Virtual Probe: A Statistically Optimal Framework for Minimum-Cost Silicon Characterization of Nanoscale Integrated Circuits

Xin Li, Rob R. Rutenbar and Ronald D. Blanton
ECE Department, Carnegie Mellon University
5000 Forbs Avenue, Pittsburgh, PA 15213
{xinli, rutenbar, blanton}@ece.cmu.edu

ABSTRACT
In this paper, we propose a new technique, referred to as *virtual probe* (VP), to efficiently measure, characterize and monitor both inter-die and spatially-correlated intra-die variations in nanoscale manufacturing process. VP exploits recent breakthroughs in *compressed sensing* [15]-[17] 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 structure in (spatial) frequency domain, VP achieves substantially lower sampling frequency than the well-known (spatial) 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 that by testing the delay of just 50 chips on a wafer, VP accurately predicts the delay of the other 219 chips on the same wafer. In this example, VP reduces the estimation error by up to 10× compared to other traditional methods.

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

General Terms
Algorithms

Keywords
Process Variation, Characterization, Integrated Circuit

1. INTRODUCTION
As integrated circuits (ICs) scale to finer feature size, it becomes increasingly difficult to control process variations for nanoscale technologies [1]-[2]. 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.

Towards this goal, various techniques have been proposed for statistical IC analysis and optimization, e.g., statistical timing analysis [3]-[6], post-silicon tuning [7]-[9], etc. All 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 heavily relies on the accuracy of the variation model (e.g., distribution, correlation, etc.) 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., ring oscillators) deployed in wafer scribe lines and/or within product chips [10]-[12]. The traditional silicon characterization suffers from three major issues:

- **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 [11].
- **Long testing time**: Physically measuring all test structures through a limited number of I/O ports consumes a large amount of testing time [12]. At nanoscale technologies, IC testing has contributed to a significant portion of the total manufacturing cost [19].
- **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 [12].

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 variation 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 [18]. Namely, if the variation contains no spatial frequency higher than f_{MAX}, the sampling frequency must be at least $2f_{MAX}$, i.e., test structures must be spaced at most $1/(2f_{MAX})$ apart.

The Nyquist sampling theorem generally assumes that all frequency components below the maximum frequency f_{MAX} may exist; this, however, is not true for our silicon characterization application. As will be demonstrated by the industrial measurement data in Section 4, spatial variation typically has 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 [22]. However, our silicon characterization problem is substantially different from image compression, as we do not want to fully sample spatial variation at Nyquist rate and then "compress" it. Instead, we want to avoid

ICCAD'09, November 2–5, 2009, San Jose, California, USA.

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* [15]-[17]) to develop a novel framework of *virtual probe* (VP) for low-cost silicon testing and characterization. Our goal is to accurately predict the spatial variation of a wafer/chip by measuring very few test structures at a set of selected locations. The proposed VP algorithm is derived from *maximum posterior estimation* (MAP) [21]. 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 prove that by exploring the sparse structure in (spatial) frequency domain, VP can fully reconstruct the spatial variation with a probability nearly equal to 1, even if the (spatial) sampling frequency is much lower than the Nyquist rate [15]-[17]. As will be demonstrated by the industrial examples in Section 4, VP reduces the estimation error by up to 10× compared to other traditional methods.

The remainder of this paper is organized as follows. In Section 2, we develop the mathematical formulation and algorithm for VP, and then discuss several possible applications of VP in Section 3. The efficacy of VP is demonstrated by a number of examples using industrial measurement data in Section 4. Finally, we conclude in Section 5.

2. VIRTUAL PROBE

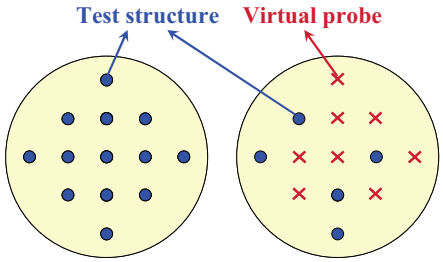

Fig 1. An example of the proposed virtual probes. (Left) Traditionally, a large number of test structures are deployed and measured to fully characterize process variations. (Right) We propose to deploy and measure very few test structures, and virtual probes are conceptually added to fully recover the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

The key idea of virtual probe (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 hardware testing. Instead, virtual probes are conceptually added at these locations to predict the variation information through the use of a numerical 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. This goal is facilitated by exploring the sparse structure in (spatial) frequency domain, as will be discussed in detail in this section.

