol{x}_{ci}$ and $\boldsymbol{x}_{cj}$ are the centers of partition $i$ and $j$ respectively. This correlated set of random variables $\{p_k\}$ can then be transformed into a set of mutually orthogonal random variables $\{p'_k\}$ using Principal Component Analysis (PCA) [13]. Thus, each random variable in the original set can be written as

$$p_k = \sum_{j=1}^{r} \sqrt{\lambda_j} v_{kj} p'_j, \quad k = 1, \ldots, r \tag{1}$$

where $\lambda_j$ is the $j$-th eigenvalue of $\boldsymbol{C}_d$ and $v_{kj}$ is the $k$-th element of the $j$-th eigenvector of $\boldsymbol{C}_d$. The accuracy of this approach improves with increasing $r$.

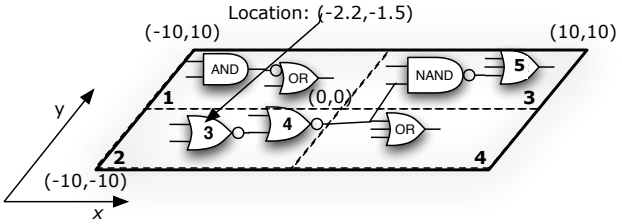



**Figure 1: Grid-based approach to model correlations**

An alternative is to represent the whole process $p(\boldsymbol{x}, \theta)$ by a Fourier-series type representation

$$p(\boldsymbol{x}, \theta) = \sum_{n=1}^{\infty} \sqrt{\lambda_n} \xi_n(\theta) f_n(\boldsymbol{x}), \tag{2}$$

where $\{\xi_n(\theta)\}$ is a set of random variables, $\lambda_n$ are some constants, and $\{f_n(\boldsymbol{x})\}$ is an orthonormal set of deterministic functions. In this work, we provide the conditions under which such an expansion can be obtained and given particular covariance functions $\boldsymbol{C}(\boldsymbol{x}_1, \boldsymbol{x}_2)$, how to find the constants $\lambda_n$ and the functions $\{f_n(\boldsymbol{x})\}$.

## 3. KARHUNEN-LOÈVE EXPANSION

DEFINITION 3.1. *[11] A second-order random variable $p(\theta)$ is one which satisfies $E[|p(\theta)|^2] < \infty$. A second order stochastic process $p(\boldsymbol{x}, \theta)$ is a family of second-order random variables.*

The random variables in our problem are the values of the physical parameters. Hence, their values are bounded above and below. Thus we can assume that the process under consideration $p(\boldsymbol{x}, \theta)$ is a second-order stochastic process.

DEFINITION 3.2. *[11] A second-order process $p(\boldsymbol{x}, \theta)$ is continuous in quadratic mean (q.m.) if $E[|p(\boldsymbol{x} + \boldsymbol{h}, \theta) - p(\boldsymbol{x}, \theta)|^2] \rightarrow 0$ as $\|\boldsymbol{h}\| \rightarrow 0$ for all $\boldsymbol{x} \in \boldsymbol{D}$.*

In the above definition, $\|\cdot\|$ is the Euclidean norm. Since, we do not have an explicit representation of $p(\boldsymbol{x}, \theta)$, we make use of the following result to guarantee its q.m. continuity.

THEOREM 3.1. *[11] A second-order process $p(\boldsymbol{x}, \theta)$ is continuous in q.m. at $\boldsymbol{x} \in \boldsymbol{D}$ if and only if, its covariance function $\boldsymbol{C}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ is continuous at $(\boldsymbol{x}, \boldsymbol{x})$.*

DEFINITION 3.3. *The functions $\{f_n(\boldsymbol{x})\}$ are eigenfunctions of the covariance kernel $\boldsymbol{C}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ if they satisfy the integral equation*

$$\int_{\boldsymbol{D}} \boldsymbol{C}(\boldsymbol{x}_1, \boldsymbol{x}_2) f_n(\boldsymbol{x}_1) d\boldsymbol{x}_1 = \lambda_n f_n(\boldsymbol{x}_2). \tag{3}$$

*The constants $\lambda_n$ are called the eigenvalues of $\boldsymbol{C}(\boldsymbol{x}_1, \boldsymbol{x}_2)$.*

Since the only information we are given regarding the process $p(\boldsymbol{x}, \theta)$ is its covariance function $\boldsymbol{C}(\boldsymbol{x}_1, \boldsymbol{x}_2)$, we need to relate the eigenfunctions and eigenvalues of $\boldsymbol{C}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ with the orthonormal functions $\{f_n(\boldsymbol{x})\}$ and constants $\lambda_n$ in (2).

