Virtual Probe: A Statistical Framework for Low-Cost Silicon Characterization of Nanoscale Integrated Circuits

Wangyang Zhang, *Student Member, IEEE,* Xin Li, *Senior Member, IEEE,* Frank Liu, *Senior Member, IEEE,* Emrah Acar, *Senior Member, IEEE,* Rob A. Rutenbar, *Fellow, IEEE,* and Ronald D. Blanton, *Fellow, IEEE*

***Abstract*—In this paper, we propose a new technique, referred to as virtual probe (VP), to efficiently measure, characterize, and monitor spatially-correlated inter-die and/or intra-die variations in nanoscale manufacturing process. VP exploits recent breakthroughs in compressed sensing to accurately predict spatial variations from an exceptionally small set of measurement data, thereby reducing the cost of silicon characterization. By exploring the underlying sparse pattern in spatial frequency domain, VP achieves substantially lower sampling frequency than the well-known Nyquist rate. In addition, VP is formulated as a linear programming problem and, therefore, can be solved both robustly and efficiently. Our industrial measurement data demonstrate the superior accuracy of VP over several traditional methods, including 2-D interpolation, Kriging prediction, and k-LSE estimation.**

***Index Terms*—Characterization, compressed sensing, integrated circuit, process variation.**

I. Introduction

As INTEGRATED circuits (ICs) scale to finer feature size, it becomes increasingly difficult to control process variations for nanoscale technologies [2], [3]. The increasing fluctuations in manufacturing process introduce unavoidable and significant uncertainties in circuit performance. Hence, modeling and analyzing these variations to ensure manufacturability and improve parametric yield has been identified as a top priority for today's IC design.

Manuscript received March 1, 2011; revised June 5, 2011; accepted July 9, 2011. Date of current version November 18, 2011. This work was supported in part by the C2S2 Focus Center, one of six research centers funded under the Focus Center Research Program, a Semiconductor Research Corporation Entity. This work was also supported in part by the National Science Foundation, under Contract CCF-0915912. This paper was presented in part at the International Conference on Computer-Aided Design in 2009 [1]. This paper was recommended by Associate Editor D. Sylvester.

W. Zhang, X. Li, and R. D. Blanton are with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: wangyan1@ece.cmu.edu; xinli@ece.cmu.edu; blanton@ece.cmu.edu).

F. Liu is with the IBM Research Laboratory, Austin, TX 78758 USA (e-mail: frankliu@us.ibm.com).

E. Acar is with the IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA (e-mail: emrah@us.ibm.com).

R. A. Rutenbar is with the Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: rutenbar@illinois.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/TCAD.2011.2164536

Toward this goal, various techniques have been proposed for statistical IC analysis and optimization, e.g., design centering, statistical timing analysis [4]–[7] and post-silicon tuning [8]–[10]. These techniques aim to predict and, consequently, minimize circuit-level performance variations in order to create a robust design with high parametric yield. The efficiency of these methods relies heavily on the accuracy of the variation model (e.g., distribution and correlation) that provides the important information about manufacturing uncertainties.

Accurately extracting the variation model, however, is not trivial. Silicon wafers/chips must be carefully tested and characterized using multiple test structures (e.g., I–V structures and ring oscillators) deployed in wafer scribe lines and/or within product chips [11]–[15]. The traditional silicon characterization suffers from three major issues.

1) *Large area overhead*: today's advanced microprocessor chips typically contain hundreds of on-chip ring oscillators to characterize and monitor parametric variations, resulting in significant overhead in silicon area [12].
2) *Long testing time*: physically measuring all test structures through a limited number of I/O ports consumes a large amount of testing time [13]. In nanoscale technologies, IC testing has contributed to a significant portion of the total manufacturing cost [27].
3) *Low testing reliability*: IC testing may even damage the wafer/chip being tested. For instance, wafer probe test may permanently damage the wafer due to mechanical stress [13].

The combination of these critical issues results in continuously growing silicon characterization cost, as more and more test structures must be added to capture the complicated spatial variations of small devices. Even though silicon characterization has been extensively studied in the past, there is an immediate need to revisit this area and develop a more efficient methodology to reduce cost.

To this end, we ask the following fundamental question: How many test structures are minimally required to fully capture the spatial variation information? A quick answer to this question can be made based on the well-known Nyquist–Shannon sampling theorem [25]. Namely, if the variations contain no spatial frequency higher than f_{MAX}, the sampling frequency must be at least $2 \cdot f_{\mathrm{MAX}}$, i.e., test structures must be spaced at most $1/(2 \cdot f_{\mathrm{MAX}})$ apart.

The Nyquist sampling theorem generally assumes that all frequency components below the maximum frequency f_{MAX} may exist. However, this is not true for most silicon characterization applications. As will be demonstrated by the industrial measurement data in Section V, spatial variations typically have a sparse representation in frequency domain (i.e., a large number of Fourier coefficients are almost zero). In this case, simply sampling at Nyquist rate generates a large number of redundant data. Such redundancy has been observed in many other application domains. For example, the key idea of image compression is to remove the redundancy and represent the information in a compact form [30]. However, our silicon characterization problem is substantially different from image compression, as we do not want to fully sample spatial variations at Nyquist rate and then "compress" them. Instead, we want to avoid redundant sampling in the first place to reduce characterization cost. The challenging issue here is how to efficiently sample few test structures on a wafer/chip and then accurately recover the essential spatial variation information.

In this paper, we exploit the recent advances in statistics (known as compressed sensing [20]–[23]) to develop a novel framework of virtual probe (VP) for low-cost silicon testing and characterization. Our goal is to accurately predict the spatial variations of a wafer/chip by measuring very few test structures at a set of selected locations. The proposed VP algorithm is derived from maximum *a posteriori* (MAP) estimation [29]. It is mathematically formulated as a linear programming problem that can be solved both robustly and efficiently. Most importantly, several theoretical studies from the statistics community have proved that by exploring the sparse pattern in spatial frequency domain, VP can fully reconstruct the spatial variations with probability nearly equal to 1, even if the spatial sampling frequency is much lower than the Nyquist rate [20]–[23]. As will be demonstrated by the industrial examples in Section V, VP shows superior accuracy over several traditional methods including 2-D interpolation [31], Kriging prediction [16], and k-LSE estimation [17].

The remainder of this paper is organized as follows. In Section II, we develop the mathematical formulation of VP, and then discuss the implementation details in Section III. Next, several possible applications of VP are briefly discussed in Section IV. The efficacy of VP is demonstrated by a number of examples with industrial measurement data in Section V. Finally, we conclude in Section VI.

II. Virtual Probe

The key idea of VP is to deploy and measure very few test structures at a set of selected locations of a wafer/chip. The parametric variations at other locations are not directly measured by silicon testing. Instead, VPs are conceptually added at these locations to predict the variation information through the use of a statistical algorithm, as shown in Fig. 1. In other words, unlike the traditional approach that uses a large number of test structures, we propose to physically monitor the variability at very few locations and then apply a "smart" algorithm to accurately predict the complete spatial variation. In this section, we first derive the mathematical formulation of VP based on spatial frequency-domain analysis. Next, we derive a MAP algorithm to solve the VP problem by exploring the unique sparse pattern in frequency domain. Finally, the accuracy of the MAP estimation is justified by studying several important theorems recently developed in the statistics community [20]–[23].

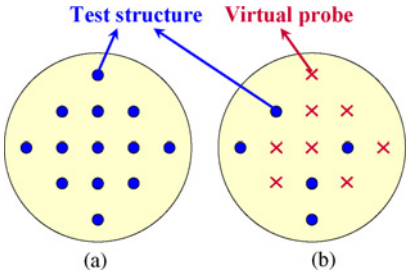


Fig. 1. Example of the proposed virtual probes. (a) Traditionally, a large number of test structures are deployed and measured to fully characterize process variations. (b) We propose to deploy and measure very few test structures, and virtual probes are conceptually added to fully recover spatial variations through the use of a statistical algorithm.

A. Mathematical Formulation

Mathematically, the spatial variations of a performance of interest (e.g., the frequency of a ring oscillator) can be expressed as a 2-D function $g(x, y)$, where x and y represent the coordinates of a spatial location on a wafer or chip. If $g(x, y)$ contains no spatial frequency higher than f_{MAX}, the Nyquist–Shannon sampling theorem [25] tells us to sample $g(x, y)$ with a frequency of $2 \cdot f_{MAX}$ in order to perfectly recover the continuous function $g(x, y)$.

Mathematically, the function $g(x, y)$ can be mapped to the frequency domain by a number of 2-D linear transforms such as Fourier transform [25], discrete cosine transform (DCT) [30], and wavelet transform [30]. In this paper, we use DCT to illustrate the basic idea of VP. It should be noted, however, that the proposed VP framework can also be implemented with other linear transformations.