In what follows, we first derive the mathematical formulation of VP by (spatial) frequency-domain analysis. Next, we propose to solve the VP problem by maximum posterior estimation (MAP) [21]. Finally, we convert the proposed MAP formulation to an equivalent linear programming problem that can be solved by the interior-point method [20] both robustly and efficiently.

2.1 Mathematical Formulation

Let $g(x, y)$ be the two-dimensional function of the performance of interest, where x and y represent the coordinate of a location within the two-dimensional plane. Depending on the test structure design, the performance g can be the frequency of a ring oscillator, the threshold of a transistor, etc. If $g(x, y)$ contains no spatial frequency higher than f_{MAX}, the Nyquist-Shannon sampling theorem [18] tells us to sample $g(x, y)$ with the sampling frequency of $2f_{MAX}$ in order to perfectly recover the continuous function $g(x, y)$.

Mathematically, the relation between the sampling values and their frequency-domain components can be represented by a number of two-dimensional linear transforms such as Fourier transform [18], discrete cosine transform (DCT) [22], wavelet transform [22], etc. 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 transforms.

We discretize the two-dimensional 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 $x \in \{1,2,...,P\}$ and $y \in \{1,2,...,Q\}$ after discretization. The DCT transform can be represented as [22]:

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

where $\{G(u, v); u = 1,2,...,P, v = 1,2,...,Q\}$ is a set of DCT coefficients and:

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

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

Equivalently, the sampling value $\{g(x, y); x = 1,2,...,P, y = 1,2,...,Q\}$ can be represented as the linear combination of $\{G(u, v); u = 1,2,...,P, v = 1,2,...,Q\}$ by the inverse discrete cosine transform (IDCT) [22]:

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

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

The proposed VP method, however, will go one step further. Our objective is to accurately recover $\{g(x, y); x = 1,2,...,P, y = 1,2,...,Q\}$ from a *very small* number of (say, M) samples at the locations $\{(x_m, y_m); m = 1,2,...,M\}$ where $M << PQ$. Towards this goal, we formulate the following linear equation:

$$A\cdot\eta = B \quad (5)$$

where

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

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

$$\eta = [G(1,1) \quad G(1,2) \quad \cdots \quad G(P,Q)]^T \quad (8)$$

$$B = [g(x_1, y_1) \quad g(x_2, y_2) \quad \cdots \quad g(x_M, y_M)]^T . \quad (9)$$

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

Solving the linear equation $A \cdot \eta = B$ in (5), however, is not trivial, since M (the number of equations) is vastly less than PQ (the number of unknowns). Namely, the linear equation in (5) is profoundly underdetermined. While Eq. (5) cannot be uniquely solved by a simple matrix inverse, we will show in the next sub-section that the solution of (5) can be statistically determined by considering additional prior information via Bayesian inference [21].

2.2 Maximum Posterior Estimation

In this sub-section, we describe an efficient algorithm using maximum posterior estimation (MAP) to statistically solve the linear equation (5). Although the result of this sub-section can be derived by applying a number of elegant statistics theorems [15]-[17], 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 [15]-[17] and [21].

To solve (5), we first need to define a so-called *prior distribution* for η [21]. 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 IC from 17 wafers, each containing 269 chips. Using this data set, we calculate the DCT coefficients. Fig 2 plots the histogram of the normalized DCT coefficient η_i (i.e., the i-th element of η). Studying Fig 2, we notice that the distribution has a sharp peak at $\eta_i = 0$. This implies that most DCT coefficients are close to 0. In general, if the performance variation $\{g(x, y); x = 1,2,...,P, y = 1,2,...,Q\}$ presents a spatial pattern, i.e., the variation is spatially correlated, the vector η that contains the corresponding DCT coefficients $\{G(u, v); u = 1,2,...,P, v = 1,2,...,Q\}$ is sparse. This sparsity has been observed in many image processing tasks [22], and motivates the compressed sensing work for image recovery using a minimum number of samples [16]-[17]. Roughly speaking, previous work in compressed sensing shows that if most of these coefficients are expected to be 0, 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 4, this sparseness assumption is also valid for silicon characterization.