THEOREM 3.2. *[11] A second-order q.m. continuous process $p(\boldsymbol{x}, \theta)$ on a closed interval $\boldsymbol{D}$ has an orthogonal decomposition*

$$p(\boldsymbol{x}, \theta) = \sum_{n=1}^{\infty} \sqrt{\lambda_n} \xi_n(\theta) f_n(\boldsymbol{x}), \text{ with} \tag{4}$$

$$E[\xi_m(\theta) \xi_n(\theta)] = \delta_{mn}, \quad \int_{\boldsymbol{D}} f_m(\boldsymbol{x}) f_n(\boldsymbol{x}) d\boldsymbol{x} = \delta_{mn}, \tag{5}$$

*if, and only if, the $\lambda_n$ are the eigenvalues and $f_n(\boldsymbol{x})$ are the orthonormalized eigenfunctions of $\boldsymbol{C}(\boldsymbol{x}_1, \boldsymbol{x}_2)$. Then the series converges in q.m. uniformly on $\boldsymbol{D}$.*

The expansion in (4) is the Karhunen-Loève expansion.

### 3.1 Example

The experimental data in [3] shows that the covariance decreases as the distance between two points increases and can be accurately modeled using a linearly decreasing function. Another covariance function that models similar behavior is the exponentially decreasing covariance function of the form $\boldsymbol{C}(\boldsymbol{x}_1, \boldsymbol{x}_2) = e^{-c_x |x_1 - x_2|} e^{-c_y |y_1 - y_2|} = C(x_1, x_2) C(y_1, y_2)$, where $c_x$ and $c_y$ are the inverse of the *correlation length* in the $x$ and the $y$ directions respectively. One example of this covariance function is shown in Fig. 2(a). Since this covariance function is continuous at all $(\boldsymbol{x}, \boldsymbol{x})$, the stochastic process that corresponds to this covariance function will be q.m. continuous (Theorem 3.1) and thus the KLE (2) is valid. It can be shown [14] that the product of the eigenfunctions of one-dimensional exponential covariance kernel are also eigenfunctions of the two-dimensional exponential covariance kernel. Thus, we need to only obtain the solutions of the equation

$$\int_{D_X} e^{-c|x_1 - x_2|} g_n(x_1) dx_1 = \lambda_n g_n(x_2). \tag{6}$$

The general solution to this integral equation has the form [14]

$$g_n(x) = a_1 \cos(\omega_n x) + a_2 \sin(\omega_n x), \quad \lambda_n = \frac{2c}{\omega_n^2 + c^2}, \quad (7)$$

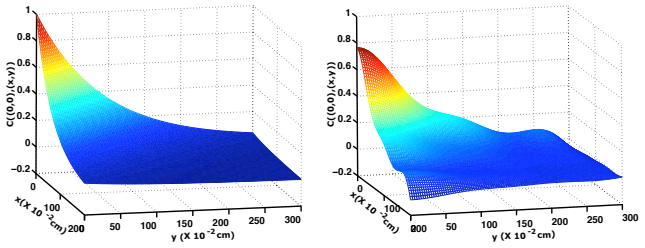
where $\omega_n$ is the solution of $c - \omega \tan(\omega a) = 0$ and $\omega + c\tan(\omega a) = 0$ for odd and even $n$ respectively. Also, $a_2 = 0$ for odd $n$ and $a_1 = 0$ for even $n$. Similarly, an analytical solution can be obtained for [14]

$$C(x_1, x_2) = \big(1 - d_x|x_1 - x_2|\big), \quad |x_1 - x_2| \in \left[0, \frac{1}{d_x}\right] \quad (8)$$

$$C(x_1, x_2) = \min\{x_1, x_2\}, \quad (x_1, x_2) \in [0, T]^2 \quad (9)$$

$$C(x_1, x_2) = \frac{\sin\big(b_x(x_1 - x_2)\big)}{\pi(x_1 - x_2)} \quad (10)$$

where $\boldsymbol{C}(\boldsymbol{x}_1, \boldsymbol{x}_2) = C(x_1, x_2)C(y_1, y_2)$ and $d_x, T$ are some constants.




(a) (b)

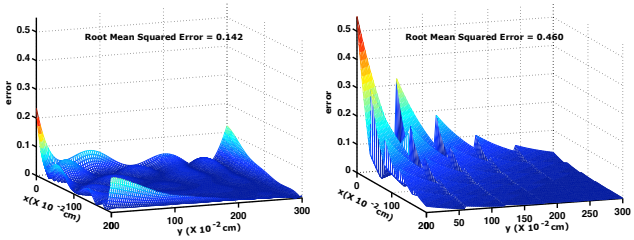
**Figure 2: (a) Exponential Covariance Function with $c_x = 0.016$ and $c_y = 0.011$, and (b) Approximation using KLE and $M = 25$.**