We discretize the 2-D function $g(x, y)$ at a spatial frequency higher than the Nyquist rate. Without loss of generality, we denote the coordinates x and y as integers $x \in \{1, 2, \ldots, P\}$ and $y \in \{1, 2, \ldots, Q\}$ after discretization. The DCT transform can be represented as [30]

$$\begin{aligned} G(u, v) &= \sum_{x=1}^{P} \sum_{y=1}^{Q} \alpha_u \cdot \beta_v \cdot g(x, y) \\ &\quad \cdot \cos \frac{\pi \cdot (2x-1) \cdot (u-1)}{2 \cdot P} \\ &\quad \cdot \cos \frac{\pi \cdot (2y-1) \cdot (v-1)}{2 \cdot Q} \end{aligned} \tag{1}$$

where

$$\alpha_u = \begin{cases} \sqrt{1/P} & (u = 1) \\ \sqrt{2/P} & (2 \le u \le P) \end{cases} \tag{2}$$

$$\beta_v = \begin{cases} \sqrt{1/Q} & (v = 1) \\ \sqrt{2/Q} & (2 \le v \le Q). \end{cases} \tag{3}$$

In (1), $\{G(u, v); u = 1, 2, \ldots, P, v = 1, 2, \ldots, Q\}$ represents a set of DCT coefficients. Equivalently, the sampling values $\{g(x, y); x = 1, 2, \ldots, P, y = 1, 2, \ldots, Q\}$ can be expressed as the linear combination of $\{G(u, v); u = 1, 2, \ldots, P, v = 1, 2, \ldots, Q\}$ by the inverse discrete cosine transform (IDCT) [30]

$$g(x, y) = \sum_{u=1}^{P} \sum_{v=1}^{Q} \alpha_u \cdot \beta_v \cdot G(u, v) \cdot \cos \frac{\pi \cdot (2x - 1) \cdot (u - 1)}{2 \cdot P} \cdot \cos \frac{\pi \cdot (2y - 1) \cdot (v - 1)}{2 \cdot Q}. \tag{4}$$

From (1)–(4), it is easy to verify that once the sampling values $\{g(x, y); x = 1, 2, \ldots, P, y = 1, 2, \ldots, Q\}$ are known, the DCT coefficients $\{G(u, v); u = 1, 2, \ldots, P, v = 1, 2, \ldots, Q\}$ are uniquely determined, and vice versa.

The proposed VP framework, however, will go one step further. Our objective is to accurately recover $\{g(x, y); x = 1, 2, \ldots, P, y = 1, 2, \ldots, Q\}$ from a very small number of (say, M) samples at the locations $\{(x_m, y_m); m = 1, 2, \ldots, M\}$ where $M << PQ$. In other words, the recovery can be formulated as the following linear equation:

$$A \cdot \eta = B \tag{5}$$

where

$$A = \begin{bmatrix} A_{1,1,1} & A_{1,1,2} & \cdots & A_{1,P,Q} \\ A_{2,1,1} & A_{2,1,2} & \cdots & A_{2,P,Q} \\ \vdots & \vdots & \vdots & \vdots \\ A_{M,1,1} & A_{M,1,2} & \cdots & A_{2,P,Q} \end{bmatrix} \tag{6}$$

$$A_{m,u,v} = \alpha_u \cdot \beta_v \cdot \cos \frac{\pi (2x_m - 1)(u - 1)}{2 \cdot P} \cdot \cos \frac{\pi (2y_m - 1)(v - 1)}{2 \cdot Q} \tag{7}$$

$$\eta = \begin{bmatrix} G(1, 1) & G(1, 2) & \cdots & G(P, Q) \end{bmatrix}^T \tag{8}$$

$$B = \begin{bmatrix} g(x_1, y_1) & g(x_2, y_2) & \cdots & g(x_M, y_M) \end{bmatrix}^T. \tag{9}$$

In (5)–(9), the DCT coefficients $\{G(u, v); u = 1, 2, \ldots, P, v = 1, 2, \ldots, Q\}$ are the problem unknowns. In other words, we need to determine $\{G(u, v); u = 1, 2, \ldots, P, v = 1, 2, \ldots, Q\}$ based on the measurement data $\{g(x_m, y_m); m = 1, 2, \ldots, M\}$. Once the DCT coefficients $\{G(u, v); u = 1, 2, \ldots, P, v = 1, 2, \ldots, Q\}$ are known, the function $\{g(x, y); x = 1, 2, \ldots, P, y = 1, 2, \ldots, Q\}$ can be easily calculated by the IDCT in (4).

Solving the linear equation $A \cdot \eta = B$ in (5), however, is not trivial, since M (i.e., the number of equations) is vastly less than PQ (i.e., the number of unknowns). Namely, the linear equation in (5) is profoundly underdetermined. While (5) cannot be uniquely solved by a simple matrix inverse, its solution can be statistically determined by considering additional prior information via Bayesian inference, as will be discussed in the next subsection.

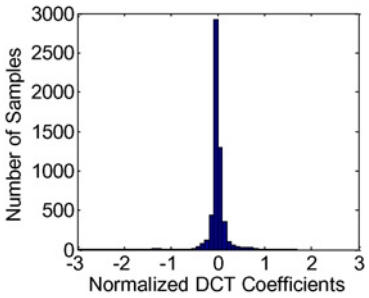


Fig. 2. Histogram of the normalized DCT coefficients calculated from 17 wafers for an industrial IC design example.

B. MAP Estimation

In this subsection, we describe an efficient algorithm using MAP estimation to statistically solve the linear equation in (5). Although the result of this subsection can be derived by applying a number of elegant statistics theorems [19]–[23], [29], we attempt to describe the MAP algorithm at a level that is intuitive to the CAD community. More mathematical details of MAP can be found in [19]–[23] and [29].

To solve (5), we first define a so-called prior distribution for η [29]. Intuitively, the prior distribution represents our prior knowledge about η without seeing any measurement data. Such prior information helps us to further constrain the underdetermined linear equation $A \cdot \eta = B$ in (5) so that a meaningful solution can be uniquely found. At first glance, this seems impossible, since we would expect that the spatial variations and, hence, the DCT coefficients in η are substantially different from wafer to wafer and from chip to chip. However, we will show in this paper that η has a unique property that we can exploit to define the prior distribution.

Before moving forward, let us first examine the following example of an industrial IC design. We measure the flush delay of this circuit from 17 wafers, each containing 282 chips. Flush delay is the time for a transition to propagate across a scan chain. We calculate the DCT coefficients and plot the histogram of them in Fig. 2. We notice that the distribution has a sharp peak at zero. This implies that most DCT coefficients are close to zero. In general, if the performance variations $\{g(x, y); x = 1, 2, \ldots, P, y = 1, 2, \ldots, Q\}$ present a spatial pattern, i.e., the variations are spatially correlated, the vector η that contains the corresponding DCT coefficients $\{G(u, v); u = 1, 2, \ldots, P, v = 1, 2, \ldots, Q\}$ is sparse. This unique property of sparseness has been observed in many image processing tasks [30], and has motivated the compressed sensing research for image recovery using a minimum number of samples [19]–[23]. The previous research in compressed sensing shows that if most of these DCT coefficients are expected to be zero, we can reconstruct the image from a surprisingly small (i.e., "compressed") set of samples. As will be demonstrated by several industrial examples in Section V, this assumption of sparseness is also valid for our silicon characterization application.

While we assume that a large number of DCT coefficients are close to zero, we do not know the exact locations of these zeros. Otherwise, solving η from the linear equation $A \cdot \eta = B$

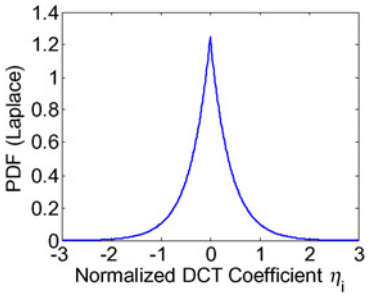


Fig. 3. Optimally-fitted Laplace distribution for the normalized DCT coefficients calculated from 17 wafers for an industrial IC design example.

in (5) becomes trivial. To find the unique solution η of the under determined linear equation $A \cdot \eta = B$, we need to statistically model the histogram in Fig. 2 by using a zero-mean Laplace distribution to approximate the probability density function (PDF) of each DCT coefficient $\{\eta_i; i = 1, 2, \ldots, PQ\}$ [29]

$$pdf\,(\eta_i) = \frac{1}{2\lambda} \cdot \exp\left(-\frac{|\eta_i|}{\lambda}\right) \quad (i = 1, 2, \cdots, PQ) \tag{10}$$

where $pdf(\eta_i)$ stands for the PDF of η_i, and $\lambda > 0$ is a parameter that controls the variance of the distribution. The parameter λ in (10) can be optimally found by maximum likelihood estimation [29]. Fig. 3 shows the optimally-fitted Laplace distribution for the data set in Fig. 2. In practice, however, it is not necessary to know the value of λ. As will be shown in (17), the solution η is independent of the actual value of λ.

To completely define the prior distribution, we further assume that all DCT coefficients in the vector $\eta \in R^{PQ}$ are mutually independent. Hence, the joint PDF of η is represented as

$$\begin{aligned} pdf\,(\eta) &= \left(\frac{1}{2\lambda}\right)^{PQ} \cdot \prod_{i=1}^{PQ} \exp\left(-\frac{|\eta_i|}{\lambda}\right) \\ &= \left(\frac{1}{2\lambda}\right)^{PQ} \cdot \exp\left(-\frac{\|\eta\|_1}{\lambda}\right) \end{aligned} \tag{11}$$

where $||\bullet||_1$ denotes the L1-norm, i.e., the summation of the absolute value of all elements in the vector. The prior PDF in (11) has a three-fold meaning.

1) The DCT coefficients $\{\eta_i; i = 1, 2, \ldots, PQ\}$ have a high probability to equal zero. This, in turn, implies the sparseness of η.
2) The prior PDF in (11) treats each η_i equally. In other words, the prior PDF does not tell us which η_i is zero or non-zero. We need a "smart" algorithm to automatically find the non-zero coefficients based on a limited number of sampling points $\{g(x_m, y_m); m = 1, 2, \ldots, M\}$.
3) The independence assumption in (11) simply means that we do not know the correlation of η in advance. The correlation information will be taken into account by the posterior distribution [see (14)], once the measurement data are available.