To mathematically model the histogram in Fig 2, we use the following zero-mean Laplace distribution [21] to approximate the probability density function (PDF) of $\{\eta_i; i = 1,2,...,PQ\}$:

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

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 (MLE) [21]. 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 at the end of this sub-section (see (15)), the final MAP solution is independent of the actual value of λ.

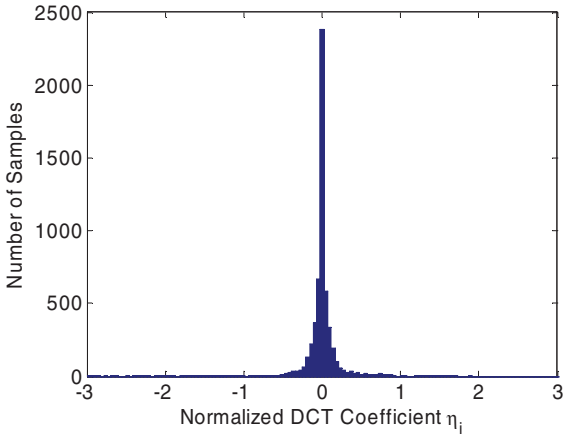

Fig 2. Histogram of the normalized DCT coefficient η_i for an industrial IC design example.

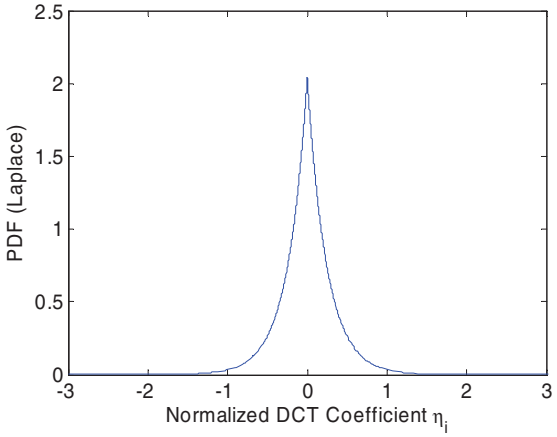

Fig 3. Optimally-fitted Laplace distribution for the normalized DCT coefficient η_i of an industrial IC design example.

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:

$$pdf(\eta) = \left(\frac{1}{2\lambda}\right)^{PQ} \cdot \prod_{i=1}^{PQ} \exp\left(-\frac{|\eta_i|}{\lambda}\right) = \left(\frac{1}{2\lambda}\right)^{PQ} \cdot \exp\left(-\frac{\|\eta\|_1}{\lambda}\right) \quad (11)$$

where $\|\eta\|_1$ denotes the L_1-norm, i.e., the summation of the absolute value of all elements in η. The prior PDF in (11) has a three-fold meaning. First, the DCT coefficients $\{\eta_i;\ i = 1,2,...,PQ\}$ have a high probability to equal zero. This, in turn, implies the sparseness of η. Second, 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 $\{g(x_m, y_m);\ m = 1,2,...,M\}$. Third, 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 (12)), once the measurement data are available. Next, we will describe the MAP algorithm to uniquely determine η based on the prior distribution in (11) as well as the measurement data $A \cdot \eta = B$ in (5).

The key idea of MAP is to find the *optimal* solution η that maximizes the *posterior distribution*, i.e., the conditional PDF $pdf(\eta \mid A \cdot \eta = B)$. Namely, given the measurement data $A \cdot \eta = B$, it aims to find the solution η that is *most likely* to occur. Based on Bayes' theorem [21], the posterior distribution $pdf(\eta \mid A \cdot \eta = B)$ is proportional to the prior distribution $pdf(\eta)$ and the likelihood function $pdf(A \cdot \eta = B \mid \eta)$:

$$pdf(\eta \mid A \cdot \eta = B) \propto pdf(\eta) \cdot pdf(A \cdot \eta = B \mid \eta). \quad (12)$$

In our case, the likelihood function is a Dirac delta function:

$$pdf(A \cdot \eta = B \mid \eta) = \begin{cases} \infty & (A \cdot \eta = B) \\ 0 & (A \cdot \eta \neq B) \end{cases}. \quad (13)$$

Hence, maximizing the posterior probability in (12) is equivalent to maximizing the prior probability in (11) subject to the constraint $A \cdot \eta = B$:

$$\begin{aligned} \underset{\eta}{\text{maximize}} \quad & 1/(2\lambda)^{PQ} \cdot \exp\left(-\|\eta\|_1/\lambda\right) \\ \text{subject to} \quad & A \cdot \eta = B \end{aligned}. \quad (14)$$