For practical use (4) has to be truncated. The covariance function of the truncated expansion can be obtained using

$$\widehat{\boldsymbol{C}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = E[\hat{p}(\boldsymbol{x}_1, \theta), \hat{p}(\boldsymbol{x}_2, \theta)] \quad (11)$$

$$= \sum_{n=1}^{M} \lambda_n f_n(\boldsymbol{x}_1) f_n(\boldsymbol{x}_2). \quad (12)$$

The eigenvalues $\lambda_n$ determine the contribution of the $n$-th random variable to the variance of $p(\boldsymbol{x}, \theta)$. Since we can always order the eigenvalues such that $\lambda_n > \lambda_{n+1}$, we truncate the expansion by finding the smallest $M$ such that $\lambda_M(\sum_{n=1}^{M} \lambda_n)^{-1} \leq \epsilon$, where $\epsilon$ is a threshold decided by the designer. In this work, we choose $\epsilon = 0.005$. Using this criteria, the KLE for $p(\boldsymbol{x}, \theta)$ having a covariance function as shown in Fig. 2(a) was obtained and truncated to obtain $M = 25$ terms. The covariance function $\widehat{\boldsymbol{C}}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ of this truncated expansion was then obtained using (11) and is shown in Fig. 2(b). The accuracy of this approximation can be seen from the absolute error between $\widehat{\boldsymbol{C}}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ and $\boldsymbol{C}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ as shown in Fig. 3(a). In comparison, if the PCA is used to represent this covariance function with the same number of partitions as the number of terms in the KLE, it introduces significant error in the approximation. For example, Fig. 3(b) shows the absolute error between the $\boldsymbol{C}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ and the covariance matrix $\boldsymbol{C}_d$ with $r = 25$ partitions. As can be seen from the figure and the respective Root Mean Squared Errors (RMSE), the KLE provides a significantly more accurate representation of the covariance function. It should be noted that this accurate representation does not require high computational complexity since we can obtain the analytical solutions for the integral equation (3). In comparison, if $M$ partitions are used in the PCA then the computational complexity is $O(M^3)$.




(a) (b)

**Figure 3: Absolute Error between the exponential covariance function and (a) the covariance function of the KLE, and (b) Covariance matrix for input to PCA.**

## 4. LEAKAGE MINIMIZATION

Let $G = (N, E)$ be a Directed Acyclic Graph (DAG) corresponding to a digital circuit. The set $N = \{1, \ldots, n\}$ corresponds to the set of nodes in the DAG, where each node represents a gate in the original circuit. The set $E$ contains the edges in the DAG $G$. Also if an edge $e_{ij} \in E$, then the gate corresponding to node $i$ fanouts to the gate corresponding to node $j$ in the circuit. Henceforth, gate $i$ would imply gate corresponding to node $i$ in the DAG. Let $W_i \in \boldsymbol{W} = \{W_1, \ldots, W_k\}$, $L_{e,i} \in \boldsymbol{L_e} = \{L_1, \ldots, L_m\}$ and $V_{t,i} \in \boldsymbol{V_t} = \{V_{t,low}, V_{t,high}\}$ be an assignment of gate size, gate length and threshold voltage, respectively for gate $i$. For each node $i \in N$, we associate a decision variable $u_i = (W_i, L_i, V_{t,i}) \in \boldsymbol{U} = \boldsymbol{W} \times \boldsymbol{L_e} \times \boldsymbol{V_t}$ that defines an implementation for the gate $i$.

Now, due to process variations, the *manufactured* values of the gate length and threshold voltage for gate $i$ are different than $L_{e,i}$ and $V_{t,i}$, respectively. Hence, we model the manufactured values of the parameters $p_i$ as a random variable

$$p_i = \mu_{p,i} + \sigma_p(\boldsymbol{x}_i) \cdot p(\boldsymbol{x}_i, \theta), \quad (13)$$

where $p(\boldsymbol{x}, \theta)$ is the stochastic process corresponding to the parameter $p$, $\boldsymbol{x}_i$ denotes the location of the gate $i$ on the die, $\mu_{p,i}$ is the assigned value of the parameter and $\sigma_p(\boldsymbol{x}_i)$ is the standard deviation of the parameter of gate $i$. In this work, we assume that $p(\boldsymbol{x}, \theta)$ is a Gaussian process. Using the KLE , $p_i$ can be written as

$$p_i = \mu_{p,i} + \sigma_p(\boldsymbol{x}_i) \cdot \sum_{n=1}^{M} \sqrt{\lambda_{p,n}} \xi_{p,n}(\theta) f_{p,n}(\boldsymbol{x}_i). \quad (14)$$