Next, to derive the MAP algorithm, we need to introduce another important terminology, likelihood function, that is mathematically defined as the conditional probability $pdf\,(B\,|\eta)$. It models the relationship between the measurement data in B and the unknown DCT coefficients in η. Given the linear equation $A \cdot \eta = B$ in (5), the measurement data in B and the DCT coefficients in η must satisfy the linear equation $A \cdot \eta = B$ in (5). In other words, it is impossible to observe a set of variables η and B for which the equation $A \cdot \eta = B$ does not hold. Hence, the likelihood function is a Dirac delta function where the conditional probability $pdf(B\,|\eta)$ is non-zero if and only if $A \cdot \eta$ equals B

$$pdf\,(B|\eta) = \begin{cases} \infty & (A \cdot \eta = B) \\ 0 & (A \cdot \eta \neq B) \end{cases} \tag{12}$$

where

$$\int_{A \cdot \eta = B} pdf\,(B|\eta) \cdot dB = 1. \tag{13}$$

After defining the prior distribution in (11) and the likelihood function in (12), we are now ready to describe the MAP algorithm to uniquely determine the unknown DCT coefficients in η. The key idea of MAP is to find the optimal solution η that maximizes the posterior distribution, i.e., the conditional PDF $pdf(\eta\,|B)$. Namely, it aims to find the solution η that is most likely to occur. Based on Bayes' theorem [29], the posterior distribution $pdf(\eta|B)$ is proportional to the prior distribution $pdf(\eta)$ and the likelihood function $pdf(B|\eta)$

$$pdf\,(\eta|B) \propto pdf\,(\eta) \cdot pdf\,(B|\eta)\,. \tag{14}$$

Hence, the MAP algorithm attempts to solve the following optimization problem:

$$\underset{\eta}{\text{maximize}} \quad pdf\,(\eta) \cdot pdf\,(B|\eta). \tag{15}$$

In our case, the likelihood function is a Dirac delta function, as shown in (12). Therefore, maximizing the posterior probability in (14) is equivalent to maximizing the prior probability in (11) subject to the linear constraint $A \cdot \eta = B$

$$\begin{aligned} \underset{\eta}{\text{maximize}} \quad & 1/(2\lambda)^{PQ} \cdot \exp\left(-\|\eta\|_1/\lambda\right) \\ \text{subject to} \quad & A \cdot \eta = B. \end{aligned} \tag{16}$$

Since the exponential function $\exp(-||\eta||_1/\lambda)$ where $\lambda > 0$ monotonically decreases in $||\eta||_1$, the optimization in (16) can be re-written as

$$\begin{aligned} \underset{\eta}{\text{minimize}} \quad & \|\eta\|_1 \\ \text{subject to} \quad & A \cdot \eta = B. \end{aligned} \tag{17}$$

Note that the optimization in (17) is independent of the parameter λ in (11).

Equation (17) is referred to as L1-norm regularization in the literature [19]–[23]. To illustrate the connection between L1-norm regularization and sparse solution, we consider a simple 2-D example (i.e., $\eta = [\eta_1\ \eta_2]^T$), as shown in Fig. 4. In this example, the equality constraint only consists of one linear equation and, hence, the feasible space of the constrained optimization problem can be represented by a line $A \cdot \eta = B$ in the 2-D space. On the other hand, the contour lines of the cost function $||\eta||_1$ correspond to a number of rotated squares. It can be seen from Fig. 4 that the optimal solution solved

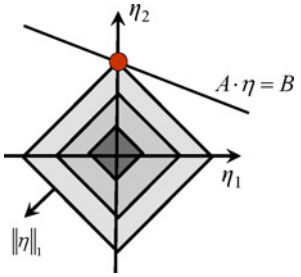


Fig. 4. Proposed L1-norm regularization results in a sparse solution η, as illustrated by the simple 2-D example.

by L1-norm regularization is located at one of the vertices of the contour lines. This observation implies that one of the coefficients (i.e., η_1 in this example) is exactly zero and, therefore, a sparse solution η is achieved.

Equation (17) can be converted to an equivalent linear programming problem and solved both robustly (i.e., with guaranteed global optimum) and efficiently (i.e., with low computational cost). The detailed algorithm of solving (17), as well as several other implementation issues, will be discussed in Section III.

C. Accuracy of MAP Estimation

Given the prior distribution in (11), the MAP estimation, i.e., the L1-norm regularization in (17), is statistically optimal, since it finds the solution η that maximizes the posterior probability, as shown in (15). However, it remains an open question if the accuracy of the MAP estimation can be quantitatively measured. In other words, we need to answer the following two questions.

1) Can the MAP estimation find the exact solution η for the underdetermined linear equation $A \cdot \eta = B$?
2) If the answer is yes, what are the sufficient conditions to guarantee the finding of the exact solution η?

In this subsection, we will answer these open questions by studying several important statistics theorems.

It has been proven in [20]–[23] that given the linear equation $A \cdot \eta = B$ in (5), the accuracy of the MAP estimation depends on the orthonormality of the column vectors of the matrix $A \in R^{M \times PQ}$. To intuitively illustrate this concept, we first consider a trivial case where the number of equations (i.e., M) equals the number of unknowns (i.e., PQ) and, hence, A is a square matrix. Furthermore, we assume that all column vectors of A are orthonormal, i.e., A is an orthogonal matrix with $A^T \cdot A = I$ where I is an identity matrix. In this trivial case, the exact solution η of $A \cdot \eta = B$ can be accurately determined as

$$\eta = A^T \cdot B. \tag{18}$$

In practice, since VP aims to predict the spatial variations from very few samples, the linear equation $A \cdot \eta = B$ in (5) is under determined and the matrix $A \in R^{M \times PQ}$ has more columns than rows (i.e., $M < PQ$). It is impossible for all columns of A to be orthonormal. In this case, it turns out that the solution η can be accurately found if the columns of A are *approximately* orthonormal. Based on the theorems of compressed sensing [20]–[23], the "orthonormality" of a matrix A can be quantitatively measured by its restricted isometry property (RIP).

Definition 1: A matrix A satisfies the restricted isometry property (RIP) of order K with constant $\delta_K < 1$, if the inequality

$$(1 - \delta_K) \cdot \|\eta\|_2^2 \leq \|A \cdot \eta\|_2^2 \leq (1 + \delta_K) \cdot \|\eta\|_2^2 \tag{19}$$

holds for every vector η that contains only K non-zero elements. In (19), $||\bullet||_2$ denotes the L2-norm, i.e., the square root of the summation of the squares of all elements in the vector.

If all columns of the matrix $A \in R^{M \times PQ}$ are almost orthonormal, RIP should be satisfied with a large K and a small δ_K. In the extreme case where A is exactly an orthogonal matrix, $||A.\eta||_2$ is equal to $||\eta||_2$ for every vector $\eta \in R^{PQ}$, since the linear transformation by an orthogonal matrix does not change the L2-norm of the vector η [26]. Hence, RIP is satisfied with $K = PQ$ and $\delta_K = 0$.

The concept of RIP has been successfully applied to assess the inherent difficulty of finding the exact solution η from the under determined linear equation $A \cdot \eta = B$ in (5). From example, the following theorem has been shown in [21].

Theorem 1: The L1-norm regularization in (17) *guarantees* to find the exact solution η of the underdetermined linear equation $A \cdot \eta = B$ in (5), if the following three conditions are all satisfied.

1) The solution vector η contains at most S non-zeros.
2) The matrix A satisfies the RIP of order $2S$ with constant $\delta_{2S} < 1$ and the RIP of order $3S$ with constant $\delta_{3S} < 1$.
3) The two RIP constants δ_{2S} and δ_{3S} further satisfy the inequality $\delta_{2S} + \delta_{3S} < 1$.

Note that the conditions in Theorem 1 are sufficient but not necessary. A number of other sufficient conditions have also been derived in the literature. More details can be found in [21].

While RIP offers a solid theoretical foundation to assess the accuracy of the MAP estimation, computing the RIP constant δ_K for a given matrix A is an NP-hard problem [20]–[23]. For this reason, an alternative metric, *coherence*, has been proposed to measure the orthonormality of a matrix A [22].

Definition 2: Given a matrix A for which every column vector has unit length (i.e., unit L2-norm), its coherence is defined as

$$\mu = \max_{i \neq j} \left| \left\langle A_i, A_j \right\rangle \right| \tag{20}$$

where A_i and A_j denote the ith and jth columns of A, respectively, and $<\bullet, \bullet>$ stands for the inner product of two vectors.

Similar to RIP, the coherence value μ in (20) offers a quantitative criterion to judge if the columns of the matrix A are approximately orthonormal. For instance, if all columns of A are orthonormal, the coherence value μ researches the minimum (i.e., zero); otherwise, the coherence value μ is always greater than zero.

While the RIP constant δ_K in (19) is difficult to compute, the coherence value μ in (20) can be easily calculated by the

inner product of column vectors. Once μ is known, the RIP constant δ_K is bounded by [22]

$$\delta_K \leq \mu \cdot (K-1) \tag{21}$$

where K denotes the order of RIP. In other words, while the exact value of the RIP constant δ_K is unknown, its upper bound can be efficiently estimated by coherence. This, in turn, offers a computationally tractable way to verify the sufficient conditions in Theorem 1. More details on coherence and its applications can be found in [22].

The aforementioned discussions summarize the theoretical framework to justify the accuracy of the MAP estimation. It demonstrates a number of sufficient conditions which guarantee to find the exact sparse solution η from the under determined linear equation $A \cdot \eta = B$. In our application of spatial variation characterization, the number of non-zeros in the vector η is not known in advance. Hence, it can be difficult to verify the conditions in Theorem 1 and then determine if the exact solution η is accurately solved. However, the theoretical results summarized in this subsection demonstrate the importance of column orthonormality for the matrix A in (5). This, in turn, motivates us to develop efficient techniques to improve the column orthonormality and, hence, enhance the accuracy of VP. The details of these implementation issues will be discussed in Section III.