Since the exponential function $\exp(-\|\eta\|_1/\lambda)$ where $\lambda > 0$ monotonically decreases in $\|\eta\|_1$, the optimization in (14) can be re-written as:

$$\begin{aligned} \underset{\eta}{\text{minimize}} \quad & \|\eta\|_1 \\ \text{subject to} \quad & A \cdot \eta = B \end{aligned}. \quad (15)$$

Note that the optimization formulation in (15) is independent of the parameter λ.

Eq. (15) is referred to as *L_1-norm regularization* [15]-[17]. Several theoretical studies from the statistics community prove that with some general assumptions, the L_1-norm regularization in (15) yields the actual value of η. Roughly speaking, if the PQ-dimensional vector η contains K non-zeros and the linear equation $A \cdot \eta = B$ is well-conditioned, the solution η can be *almost* uniquely determined (with a probability nearly equal to 1) from M sampling points, where M is in the order of $O(K \cdot \log(PQ))$ [15]-[17]. Note that M (the number of sampling points) is a logarithmic function of PQ (the number of problem unknowns). It, in turn, provides the theoretical foundation we need: by solving the sparse solution η using MAP, all DCT coefficients $\{G(u, v);\ u = 1,2,...,P,\ v = 1,2,...,Q\}$ can be *almost* uniquely determined (with a probability nearly equal to 1) and, hence, the performance variation $g(x, y)$ can be *almost* completely recovered from a small number of sampling points.

2.3 Linear Programming

Studying (15), we notice that the cost function $\|\eta\|_1$ is not smooth. To efficiently solve the optimization, we convert (15) to an equivalent linear programming problem, as in [15].

Introduce a set of slack variables $\{\theta_i;\ i = 1,2,...,PQ\}$ and re-write (15) as the following form:

$$\begin{aligned} \underset{\eta,\theta}{\text{minimize}} \quad & \theta_1 + \theta_2 + \cdots + \theta_{PQ} \\ \text{subject to} \quad & A \cdot \eta = B \\ & -\theta_i \leq \eta_i \leq \theta_i \qquad (i = 1,2,\cdots,PQ) \end{aligned}. \quad (16)$$

Intuitively, by minimizing the cost function in (16), all constraints $\{-\theta_i \leq \eta_i \leq \theta_i;\ i = 1,2,...,PQ\}$ will become active, i.e., $\{|\eta_i| = \theta_i;\ i = 1,2,...,PQ\}$. For this reason, the optimizations in (15) and (16) are equivalent. This conclusion can be formally proven by using the Karush-Kuhn-Tucker condition from optimization theory [20]. Note that both the cost function and the constraints in (16) are linear. Therefore, it is a linear programming problem and can be solved by various efficient and robust algorithms, e.g., the interior-point method [20].

2.4 Summary

Algorithm 1: Virtual Probe (VP)

1. Randomly select M sampling locations $\{(x_m, y_m);\ m = 1,2,...,M\}$.
2. Collect the measurement data $\{g(x_m, y_m);\ m = 1,2,...,M\}$ at these locations.
3. Formulate the linear equation in (5)-(9).
4. Solve the linear programming problem in (16) to determine η, i.e., the DCT coefficients $\{G(u, v);\ u = 1,2,...,P,\ v = 1,2,...,Q\}$.
5. Apply IDCT in (4) to recover the performance function $\{g(x, y);\ x = 1,2,...,P,\ y = 1,2,...,Q\}$ spatially across the wafer/chip.

Algorithm 1 summarizes the major steps of the proposed VP method. It starts from very few (i.e., M) random sampling points $\{g(x_m, y_m);\ m = 1,2,...,M\}$. Given these measurement data, VP solves a linear programming problem to determine all DCT coefficients $\{G(u, v);\ u = 1,2,...,P,\ v = 1,2,...,Q\}$ and, consequently, recover the spatial variation $\{g(x, y);\ x = 1,2,...,P,\ y = 1,2,...,Q\}$. It is worth mentioning that the random sampling scheme used by Algorithm 1 may not result in maximum accuracy. In our future research, we will further study this sampling issue and develop improved algorithm for optimal sampling.