where $f_{p,n}(x)$ and $\lambda_{p,n}$ are the eigenfunctions and the eigenvalues of the covariance matrix $\boldsymbol{C}_p(\boldsymbol{x}_1, \boldsymbol{x}_2)$ of $p(\boldsymbol{x}, \theta)$ and $\{\xi_n(\theta)\}$ are standard normal random variables. Let $I_i : \boldsymbol{U} \rightarrow \boldsymbol{R}$ be the function defining the leakage of the gate corresponding to node $i$, and $I_T : \boldsymbol{U}^n \rightarrow \boldsymbol{R}$ be the total leakage of the circuit. The total leakage of the circuit is the sum of the leakage of the individual gates, that is $I_T = \sum_{i \in N} I_i$. Similarly, let $D_P : \boldsymbol{U}^n \rightarrow \boldsymbol{R}$ be the delay function of a path $P \in \mathcal{P}$ of the circuit, where $\mathcal{P}$ is the set of all paths in the circuit. Since, leakage and delay are functions of random variables, they are also random variables. In this work, we maximize the leakage yield at a target leakage $I_{req}$ with constraints on the timing yield of every path. An alternative to this approach is to constraint the *circuit timing yield*

and use a Statistical Static Timing Analysis [4] engine in the optimization. Hence, the discrete delay constrained leakage minimization problem can be formulated as

$$\begin{aligned} \max_{\boldsymbol{u}\in \boldsymbol{U}^n} \quad & Pr(I(\boldsymbol{u}) \le I_{req}) \\ \text{subject to} \quad & Pr(D_P(\boldsymbol{u}) \le T_{req}) \ge \alpha, \ \ \forall P \in \mathcal{P}. \end{aligned} \quad (15)$$

Also, $\boldsymbol{u} = (u_1, \ldots, u_n) \in \boldsymbol{U}^n$ corresponds to an implementation of a circuit with gate $i$ assigned the configuration $u_i$. Since the path delay is the sum of the delays of the gates on the path, it can be well approximated by a normal random variable even if the individual gate delays are not normally distributed (Central Limit Theorem [11]). Hence, the delay constraints can be rewritten as

$$z_\alpha(D(\boldsymbol{u})) = E[D_P(\boldsymbol{u})] + z_\alpha \sigma(D_P(\boldsymbol{u})) \le T_{req} \ \ \forall P \in \mathcal{P}, \quad (16)$$

where $z_\alpha$ is the $\alpha$-percentile of $N(0,1)$. The optimization approach is based on the method used in [15].

## 4.1 Leakage Model

We use the following model for sub-threshold leakage [16].

$$\begin{aligned} I_T &= \sum_{i\in N} I_o \frac{W_i}{L_{e,o}^k} e^{\left(\frac{-(V_{t,i}+a_1 L_{e,i})}{S}\right)}, \quad k > 1 \\ &= \sum_{i\in N} I'_{o,i} \exp\Big(\sum_{n=1}^{M} \alpha_{V_t,ni}\xi_{V,n} + \alpha_{L_e,ni}\xi_{L,n}\Big) \end{aligned} \quad (17)$$

where $I_o$ is the nominal sub-threshold leakage, $k$ and $S$ are positive fitting parameters, $I'_{o,i} = I_o \frac{W_i}{L_{e,o}^k} \exp(-(\mu_{V_t,i} + a_1\mu_{L_e,i})/S)$, $\alpha_{p,ni} = S^{-1}\sigma_p(\boldsymbol{x}_i)\sqrt{\lambda_{p,ni}} f_{p,n}(\boldsymbol{x}_i)$ for $p = V_t$ and $p = L_e$. The summation in (17) is over all the gates in the circuit. The above model captures the dependence of the sub-threshold leakage on the *all* the decision variables. This model was fitted to the data from SPICE to obtain the parameters. Hence, the leakage given in (17) can be written in the following general form:

$$I_T = \sum_{i\in N} I'_{o,i} \exp\Big(-\sum_{k=1}^{2M} \beta_{ik}\zeta_k\Big) \quad (18)$$

where, the random variables $\zeta_k$ are independent standard normal random variables and the coefficients $\beta_{ik}$'s are dependent on the current assignment of $W, L_e$ and $V_t$ to gate $i$. Since the expectation operator $E[\cdot]$ is linear and the random variables $\zeta_k$'s are independent, the expected leakage, $E[I_T]$ and the second moment of the leakage, $E[I_T^2]$ can be computed as:

$$E[I_T] = \sum_{i\in N} I'_{o,i} \prod_{k=1}^{2M} E[\exp(-\beta_{ik}\zeta_k)] \quad (19)$$

$$E[I_T^2] = \sum_{i,j\in N} I'_{o,i} I'_{o,j} \prod_{k=1}^{2M} E[\exp(-(\beta_{ik}+\beta_{jk})\zeta_k)]. \quad (20)$$