III. Implementation Details

Our proposed VP technique is made of practical utility by carefully addressing several important implementation issues, including the following:

1) a column normalization scheme to improve the orthonormality of the matrix A for the underdetermined linear equation $A \cdot \eta = B$ in (5);
2) a linear programming formulation to efficiently solve the L1-norm regularization problem in (17);
3) a modified Latin hypercube sampling (M-LHS) scheme to randomly select the spatial sampling locations with small coherence;
4) a DCT coefficient pre-selection scheme to further improve the prediction accuracy by carefully removing non-critical high-frequency DCT coefficients.

In this section, we describe these implementation details and highlight their novelties.

A. Normalization

As discussed in Section II-C, all columns of the matrix A of the linear equation $A \cdot \eta = B$ should be approximately orthonormal so that the MAP estimation is capable of accurately finding the solution η. The requirement on orthonormality has a two-fold meaning. First, the columns of the matrix A should be approximately orthogonal. Second, all these columns should have unit length (i.e., unit L2-norm).

It is important to note that the column orthogonality of the matrix A cannot be enhanced by applying a simple orthogonalization algorithm, e.g. the Gram–Schmidt orthogonalization [26]. Such an orthogonalization process will change the solution vector η and compromise its unique sparse pattern.

The requirement on unit length, however, can be easily satisfied, if we normalize each column of the matrix A by its L2-norm. It has been demonstrated by the statistics community that the aforementioned normalization can efficiently improve the accuracy of the MAP estimation [22]. Hence, it is adopted in this paper and applied to (17), before the L1-norm regularization problem is solved by a numerical solver.

B. Linear Programming

Once all columns of the matrix A are normalized to unit length, a numerical solver should be applied to solve the L1-norm regularization problem in (17) and find the optimal solution η. From (17), we would notice that the cost function $||\eta||_1$ is not smooth and, hence, it cannot be easily minimized by a simple gradient-based algorithm [28]. To address this issue, we introduce a set of slack variables $\{\theta_i; i = 1, 2, \ldots, PQ\}$ and re-write (17) as

$$\begin{aligned} \underset{\eta,\theta}{\text{minimize}} \quad & \theta_1 + \theta_2 + \cdots + \theta_{PQ} \\ \text{subject to} \quad & A \cdot \eta = B \\ & -\theta_i \leq \eta_i \leq \theta_i \qquad (i = 1, 2, \cdots, PQ). \end{aligned} \tag{22}$$

Intuitively, by minimizing the cost function in (22), all constraints $\{-\theta_i \leq \eta_i \leq \theta_i;\ i = 1, 2, \ldots, PQ\}$ will become active, i.e., $\{|\eta_i| = \theta_i;\ i = 1, 2, \ldots, PQ\}$. For this reason, the optimizations in (17) and (22) are equivalent, i.e., they share the same optimal solution η. This conclusion can be formally proven based on the Karush–Kuhn–Tucker condition from optimization theory [28].

Note that both the cost function and the constraints in (22) are linear. Therefore, it is a linear programming problem and can be solved both robustly (i.e., with guaranteed global optimum) and efficiently (i.e., with low computational cost), e.g., by using the interior-point method [28]. For large-scale problems, there exist a number of fast algorithms (e.g., [18]) that can solve (22) with millions of variables in a few minutes.

C. Latin Hypercube Sampling

The study in Section II-C shows that the accuracy of the MAP estimation depends on the orthonormality of the matrix A in (5). According to the definition of the matrix A, it can be easily seen that the value of A is determined by the sampling locations $\{(x_m, y_m);\ m = 1, 2, \ldots, M\}$. In other words, different choices of sampling locations will provide different values of the matrix A and, hence, different results of the MAP estimation. It, in turn, motivates us to develop an efficient algorithm to find a set of "good" sampling locations. As such, the orthonormality of the matrix A is well approximated, thereby resulting in high prediction accuracy for the MAP estimation.

While directly optimizing the orthonormality of the matrix A, e.g., minimizing the coherence value μ in (20), is not trivial, it has been proven that random sampling is able to result in a good matrix A [20]–[23]. In particular, the theoretical results in [20]–[23] demonstrate that if the vector $\eta \in R^{PQ}$ contains at most S ($S << PQ$) non-zeros and M

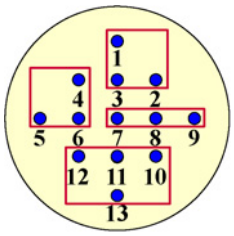


Fig. 5. Simple partition example for the M-LHS algorithm where 13 possible sampling locations are labeled as $\{(x_n, y_n), n = 1, 2, \ldots, 13\}$ and divided into four subsets $\{(x_n, y_n), n = 1, 2, 3\}$, $\{(x_n, y_n), n = 4, 5, 6\}$, $\{(x_n, y_n), n = 7, 8, 9\}$, and $\{(x_n, y_n), n = 10, 11, 12, 13\}$.

sampling locations are randomly selected where M is in the order of $O(S{\cdot}\log(PQ))$, the sufficient conditions in Theorem 1 are almost guaranteed to hold (i.e., with probability nearly equal to 1), implying that the exact value of η can be accurately determined with extremely high probability. To the best of our knowledge, there is no other sampling scheme that clearly outperforms random sampling.

Based on these observations, we adopt the random sampling strategy in this paper. Our objective is to evenly distribute M random sampling points over the entire wafer/chip. To achieve this goal, we borrow the idea of Latin hypercube sampling (LHS) from the statistics community [24] and develop a M-LHS algorithm to generate well-controlled random samples.

Starting from N possible sampling locations on a wafer/chip, we first label each sampling location by an index $n \in \{1, 2, \ldots, N\}$. For illustration purposes, Fig. 5 shows a simple example where 13 sampling locations are sequentially labeled. This simple example reveals two important properties of our labeling scheme. First, if two different sampling locations (x_i, y_i) and (x_j, y_j) are in the same neighborhood, their indexes i and j should be close to each other. Second, the number of possible sampling locations (i.e., N) can be less than the total number of DCT coefficients (i.e., PQ), since the shape of a wafer is close to a circle, instead of a rectangle.

Next, to evenly distribute the M sampling locations among the N possible choices, we partition the index set $\{1, 2, \ldots, N\}$ into M non-overlapped subsets. There are two possible scenarios when we construct these subsets.

1) M is a factor of N (i.e., M divides N without leaving a remainder). In this case, the partition of the set $\{1, 2, \ldots, N\}$ is simply determined by the order of the indexes. Namely, the M subsets are: $\{1, 2, \ldots, N/M\}$, $\{N/M + 1, N/M + 2, \ldots, 2.N/M\}$, and so on.
2) M is not a factor of N and M divides N leaving a non-zero remainder R. In this case, we again divide the set $\{1, 2, \ldots, N\}$ into M non-overlapping subsets based on the order of the indexes. However, the sizes of all subsets are different: either $\lfloor N/M \rfloor$(the largest integer that is less than or equal to N/M) or $\lfloor N/M \rfloor + 1$. The choices between $\lfloor N/M \rfloor$ and $\lfloor N/M \rfloor + 1$ are randomly selected for each subset with the constraint that only R subsets have the size of $\lfloor N/M \rfloor + 1$ and the other $M - R$ subsets have the size of $\lfloor N/M \rfloor$. Hence, each of the N possible sampling locations belongs to one of the M subsets.

Algorithm 1 Modified Latin hypercube sampling (M-LHS)

1: Start from N possible sampling locations on a wafer/chip.
2: Label each sampling location by an index $n \in \{1, 2, \ldots, N\}$ so that the indexes i and j of two different sampling locations (x_i, y_i) and (x_j, y_j) are close to each other if (x_i, y_i) and (x_j, y_j) are in the same neighborhood.
3: Partition the index set $\{1, 2, \ldots, N\}$ into M non-overlapped subsets. If M is a factor of N, the partition is simply determined by the order of the indexes. Otherwise, if M divides N leaving a non-zero remainder R, the set $\{1, 2, \ldots, N\}$ is divided into M non-overlapped subsets based on the order of the indexes where only R subsets have the size of $\lfloor N/M \rfloor + 1$ and the other $M - R$ subsets have the size of $\lfloor N/M \rfloor$.
4: Randomly select one sampling location from each of the M subsets, resulting in M sampling locations in total.

Fig. 5 shows a simple partition example with 13 possible sampling locations and four subsets. In this case, since $M = 4$ is not a factor of $N = 13$, the sizes of these four subsets are not identical. Studying Fig. 5, we would notice that each of the M subsets generated by our partition scheme contains the sampling locations in the same neighborhood. In other words, the M subsets conceptually represent M local clusters spatially distributed over the wafer/chip.

As the final step of the proposed M-LHS algorithm, we randomly select one sampling location from each of the M subsets, thereby resulting in M sampling locations in total. Algorithm 1 summarizes the major steps of our M-LHS algorithm. Unlike the traditional LHS algorithm that aims to sample continuous random variables [24], Algorithm 1 has been particularly tuned to randomly select M choices out of N possible candidates. Hence, it is referred to as M-LHS in this paper.

Compared to brute force random sampling, M-LHS guarantees to distribute the M sampling locations over different local regions on the wafer/chip. Namely, it eliminates the possibility that many sampling locations are selected from the same local space. This, in turn, leads to superior prediction accuracy over brute-force random sampling, as will be demonstrated by the numerical examples in Section V.