The expectation of $e^{-aX}$ where $X$ is a standard normal random variable is given by: $E[e^{-aX}] = e^{a^2/2}$. Using (19) and (20), the variance of the leakage can be computed as: $\sigma^2(I_T) = E[I_T^2] - E^2[I_T]$. The cost of computing the second moment (and variance) of the leakage is $O(n^2M)$ because it involves the product of terms. The coefficients $\beta_{ij}$'s, $i \in N$ and $j \in \{1, \ldots, 2M\}$ are dependent on the values of $L_{e,i}$ and $V_{t,i}$ due to the dependence of the variance of $L_{e,i}$ and $V_{t,i}$ on the assigned value. The leakage is assumed to be a lognormal random variable and its yield is computed using

$$Y_I = Pr(I_T(\boldsymbol{u}) \le I_{req}) = Pr(\log(I_T(\boldsymbol{u})) \le log(I_{req})). \quad (21)$$

Since $\log(I_T(\boldsymbol{u}))$ is a normal random variable, the expression in (21) can be computed in constant time using the moments of $I_T$ [6].

## 4.2 Delay Model

The delay is modeled using an improved alpha-power law based physical delay model proposed in [17]. Assuming that the transistors operate in the saturation mode, the proposed model can be simplified into the form shown in (22).

$$E[d_i] = \alpha\left(\frac{\beta_1}{w_i} + \beta_2\right)\frac{L_{e,i}V_{dd}}{(V_{dd}-V_{t,i})^2}\left(1 + \frac{(V_{dd}-V_{t,i})}{\gamma L_{e,i}}\right) \quad (22)$$

$$\sigma^2(d_i) = \frac{\partial E[d_i]}{\partial L_{e,i}}\sigma_L^2 + \frac{\partial E[d_i]}{\partial V_{t,i}}\sigma_{V_t}^2 \quad (23)$$

where $d_i$ represents the delay of the gate corresponding to node $i$. The model accounts for the short channel effects and the DIBL effect, which improve the model scalability to process and design variables. The variations in the threshold voltage are modeled using the Pelgrom's model [18] as $\sigma_{V_{t,i}}^2 = \kappa \cdot W_i^{-1} L_{e,i}^{-1}$, whereas, the variation in $L_e$ are assumed to be independent of the assigned value of the gate length. The parameters for these models are obtained by performing SPICE simulation and by fitting these models to SPICE data. The average error compared to the data from SPICE simulations is around 3-4% over ±25% range of $L_{e,i}$ for different values of the supply voltage and threshold voltage. For a given path $P \in \mathcal{P}$, the path delay is the sum of the delays of the gates on that path. A simple upper bound on the $\alpha$-percentile of the path delay is given by [19].

$$z_\alpha[D_P] \le z_\alpha^U[D_P] = \sum_{i\in P} E[d_i] + z_\alpha \sum_{i\in P} \sigma(d_i). \quad (24)$$

The authors of [19] show that the an assignment of gate sizes that satisfies $z_\alpha[D_P]$ for a path $P$, also satisfies $z_\alpha^U[D_p]$. That result can be easily extended to our case where we have more than one type of decision variables $(W, L, V_t)$. Hence, given a circuit configuration (or $\boldsymbol{x} \in \boldsymbol{X}^n$) that satisfies the delay constraint $z_\alpha^U[D_P]$ for all $P \in \mathcal{P}$, will also satisfy the constraint $z_\alpha[D_p]$ on $P$.

## 4.3 Optimization Algorithm

The optimization algorithm used in this work is based on the work in [20]. The efficiency of this algorithm is dependent on efficient (discrete) *gradient* computation of the objective and the constraint functions with respect to the decision variables. Since, the objective as well as the constraint functions of our problem are significantly different compared to [20], we propose *efficient* techniques that can give accurate estimates of the gradients of the functions involved with respect to the decision variables.

In order to maximize the leakage yield, first the unconstrained minimum of the $\alpha$-percentile of the delay is computed using the algorithm in [20], then the solution is moved in the direction such that the delay is increased and the leakage of the circuit is decreased (thus increasing the leakage yield) till the delay constraint is no longer met. The final result is the constrained minimum of the leakage function.