D. DCT Coefficient Pre-Selection

It is shown in the previous subsection that the number of samples required for VP depends on the logarithm of the total number of DCT coefficients, i.e., $\log(PQ)$. If we know that a subset of DCT coefficients must be zero, we no longer need to consider these DCT coefficients as problem unknowns. Given a limited number of sampling points, such a pre-selection strategy for DCT coefficients can further improve the accuracy of the proposed VP algorithm.

As is demonstrated in the literature [15], spatial patterns of process variations are often smooth. It, in turn, implies that the spatial variation patterns may be accurately represented by a few dominant DCT coefficients at low frequencies. Namely, the low-frequency DCT coefficients are more important than the high-frequency ones when predicting spatial variations.

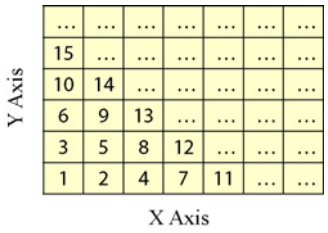


Fig. 6. Importance ranking of DCT coefficients defined in [17] where the rank "1" corresponds to the most important DCT coefficient.

Motivated by this observation, we follow the idea proposed in [17] to rank the importance of all DCT coefficients, as shown in Fig. 6. When solving the linear equation $A \cdot \eta = B$ by (17), we only consider the first K low-frequency DCT coefficients as problem unknowns. All other high-frequency DCT coefficients are simply set to zero. In other words, the optimization in (17) only needs to solve the sparse solution for K (instead of PQ) problem unknowns.

In our implementation, cross-validation [29] is further used to estimate the optimal value of K. An F-fold cross-validation partitions the sampling points into F groups. Prediction error is estimated from F independent runs. In each run, one of the F groups is used to estimate the prediction error and all other groups are used to solve the L1-norm regularization problem in (17) to determine the unknown DCT coefficients. Note that the training data for coefficient estimation and the testing data for error estimation are not overlapped. Hence, over-fitting can be easily detected. In addition, different groups are selected for error estimation in different runs. As such, each run results in an error value ε_f ($f = 1, 2, \ldots, F$) that is measured from a unique set of testing data. The final prediction error is computed as the average of $\{\varepsilon_f;\ f = 1, 2, \ldots, F\}$, i.e., $\varepsilon = (\varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_F)/F$. The optimal value of K is determined to minimize the cross-validation error ε.

E. Summary

Algorithm 2 summarizes the major steps of the proposed VP method. It starts from very few (i.e., M) sampling locations $\{g(x_m, y_m); m = 1, 2, \ldots, M\}$ determined by M-LHS. Next, it formulates an underdetermined linear equation based on these measurement data, and solves all DCT coefficients. Finally, the spatial variations $\{g(x, y); x = 1, 2, \ldots, P, y = 1, 2, \ldots, Q\}$ are recovered by the IDCT in (4).

In summary, the proposed VP method offers a number of important advantages over other traditional techniques.

1) *Low cost*: VP is developed to minimize the number of test structures required to fully extract the spatial variation information. It, in turn, reduces the testing and measurement cost, e.g., area overhead, testing and characterization time, yield loss during testing, etc. In addition, the VP formulation in (22) is a linear programming problem and it can be solved both robustly (i.e., with guaranteed global optimum) and efficiently (i.e., with low computational cost).
2) *High accuracy*: the accuracy of VP is guaranteed by the theoretical studies from the statistics community [19]–[23], as discussed in Section II-C. Namely, with several general assumptions, VP can fully reconstruct the spatial variations with probability nearly equal to 1. In addition, the accuracy of VP can be verified in real time by the cross-validation method mentioned in Section III-D. This error estimation scheme is extremely important, since it provides a practical, quantitative criterion to determine whether the result of VP is sufficiently accurate or not. Additional sampling points can be further collected to improve accuracy, until the prediction error is sufficiently small.
3) *General purpose*: VP can be used to predict the spatial pattern of both inter-die and spatially-correlated intra-die variations. The prediction by VP is based on the measurement data collected from the current wafer/chip only. It does not require any historical data for training and, hence, can efficiently handle the non-stationary effects, e.g., process drifting caused by equipment aging. The only assumption posed by VP is that the spatial variations must have a sparse representation in frequency domain. This assumption is valid for spatially-correlated process variations. In other words, the variations of interest are not dominated by independent random mismatches (e.g. random dopant fluctuations). As will be demonstrated by the experimental examples in Section V, such a sparseness assumption holds for a number of performance variations (e.g., ring oscillator delay, full-chip leakage, and so on). The impact of independent random mismatches on these performance metrics is averaged out and, hence, becomes non-dominant. In practice, the sparseness assumption can be verified by the error estimation scheme we previously mentioned. Namely, if the frequency-domain representation is not sparse, we will observe a large prediction error reported by VP.

Algorithm 2 Virtual probe (VP)

1: Select M sampling locations $\{(x_m, y_m);\ m = 1, 2, \ldots, M\}$ by M-LHS (i.e., Algorithm 1).
2: Collect the measurement data $\{g(x_m, y_m);\ m = 1, 2, \ldots, M\}$ at these locations.
3: Formulate the underdetermined linear equation $A \cdot \eta = B$ in (5)-(9) for the first K DCT coefficients where the optimal value of K is determined by cross-validation (see Section III-D).
4: Normalize all columns of the matrix A (see Section III-A) and formulate the linear programming problem in (22).
5: Solve the optimization problem in (22) to determine η, i.e., the first K DCT coefficients. All other DCT coefficients are set to zero.
6: Apply the IDCT in (4) to recover the spatial variations $\{g(x, y);\ x = 1, 2, \ldots, P,\ y = 1, 2, \ldots, Q\}$ across the wafer/chip.

IV. Applications of VP

The proposed VP method can be applied to a broad range of applications related to integrated circuits. In this section,

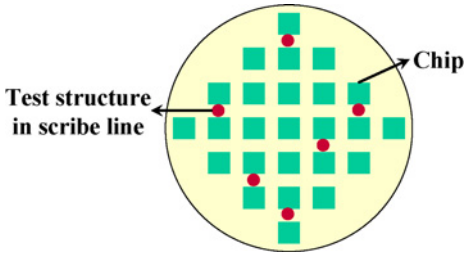


Fig. 7. Test structures are deployed in wafer scribe lines to measure and characterize inter-die variations at wafer level.

we will briefly discuss these possible applications, including: 1) wafer-level silicon characterization (for inter-die variations); 2) chip-level silicon characterization (for intra-die variations); and 3) testing and self-healing of integrated circuits. Note that the main objective of this section is to motivate several possible future research directions based upon our VP technique. The details of these new research problems are beyond the scope of this paper and, hence, are not discussed here.

A. Wafer-Level Silicon Characterization

To characterize parametric variations at wafer level (i.e., inter-die variations), test structures are deployed in wafer scribe lines [11]–[13], as shown in Fig. 7. These test structures do not have area overhead, as they are not within a product chip. However, it does not simply mean that the characterization is free. Instead, wafer-level characterization can still be expensive due to the following two reasons.

First, test structures in scribe lines must be measured by wafer probe test, as these devices will be completely destroyed during wafer dicing before packaging. Within this testing process, a probe card will contact the I/O pads of the test structures to measure currents, voltages or frequencies. Such a wafer probe testing, however, is not perfectly safe. It may break the wafer being tested due to mechanical stress, create additional yield loss, and eventually increase manufacturing cost. Second, wafer probe test (e.g., aligning the probe card with the I/O pads and collecting all measurement data) is time-consuming. This, in turn, further increases manufacturing cost, as the overall manufacturing time is increased.

For these two reasons, it is crucial to reduce the number of measured test structures so that the overall testing and characterization cost is minimized. Our proposed VP method perfectly fits this need. Namely, we propose to deploy and measure very few test structures randomly distributed over the scribe lines of a wafer. Once the measurement data are collected, Algorithm 2 is applied to reconstruct the spatial variations across the wafer. Note that since the test structures are constrained within scribe lines, the aforementioned wafer-level characterization may not provide sufficient resolution to predict intra-die variations. It, therefore, implies that additional test structures are required for chip-level silicon characterization, as will be discussed in the next subsection.

B. Chip-Level Silicon Characterization

On-chip test structures are typically used to characterize intra-die variations at chip level [11]–[13], as shown in Fig. 8.

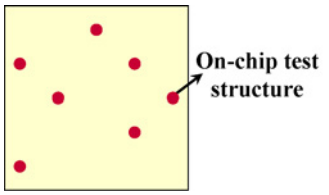


Fig. 8. Test structures are deployed within a product chip to measure and characterize intra-die variations at chip level.

The cost of chip-level characterization consists of two major portions: 1) area overhead, and 2) testing time.

First, on-chip test structures are deployed within a product chip at a number of pre-selected locations. If too many test structures are used, they lead to significant area overhead and, hence, become financially intractable. Second, all on-chip test structures must be measured through a limited number of I/O pads. This testing process is time-consuming and directly increases manufacturing cost.

Motivated by these observations, we propose to deploy and measure very few on-chip test structures and then apply VP to reconstruct the complete spatial variation pattern at chip level. As such, the characterization cost is substantially reduced.

C. Beyond Silicon Characterization

The silicon characterization results extracted by VP can be efficiently applied to a number of practical applications. In this subsection, we briefly discuss two important application examples: 1) speed binning, and 2) post-silicon tuning.

In traditional speed binning, all manufactured chips are tested individually to determine the maximum operation frequency [27]. This is expensive, since each chip must be repeatedly tested with different speed setups. Given the proposed VP framework, we can potentially test a small number of chips to find their speed bins, and then use VP to predict the speed of other chips on the same wafer. Note that even if the prediction by VP is not exact, it can still be used to optimize the testing scheme to reduce cost. For instance, if the speed of an untested chip is estimated by VP, the speed test should start from the nearest bin since this chip is most likely to fall in that speed bin. Such a strategy helps us to find the appropriate speed bin quickly and, hence, reduce testing cost.

On the other hand, post-silicon tuning is a recently-developed technique to improve parametric yield in the presence of large-scale process variations [8]–[10]. It adaptively configures a number of tunable parameters (e.g., supply voltage, body bias, and so on) so that a given circuit can work properly under different process conditions. An important component of post-silicon tuning is to accurately measure the process condition of a given chip so that the tunable parameters can be appropriately configured to adjust the circuit behavior. Such measurement, however, is not trivial, since it often requires a large number of on-chip "sensors." We believe that the proposed VP framework can be used to predict the process condition from a significantly reduced number of on-chip sensors. By minimizing the number of required sensors, both the design complexity and the manufacturing cost can be reduced.

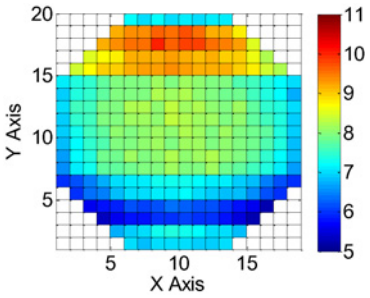


Fig. 9. Measured flush delay values (normalized by a randomly selected constant) of 282 industrial chips from the same wafer show significant spatial variations.

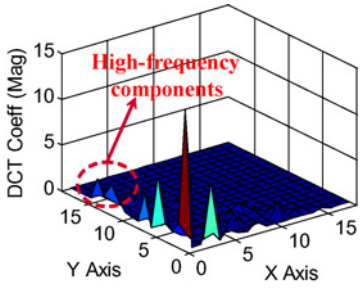


Fig. 10. DCT coefficients (magnitude) of the normalized flush delay measurement show a unique sparse pattern.

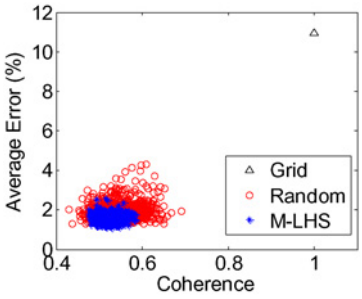


Fig. 11. Coherence and average error calculated from 128 chips with 1000 repeated runs for three different sampling techniques: 1) grid sampling (grid); 2) brute-force random sampling (random); and 3) M-LHS.

TABLE I

STATISTICS OF COHERENCE AND AVERAGE ERROR CALCULATED FROM 128 CHIPS WITH 1000 REPEATED RUNS

Method	Coherence		Average Error (%)	
	Mean	Std	Mean	Std
Grid	1.000	N/A	10.96	N/A
Random	0.545	0.040	1.90	0.41
M-LHS	0.521	0.020	1.54	0.20

V. NUMERICAL EXPERIMENTS

In this section, we demonstrate the efficacy of VP using several examples based on industrial measurement data. All numerical experiments are performed on a 2.8 GHz Linux server.

A. Flush Delay Measurement Data

We consider the flush delay values measured from 282 industrial chips on the same wafer, as shown in Fig. 9. In this example, the measured delay significantly varies from chip to chip due to process variations. Our goal is to capture these wafer-level delay variations. We use a 2-D function $g(x, y)$ to model the delay, where $x \in \{1, 2, \ldots, 18\}$ and $y \in \{1, 2, \ldots, 19\}$. Each coordinate point (x, y) corresponds to a chip. Next, we apply a 2-D DCT to $g(x, y)$, yielding the frequency-domain components $G(u, v)$ shown in Fig. 10.

Two important observations can be made from the result in Fig. 10. First, $G(u, v)$ contains substantial high-frequency components, implying that the spatial sampling rate cannot be drastically reduced according to the well-known Nyquist–Shannon sampling theorem. Second, $G(u, v)$ is sparse, as its magnitude is almost zero at a large number of frequencies. This sparse pattern is the essential necessary condition that makes the proposed VP framework applicable to this example.

In what follows, we first use the data set in Fig. 9 to compare the modified Latin hypercube sampling algorithm (i.e., M-LHS summarized in Algorithm 1) with two brute-force sampling methods (i.e., grid sampling and random sampling), thereby demonstrating the superior accuracy achieved by M-LHS. Next, we further apply the proposed VP technique (i.e., Algorithm 2) and several traditional methods to predict the spatial delay variations and compare the accuracy of these different approaches.

1) *Spatial Sample Generation:* For testing and comparison purposes, we implement three different sampling schemes to select the spatial locations for silicon testing: 1) grid sampling; 2) brute-force random sampling; and 3) M-LHS. Grid sampling deterministically picks up a set of spatial locations from a uniform 2-D grid. Brute-force random sampling simply selects random spatial locations by using a pseudo-random number generator. Finally, M-LHS follows the partition and selection steps summarized in Algorithm 1 to determine random sampling locations.

We apply the aforementioned three sampling schemes to select 128 chips out of 282 possible candidates on the same wafer. Next, we collect the flush delay data for these 128 selected chips and apply the MAP estimation (see Algorithm 2) to predict the spatial variations of the entire wafer. In Step 5 of Algorithm 2, the linear optimization is efficiently solved by the l1-MAGIC package developed by the California Institute of Technology, Pasadena, taking about 8s to finish in this example. Such an experiment is repeated for 1000 times in order to accurately estimate the statistics of prediction error.

Fig. 11 shows the coherence (see Definition 2) and the average error for these 1000 repeated runs. As discussed in Section II-C, the value of coherence provides a quantitative measure to assess the orthonormality of the matrix A in (5). On the other hand, the average error of the MAP estimation is calculated by

$$Error_{AVG} = \sqrt{\frac{\sum_x \sum_y \left[g(x, y) - \tilde{g}(x, y)\right]^2}{\sum_x \sum_y \left[g(x, y)\right]^2}} \tag{23}$$

where $g(x, y)$ and $\tilde{g}(x, y)$ denote the exact value and the estimated value of the flush delay at the location (x, y),

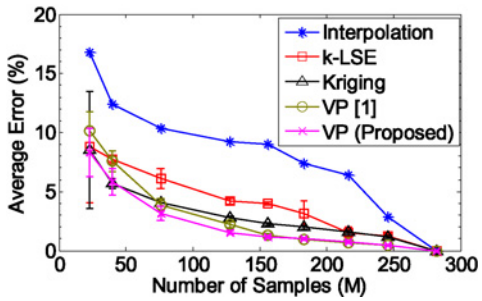


Fig. 12. Average prediction error (both mean and standard deviation) of different algorithms estimated by 100 repeated runs.

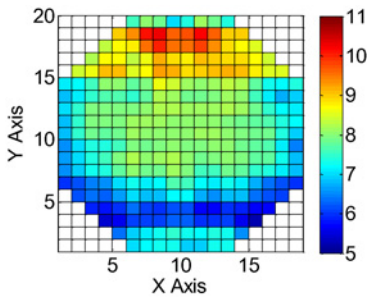


Fig. 13. Flush delay values predicted from 60 tested chips by the proposed VP algorithm.

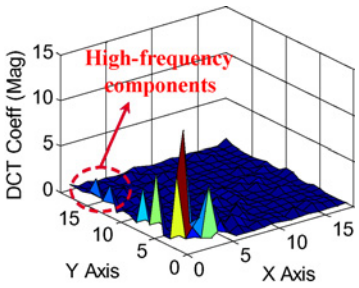


Fig. 14. Proposed VP algorithm accurately captures the low-frequency and high-frequency DCT coefficients by using 60 tested chips.

respectively. Note that since grid sampling is deterministic, all 1000 runs yield the same result. Hence, only one data point is plotted in Fig. 11. Table I further shows the statistics of both coherence and error calculated from these 1000 repeated runs.

Studying Fig. 11 and Table I, we would have two important observations. First, grid sampling results in a large coherence value. It, in turn, implies that the columns of the matrix A in (5) are not approximately orthonormal. According to the discussions in Section II-C, the MAP estimation cannot accurately predict the spatial variations in this case. This conclusion is consistent with the results in Fig. 11 where the average error associated with grid sampling is extremely large.

Second, both brute-force random sampling and M-LHS yield small coherence values and, consequently, small prediction errors. These results demonstrate the fact that a randomized algorithm can efficiently generate good spatial samples. In addition, comparing brute-force random sampling with M-LHS, we would notice that both methods result in similar mean values for coherence and error. However, M-LHS is able to reduce the standard deviation of both coherence and error by about $2\times$ in this example. Such a reduction in standard deviation occurs, because M-LHS well controls the random samples by the partition and selection steps summarized in Algorithm 1. This is an important benefit offered by M-LHS, since it reduces the probability that a set of "bad" samples are randomly selected and, hence, result in large prediction error.

2) *Spatial Variation Prediction:* To quantitatively evaluate the accuracy of the proposed VP technique, we repeatedly apply Algorithm 2 with M-LHS to predict the wafer-level spatial variations with different numbers of spatial samples. For testing and comparison purposes, we implemented a number of traditional methods: 1) the 2-D interpolation method with uniform grid sampling [31]; 2) the Kriging method with exponential correlation function [16]; 3) the k-LSE method based on DCT analysis [17]; and 4) the simple VP implementation without DCT coefficient pre-selection [1].