## 4.4 Delay Gradient

In [20], the author shows that the change in delay of a circuit on perturbing the configuration of a gate $i$ can be estimated accurately by considering only a subcircuit around

node $i$ consisting of two levels of transitive fanouts and two levels of transitive fanins of $i$. Using the upper bound of the delay as shown in (24) has an important implication that delay of the circuit (or a subcircuit) can be computed simply by assigning to each gate a delay of $z_\alpha^U(d_i)$. This assignment can be used to identify the path having the largest value of $z_\alpha^U[D_P]$, which can be used as an estimate for the $\alpha$-percentile of the circuit delay. On perturbing the configuration of gate $i$, its delay changes to $d_i^*$ and the path having the largest value of $z_\alpha^U[D_{P'}^*]$ can be identified. $D_{P'}^*$ corresponds to the delay of the path $P'$ that has the largest value of the upper bound on $\alpha$-percentile, when the configuration of a node is changed. Thus we can approximate the delay gradient $\Delta z_\alpha^U[D]$ as:

$$\Delta z_\alpha^U[D] \approx z_\alpha^U[D_P] - z_\alpha^U[D_{P'}^*] \quad (25)$$

This approach provides an efficient method for taking the variations into account as well as guarantees that we do not miss any feasible solutions.

## 4.5 Leakage Gradient

In the deterministic leakage minimization problem, the change in the leakage gradient can be computed in $O(1)$ time because of the *additiveness* of the leakage function. However, since we want to maximize the leakage yield, the objective function is no longer additive in the decision variables. In order to compute the leakage yield gradient, we need to compute the gradient of the first and the second moment of the leakage. Now, $\Delta E^2[I_T] = E^2[I_T] - E^2[I_T^*]$, where $I_T^*$ is the leakage of the circuit after perturbing the node $i$, can be rewritten as in (26).

$$\begin{aligned} \Delta E^2[I_T] &= \big(E[I_T] + E[I_T^*]\big)\big(E[I_T] - E[I_T^*]\big) && (26)\\ &= E[I_T + I_T^*] \cdot E[I_T - I_T^*] && (27)\\ &= 2E[I_T] \cdot E[\Delta I_i] - E^2[\Delta I_i] && (28) \end{aligned}$$

where $\Delta I_i = I_i - I_i^*$ and $I_i^*$ is the leakage of the node $i$ after changing its configuration. Hence, $\Delta E^2[I_T]$ can also be computed in $O(M)$ time. Similarly, we can write $\Delta E[I_T^2] = E[I_T^2] - E[I_T^{*2}]$ as in (29).

$$\begin{aligned} \Delta E[I_T^2] &= E[(I_T + I_T^*)(I_T - I_T^*)] && (29)\\ &= 2E[I_T \cdot \Delta I_i] - E[\Delta I_i^2] && (30) \end{aligned}$$

Now, $E[\Delta I_i^2]$ can be computed in $O(M)$ time. The first term on the RHS of (30) is the equal to $2E[I_T I_i] - 2E[I_T I_i^*]$. (31) shows how to compute $E[I_T I_i]$, the other term can be computed similarly.

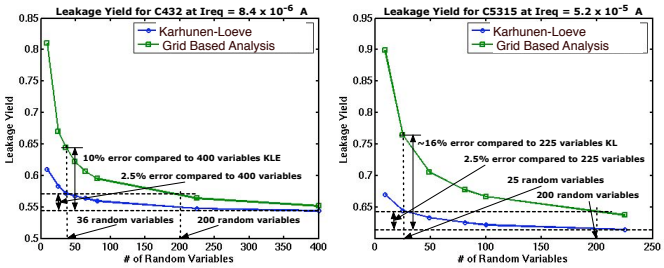
$$\begin{aligned} E[I_T I_i] &= E\Big[\sum_{j\in N} I'_{o,j} I'_{o,i} \exp\big(-\sum_{k=1}^{2M}(\beta_{jk} + \beta_{ik})\zeta_k\big)\Big] \\ &= \sum_{j\in N} I'_{o,j} I'_{o,i} \exp\Big(\sum_{k=1}^{2M} \frac{\beta_{jk}^2 + \beta_{ik}^2 + 2\beta_{jk}\beta_{ik}}{2}\Big) \quad (31) \end{aligned}$$

Thus, the gradient of the second moment can be computed only in $O(nM)$ time.

# 5. EXPERIMENTAL RESULTS

The proposed leakage reduction methodology was implemented as a tool in C++. The results were obtained on ISCAS85 benchmark circuits on dual-core 3.2GHz machines with 4GB of RAM. The circuits are first processed by mapping them using Berkeley SIS and placing them using UMpack [21]. The gates in the library were characterized using SPICE for $90nm$ technology to obtain the parameters for the leakage model as well as the mean and the variance of the delay. For a given circuit, the program takes in the following inputs: the gate level netlist (in BLIF format), placement data, a gate library consisting of different implementations of each gate type, the type of analysis (KLE or grid-based) to be performed, the number of random variables to be considered $M$, the required time $T_{req}$ and the required leakage $L_{req}$. The output is a set of optimal assignments of the decision variables, $W$, $L_e$ and $V_{th}$ for each gate. We now demonstrate the increase in accuracy as a result of using KLE compared to a grid-based approach.

Figures 4 (a) and 4 (b) show how the accuracy of yield analysis changes with the change in the number of random variables used for representing intra-die variations for C432 and C5315 respectively. The figures show that the leakage yield $P(I \le I_{req})$ obtained by using PCA (upper curve) always overestimates the leakage yield obtained using KLE (lower curve). This difference can be significantly large when a small number of random variables are considered. For example, for C5315, the difference between the grid-based based analysis and KLE based analysis for $M = 25$ random variables is 16% and 2.5%, respectively, with respect to a $M = 200$ random variables analysis using KLE. The reason for the large error is that the grid-based approach tries to reduce the number of random variables by discretizing the covariance function beforehand rather than decorrelating at a fine level first and then finding out the principal components as in the case of KLE.




(a) (b)

**Figure 4: Leakage Yield as a function of the number of random variables for (a) C432 and (b) C5315**

## 5.1 Leakage Optimization Results

Since the complexity of the leakage gradient computation is linearly dependent on the number of random variables, a lower value of $M$ is desirable. From the leakage analysis of the benchmark circuits, we found that KLE based analysis considering $M = 25$ random variables provides an accuracy of within $1 - 3\%$ compared to an analysis considering $M = 200$ random variables. Also, $M = 25$ was found to be greater than the number of random variables obtained using the truncation criteria outlined in Section 3.1. Hence, for leakage optimization purposes, we use $M = 25$ terms as the basis for performing KLE based optimization. Figure 5 shows the comparison of the KLE based optimization (KL25) with $M = 25$ terms in the expansion with three grid-based (GB) based optimizations, GB25, GB49 and GB81 where the circuits are partitioned into $25, 49$ and $81$ grids respectively. The *closeness* of two circuit implementations is determined by finding the fraction of gates that have different configuration (ie. different values of $W$, $L_e$ or $V_{th}$) in the two implementations. As the figure shows, the optimal circuit configuration obtained using KL25 matches closely with the optimal configuration obtained using GB81 for the benchmarks shown (average difference $\sim 12\%$), whereas the

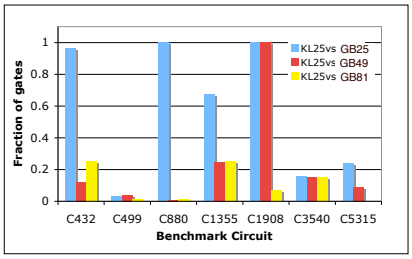



**Figure 5: Fraction of gates having a different configuration for a grid-based vs. KLE based optimization.**

configuration obtained using GB25 was significantly different than that obtained using GB25 (average difference $\sim$ 58%). The reason for this large difference is that as shown in the previous section, for the same circuit configuration, the leakage yield using GB25 is significantly overestimated as compared to KL25. Hence, in GB25, the estimate of the leakage yield for the initial circuit configuration is high. Hence, the optimization terminates in a small number of steps without trying to reduce the leakage significantly. In comparison, for KL25 the estimate of the leakage yield for the initial configuration is much lower. Hence, the optimization procedure keeps minimizing the leakage until it achieves a high leakage yield. Since the leakage yield estimated from GB81 is close to that estimated using KL25, the optimization in these two cases proceeds in similar direction and hence results in similar optimal circuit configurations. Thus, by using KLE, we can obtain the same optimal configuration by using a smaller number of random variables as compared to GB. Also, for C5315, the GB81 optimization did not terminate in 24hrs and hence the result has not been reported for this case.

Table 1 outlines the results of the leakage optimization for the benchmark circuits for two different values of the required time, $T_{req}$ to show the leakage delay trade-offs for each circuit. The mean leakage, standard deviation of the leakage and the leakage yield are also shown. As can be seen from the table, significant improvements in the leakage yield can be obtained for a small increase in the delay. For example, for circuit C1908, a change of 1.3% in $T_{req}$ results in an increase of 86% in the leakage yield. Also, the leakage

| Circuit | $T_{req}$ (ns) | $I_{req}$ ($\mu A$) | $E[I]$ | $\sigma(I)$ | Leak. Yield | Runtime (min.) |
|---|---|---|---|---|---|---|
| C1355 | 0.834 | 20.0 | 20.0 | 2.20 | 0.500 | 3.2hr |
| | 0.837 | 20.0 | 7.63 | 0.78 | 0.999 | 3.7hr |
| C1908 | 0.869 | 13.5 | 15.3 | 1.66 | 0.135 | 25m |
| | 0.881 | 13.5 | 4.8 | 0.48 | 0.999 | 30m |
| C3540 | 1.211 | 41.0 | 41.1 | 4.59 | 0.504 | 3.8hr |
| | 1.240 | 41.0 | 33.5 | 3.70 | 0.969 | 4.05hr |
| C5315 | 1.010 | 65.0 | 59.6 | 6.63 | 0.799 | 12.4hr |
| | 1.061 | 65.0 | 54.3 | 6.00 | 0.953 | 13.3hr |
| C6228 | 2.979 | 82.0 | 80.7 | 9.22 | 0.576 | 7.6hr |
| | 3.053 | 82.0 | 76. | 8.67 | 0.755 | 8.2hr |
| C7552 | 1.248 | 75.0 | 81.6 | 9.20 | 0.239 | 12.6hr |
| | 1.457 | 75.0 | 72.1 | 8.01 | 0.660 | 13.2hr |