Fig. 12 shows the average error calculated by (23) as a function of the number of samples (i.e., M) for different algorithms. M-LHS is applied to select the sampling locations for all methods except 2-D interpolation. To account for the inherent randomness of M-LHS sampling, we repeatedly run each algorithm for 100 times and plot the mean and the standard deviation of the average error in Fig. 12. Note that Algorithm 2 achieves the highest accuracy in this example. Compared to the simple VP implementation developed in [1], our new implementation (i.e., Algorithm 2) achieves superior accuracy by carefully pre-selecting the important DCT coefficients at low frequencies. Such a pre-selection scheme is particularly important, if the number of available samples (i.e., M) is small and, hence, it is difficult to accurately find the non-zeros from all DCT coefficients by L1-norm regularization. On the other hand, the proposed VP algorithm outperforms other traditional techniques (i.e., 2-D interpolation, Kriging prediction, and k-LSE estimation), because all these traditional methods assume a smooth spatial variation pattern and, therefore, cannot accurately capture the high-frequency components of our measurement data shown in Fig. 10.

Fig. 13 shows the flush delay values predicted from 60 tested chips (i.e., $M = 60$) by the proposed VP algorithm. In this example, 120 DCT coefficients are pre-selected by the cross-validation algorithm in Section III-D, before the final L1-norm regularization step is applied. Fig. 14 further plots the DCT coefficients associated with the spatial variation pattern in Fig. 13. Comparing Figs. 10 and 14, we would notice that both the low-frequency and the high-frequency DCT coefficients are accurately captured in this example.

To quantitatively assess the prediction accuracy of each chip, we calculate the following relative error:

$$Error_{REL}(x, y) = \left| \frac{g(x, y) - \tilde{g}(x, y)}{g(x, y)} \right| \qquad (24)$$

where $g(x, y)$ and $\tilde{g}(x, y)$ are similarly defined as those in (23). The error metric in (24) measures the difference between the measurement data (i.e., Fig. 9) and the prediction results (i.e., Fig. 13) for every chip. Fig. 15 shows the histogram of the relative error calculated for all chips on the same wafer. Note that the relative error is less than 10% for most chips in this example.

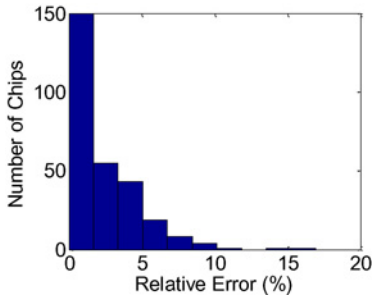


Fig. 15. Histogram of the relative error of the proposed VP algorithm calculated by (24) for all chips on the same wafer.

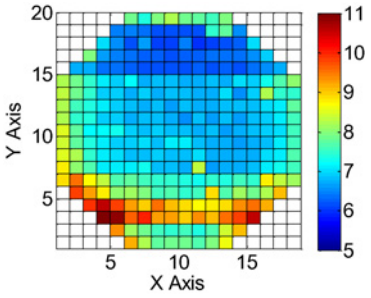


Fig. 16. Measured leakage current values $\log_{10}(I_{LEAK})$ (normalized by a randomly selected constant) of 282 industrial chips from the same wafer show significant spatial variations.

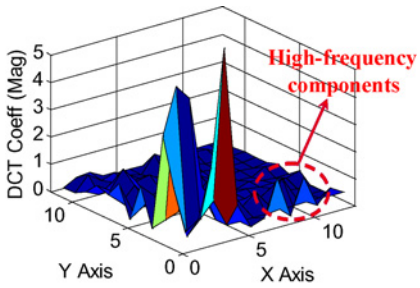


Fig. 17. DCT coefficients (magnitude) of the normalized leakage current measurement $\log_{10}(I_{LEAK})$ show a unique pattern that is approximately sparse.

B. Leakage Current Measurement Data

We consider the leakage current measurement data collected by IDDQ test for the same industrial circuit design. Fig. 16 shows the normalized leakage current values $\log_{10}(I_{LEAK})$ (after logarithmic transform) as a function of the location (x, y). Fig. 17 further shows the frequency-domain components after DCT. Similar to the flush delay example, the DCT coefficients contain important high-frequency components. In addition, a large number of small DCT coefficients are observed and, hence, the frequency-domain representation is approximately (but not exactly) sparse. This observation is consistent with the fact that the full-chip leakage current partially depends on V_{TH} mismatches that are not spatially correlated.

We apply different algorithms to predict the spatial variations based on a few (i.e., M) sampling points. Fig. 18 shows the average error calculated by (23). Similar to the previous example, both the mean and the standard deviation of the average error are calculated from 100 repeated runs and they are plotted in Fig. 18. Note that the proposed VP method (i.e., Algorithm 2) achieves better accuracy than three traditional

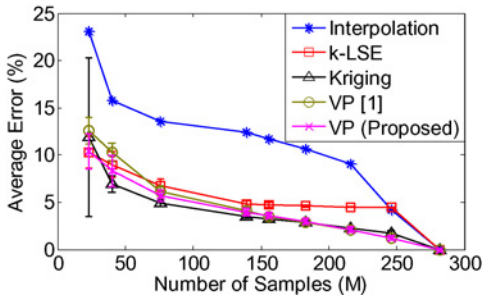


Fig. 18. Average prediction error (both mean and standard deviation) of different algorithms estimated by 100 repeated runs.

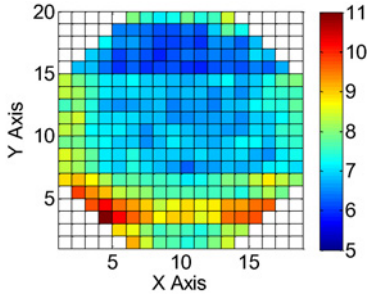


Fig. 19. Leakage current values $\log_{10}(I_{LEAK})$ predicted from 100 tested chips by the proposed VP algorithm.

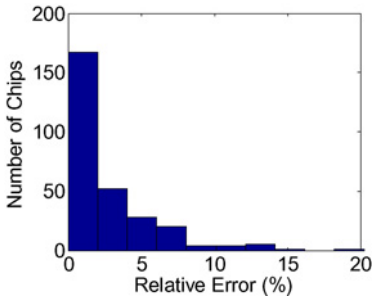


Fig. 20. Histogram of the relative error of the proposed VP algorithm calculated by (24) for all chips on the same wafer.

techniques: 1) 2-D interpolation; 2) k-LSE estimation; and 3) the simple VP implementation in [1]. However, the accuracy of VP is slightly worse than Kriging prediction, because the frequency-domain representation is not exactly sparse in this example.

Fig. 19 shows the leakage current values $\log_{10}(I_{LEAK})$ (after logarithmic transform) predicted from 100 tested chips by the proposed VP algorithm. In this example, 240 DCT coefficients are pre-selected by the cross-validation algorithm in Section III-D, before the final L1-norm regularization step is applied. Fig. 20 further shows the histogram of the relative error calculated for all chips using (24). Note that the relative error is less than 10% for most chips in this example.

C. Ring Oscillator Period Measurement Data

We consider the ring oscillator (RO) period measurement data collected from a wafer at an advanced technology node. These RO measurement data are strongly correlated with the final chip performance and, hence, are often used for process monitoring and control [11], [12]. Our wafer contains 117 ROs distributed over different spatial locations. Fig. 21 shows

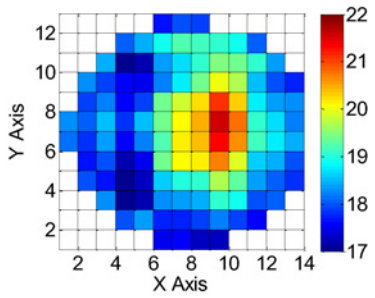


Fig. 21. Measured RO period values (normalized by a randomly selected constant) of 117 ROs from the same wafer show significant spatial variations.

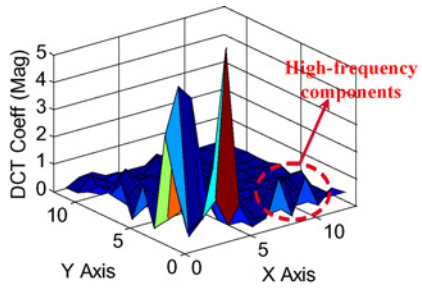


Fig. 22. DCT coefficients (magnitude) of the normalized RO period show a unique pattern that is approximately sparse.

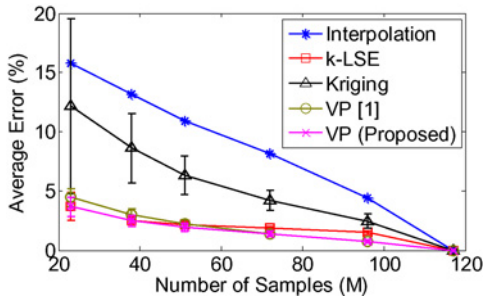


Fig. 23. Average prediction error (both mean and standard deviation) of different algorithms estimated by 100 repeated runs.

the normalized RO period values as a function of the location (x, y). Fig. 22 further shows the frequency-domain components after DCT. Similar to the leakage current example, the DCT coefficients are approximately sparse and contain significant high-frequency components.

We apply different algorithms to recover the spatial variations based on a few (i.e., M) sampling points. Fig. 23 compares the average error calculated by (23) for different methods. Both the mean and the standard deviation of the average error are calculated from 100 repeated runs and they are plotted in Fig. 23. Note that the proposed VP technique (i.e., Algorithm 2) achieves the best accuracy in this example. The Kriging method shows large error, because it assumes an exponential correlation model while the actual spatial correlation does not match the model template. In general, the Kriging method needs to know the correlation template in advance. If the prior knowledge of the correlation function is not correct, the Kriging method may fail to predict the spatial variations accurately.