**Table 1: Leakage-Delay trade-off for ISCAS'89 circuits.**

minimization algorithm starts from a minimum delay solution and keeps increasing the leakage until the delay constraint is violated. Hence, for a given $T_{req}$, we can obtain the optimal configurations for required times less than $T_{req}$ as well and so the designer can choose an optimal point on the leakage delay trade-off curve.

# 6. CONCLUSIONS

We proposed the use of Karhunen-Loève Expansion (KLE) for accurate and efficient representation of correlations in the semiconductor manufacturing process as compared to existing grid-based approaches. We showed that the use of KLE provides a much better approximation of the covariance function of the spatial stochastic process characterizing the device parameters. We showed that for obtaining similar accuracy, the grid-based approach would require up to $4-5\times$ the number of random variables used in KLE. Thus we provide an efficient technique for reducing the number of random variables while maintaining the accuracy.

# 7. REFERENCES

[1] International technology roadmap for semiconductors. 2003.
[2] D. Boning and S. Nassif. *Models of process variations in device and interconnect, Design of High- Performance Microprocessor Circuits*, chapter 6. IEEE Press, 2000.
[3] Paul Friedberg et al. Modeling within-die spatial correlation effects for process-design co-optimization. In *Proc. of ISQED*, 2005.
[4] Hongliang Chang and Sachin Sapatnekar. Statistical timing analysis considering spatial correlations using a single pert-like traversal. In *IEEE ICCAD*, 2003.
[5] Debjit Sinha and Hai Zhou. A unified framework for statistical timing analysis with coupling and multiple input switching. In *IEEE/ACM ICCAD*, 2005.
[6] Rajeev Rao, Anirudh Devgan, David Blaauw, and Dennis Sylvester. Parametric yield estimation considering leakage variability. In *Proc. of DAC*, 2004.
[7] Jaskirat Singh, Vidyasagar Nookala, Zhi-Quan Luo, and Sachin Sapatnekar. Robust gate sizing using geometric programming. In *Proc. of DAC*, 2005.
[8] Murari Mani, Anirudh Devgan, and Michael Orshansky. An efficient algorithm for statistical minimization of total power under timing yield constraints. In *Proc. of DAC*, 2005.
[9] Kaviraj Chopra et al. Parametric yield maximization using gate sizing based on efficient statistical power and delay gradient computation. In *IEEE/ACM ICCAD*, 2005.
[10] Matthew R. Guthaus, Natesan Venkateswaran, Chandu Visweswariah, and Vladimir Zolotov. Gate sizing using incremental parameterized statistical timing analysis. In *IEEE/ACM ICCAD*, 2005.
[11] M. Loève. *Probability Theory*. D. Van Nostrand Company Inc., 1960.
[12] Aseem Agarwal, David Blaauw, and Vladimir Zolotov. Statistical timing analysis for intra-die process variations with spatial correlations. In *IEEE ICCAD*, 2003.
[13] T W Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley and Sons, 1958.
[14] R G Ghanem and P Spanos. *Stochastic Finite Elements: A Spectral Approach*. Springer-Verlag, 1991.
[15] S. Bhardwaj, S. Vrudhula, and Yu Cao. LOTUS: Leakage Optimization under Timing Uncertainty for Standard-cell designs. In *Proc. of ISQED*, 2006.
[16] UC Berkeley Device Group. *BSIM 4.2.1 MOSFET Model - User's Manual*, 2004.
[17] Y. Cao and L T Clark. Mapping statistical process variations toward circuit performance variability: An analytical modeling approach. In *Proc. of DAC*, 2005.
[18] M J M Pelgrom, A C J Duinmaijer, and A P G Welbers. Matching properties of mos transistors. *IEEE Journal of Solid-State Circuits*, 24(5):1433–1439, Oct 1989.
[19] S. Bhardwaj and S. Vrudhula. Leakage minimizatio of nano-scale circuits in the presence of systematic and random variations. In *Proc. of DAC*, pages 541–546, 2005.
[20] Olivier Coudert. Gate sizing for constrained delay/power/area optimization. *IEEE Transactions on VLSI Systems*, 5(4):465–472, December 1997.
[21] http://vlsicad.eecs.umich.edu/bk/pdtools/.