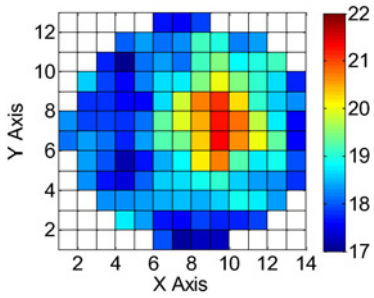


Fig. 24. RO period values predicted from 40 tested ROs by the proposed VP algorithm.

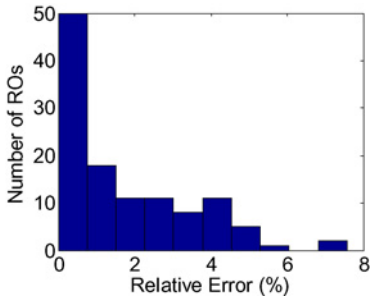


Fig. 25. Histogram of the relative error of the proposed VP algorithm calculated by (24) for all ring oscillators on the same wafer.

Fig. 24 further shows the RO period values predicted from 40 tested ROs by the proposed VP algorithm. The cross-validation algorithm in Section III-D pre-selects all DCT coefficients in this example. Fig. 25 shows the histogram of the relative error calculated for all ROs using (24). Note that the relative error is less than 5% for most chips in this example.

VI. Conclusion

In this paper, we proposed a novel VP framework to efficiently and accurately recover full-wafer/chip spatial variations from an extremely small set of measurement data, thereby reducing the cost of silicon characterization and testing. VP exploits recent breakthroughs in compressed sensing [20]–[23]. It is formulated as a MAP problem and can be efficiently solved via linear programming. Our numerical examples based on industrial measurement data demonstrate that VP provides superior accuracy over other traditional methods, including 2-D interpolation, Kriging prediction, and k-LSE estimation.

References

[1] X. Li, R. Rutenbar, and R. Blanton, "Virtual probe: A statistically optimal framework for minimum-cost silicon characterization of nanoscale integrated circuits," in *Proc. Int. Conf. Comput.-Aided Des.*, 2009, pp. 433–440.

[2] S. Nassif, "Delay variability: Sources, impacts and trends," in *Proc. Int. Solid-State Circuits Conf.*, 2000, pp. 368–369.

[3] Semiconductor Industry Associate. (2007). *International Technology Roadmap for Semiconductors* [Online]. Available: www.itrs.net/Links/2007ITRS/Home2007.htm

[4] H. Chang and S. Sapatnekar, "Statistical timing analysis under spatial correlations," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 24, no. 9, pp. 1467–1482, Sep. 2005.

[5] C. Visweswariah, K. Ravindran, K. Kalafala, S. Walker, and S. Narayan, "First-order incremental block-based statistical timing analysis," in *Proc. Des. Autom. Conf.*, 2004, pp. 331–336.

[6] Y. Zhan, A. Strojwas, X. Li, L. Pileggi, D. Newmark, and M. Sharma, "Correlation aware statistical timing analysis with non-Gaussian delay distributions," in *Proc. Des. Autom. Conf.*, 2005, pp. 77–82.

[7] K. Heloue and F. Najm, "Statistical timing analysis with two-sided constraints," in *Proc. Int. Conf. Comput.-Aided Des.*, 2005, pp. 829–836.

[8] M. Mani, A. Singh, and M. Orshansky, "Joint design-time and post-silicon minimization of parametric yield loss using adjustable robust optimization," in *Proc. Int. Conf. Comput.-Aided Des.*, 2006, pp. 19–26.

[9] S. Kulkarni, D. Sylvester, and D. Blaauw, "A statistical framework for post-silicon tuning through body bias clustering," in *Proc. Int. Conf. Comput.-Aided Des.*, 2006, pp. 39–46.

[10] Q. Liu and S. Sapatnekar, "Synthesizing a representative critical path for post-silicon delay prediction," in *Proc. Int. Symp. Phys. Des.*, 2009, pp. 183–190.

[11] M. Ketchen, M. Bhushan, and D. Pearson, "High speed test structures for in-line process monitoring and model calibration," in *Proc. IEEE Int. Conf. Microelectron. Test Structures*, Apr. 2005, pp. 33–38.

[12] M. Bhushan, A. Gattiker, M. Ketchen, and K. Das, "Ring oscillators for CMOS process tuning and variability control," *IEEE Trans. Semicond. Manuf.*, vol. 19, no. 1, pp. 10–18, Feb. 2006.

[13] W. Mann, F. Taber, P. Seitzer, and J. Broz, "The leading edge of production wafer probe test technology," in *Proc. Int. Test Conf.*, 2004, pp. 1168–1195.

[14] F. Koushanfar, P. Boufounos, and D. Shamsi, "Post-silicon timing characterization by compressed sensing," in *Proc. Int. Conf. Comput.-Aided Des.*, 2008, pp. 185–189.

[15] S. Reda and S. Nassif, "Analyzing the impact of process variations on parametric measurements: Novel models and applications," in *Proc. Des. Autom. Test Eur. Conf.*, 2009, pp. 375–380.

[16] F. Liu, "A general framework for spatial correlation modeling in VLSI design," in *Proc. Des. Autom. Conf.*, 2007, pp. 817–822.

[17] A. Nowroz, R. Cochran, and S. Reda, "Thermal monitoring of real processors: Techniques for sensor allocation and full characterization," in *Proc. Des. Autom. Conf.*, 2010, pp. 56–61.

[18] S. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large-scale l1-regularized least squares," *IEEE J. Sel. Top. Signal Process.*, vol. 1, no. 4, pp. 606–617, Dec. 2007.

[19] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. Roy. Statist. Soc.*, vol. 58, no. 1, pp. 267–288, 1996.

[20] D. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[21] E. Candes, "Compressive sampling," in *Proc. Int. Congr. Math.*, vol. 3. 2006, pp. 1433–1452.

[22] J. Tropp and S. Wright, "Computational methods for sparse solution of linear inverse problems," *Proc. IEEE*, vol. 98, no. 6, pp. 948–958, Jun. 2010.

[23] D. Donoho and J. Tanner, "Precise undersampling theorems," *Proc. IEEE*, vol. 98, no. 6, pp. 913–924, Jun. 2010.

[24] M. McKay, R. Beckman, and W. Conover, "A comparison of three methods for selecting values of input variables in the analysis of output from computer code," *Technometrics*, vol. 42, no. 1, pp. 55–61, 1979.

[25] A. Oppenheim, *Signals and Systems*. Upper Saddle River, NJ: Prentice-Hall, 1996.

[26] G. Golub and C. Loan, *Matrix Computations*. Baltimore, MD: Johns Hopkins University Press, 1996.

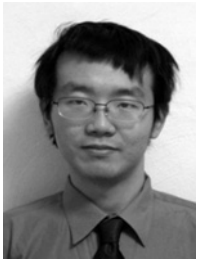
[27] M. Bushnell and V. Agrawal, *Essentials of Electronic Testing for Digital, Memory and Mixed-Signal VLSI Circuits*. Norwell, MA: Kluwer Academic, 2000.

[28] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge University Press, 2004.

[29] C. Bishop, *Pattern Recognition and Machine Learning*. Upper Saddle River, NJ: Prentice-Hall, 2007.

[30] R. Gonzalez and R. Woods, *Digital Image Processing*. Upper Saddle River, NJ: Prentice-Hall, 2007.

[31] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes: The Art of Scientific Computing*. Cambridge, U.K.: Cambridge University Press, 2007.



Wangyang Zhang (S'10) received the B.S. and M.S. degrees in computer science from Tsinghua University, Beijing, China, in 2006 and 2008, respectively. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA.

His current research interests include statistical methods for computer-aided design under process variations.

Mr. Zhang received the Best Paper Award from the Design Automation Conference in 2010.

Xin Li (S'01–M'06–SM'10) received the Ph.D. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, in 2005.

He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, Carnegie Mellon University. His current research interests include computer-aided design, neural signal processing, and power system analysis and design.

Dr. Li received the Best Paper Award from the Design Automation Conference in 2010 and the IEEE/ACM William J. McCalla ICCAD Best Paper Award in 2004.

Frank Liu (S'95–M'99–SM'09) received the Ph.D. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA.

Currently, he is a Research Staff Member with the IBM Research Laboratory, Austin, TX. He has authored and co-authored over 50 conference and journal papers. His current research interests include circuit analysis, model order reduction, numerical analysis, variability characterization, and computational lithography.

Dr. Liu is the co-recipient of the Best Paper Award at the Asia and South Pacific Design Automation Conference.

Emrah Acar (S'95–M'01–SM'07) received the Ph.D. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, in 2001.

He is currently with the IBM T. J. Watson Research Center, Yorktown Heights, NY. His current research interests include circuit simulation, timing analysis, low-power design methods, statistical analysis, and design for manufacturability tools.

Rob A. Rutenbar (S'77–M'84–SM'90–F'98) received the Ph.D. degree from the University of Michigan, Ann Arbor, in 1984.

He is currently the Abel Bliss Professor of Engineering and the Head of the Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana. His research pioneered much of today's commercial analog circuit and layout synthesis technology.

Dr. Rutenbar has received many awards, including several Best Paper Awards (e.g., DAC 1987, 2002, and 2010), the Semiconductor Research Corporation Aristotle Award for Excellence in Education in 2001, and the IEEE Circuits and Systems Industrial Pioneer Award in 2007. He is a fellow of the ACM.

Ronald D. Blanton (S'93–M'95–SM'03–F'09) received the Ph.D. degree in computer science and engineering from the University of Michigan, Ann Arbor, in 1995.

He is currently a Professor with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA. His current research interests include the test and diagnosis of integrated, heterogeneous systems and design, manufacture, and test information extraction from tester measurement data.