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

- **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/characterization time, yield loss during testing, etc. In addition, the VP formulation in (16) is a linear programming problem and it can be solved both robustly and efficiently. Hence, the computation overhead of Algorithm 1 is almost negligible, as will be demonstrated by our examples based on industrial measurement data in Section 4.

- **High accuracy**: The prediction accuracy of VP is guaranteed by the theoretical studies from the statistics community [15]-[17]. Namely, with some general assumptions, VP can fully

reconstruct the spatial variation with a probability nearly equal to 1. In addition, the accuracy of VP can be verified in real time using several efficient techniques [15]-[17], e.g., cross validation, Bayesian inference, etc. These error estimation schemes are extremely important, since they provide quantitative criteria to determine whether the result of VP is sufficiently accurate or not. In practice, additional sampling points can be further collected to improve accuracy, until the prediction error is sufficiently small. More details on error estimation can be found in [15]-[17]. They are not discussed in this paper due to the limited number of available pages.

- **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 variation has a sparse representation in frequency domain. This assumption is typically valid, as process variations are spatially correlated. In practice, such a sparseness assumption can be easily verified by the error estimation schemes we previously mentioned. Namely, if the frequency-domain representation is not sparse, we will observe large prediction error reported by VP.

3. APPLICATION OF VIRTUAL PROBE

The proposed VP method can be applied to a broad range of applications related to integrated circuits. In this section, we first illustrate how to apply VP to silicon characterization at both wafer level (for inter-die variations) and chip level (for intra-die variations). Next, we briefly discuss several additional application areas including speed binning and post-silicon tuning.

3.1 Wafer-Level Silicon Characterization

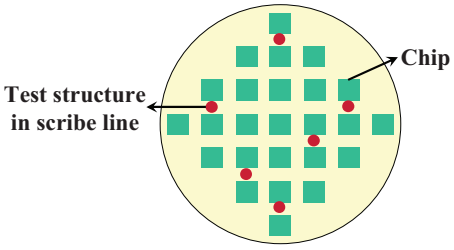

Fig 4. Test structures are deployed in wafer scribe lines to measure and characterize inter-die variations at wafer level.

To characterize parametric variations at wafer level (i.e., inter-die variations), test structures are deployed in wafer scribe lines [10]-[12], as shown in Fig 4. 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 and/or voltages. 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. It, 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 1 is applied to reconstruct the spatial variation 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 detail in the next sub-section.

3.2 Chip-Level Silicon Characterization

On-chip test structures are typically used to characterize intra-die variations at chip level [10]-[12], as shown in Fig 5. 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 of the chip. This testing process is time-consuming and directly increases manufacturing cost.

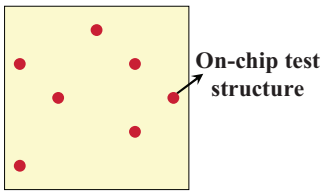

Fig 5. Test structures are deployed within a product chip to measure and characterize intra-die variations at chip level.

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 information within a chip. As such, the characterization cost is substantially reduced.

3.3 Beyond Silicon Characterization

The silicon characterization results extracted by VP can be efficiently applied to a number of practical applications. In this sub-section, 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 [19]. 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 [7]-[9]. It aims to adaptively configure a number of tunable parameters (e.g., supply voltage, body bias, etc.) so that a given circuit can work properly under different process conditions. An important component of post-silicon tuning is accurate on-chip measurement of the current process condition so that the tunable parameters can be appropriately configured to adjust the circuit behavior. Such measurement, however, is not trivial, as 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 much reduced number of on-chip sensors. By minimizing the number of the required sensors, both the design complexity and the manufacturing cost can be significantly reduced.

4. NUMERICAL EXAMPLES

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

4.1 Flush Delay Measurement Data

We consider the flush-delay values measured from 269 industrial chips on the same wafer, as shown in Fig 6. In this example, the measured delay is not a constant, but significantly varies from chip to chip due to process variations. Our goal is to capture these wafer-level delay variations. We use a two-dimensional function $g(x, y)$ to model the delay, where $x \in \{1,2,...,18\}$ and $y \in \{1,2,...,19\}$. Each coordinate point (x, y) corresponds to a chip. Next, we apply a two-dimensional DCT to $g(x, y)$, yielding the frequency-domain components $G(u, v)$ shown in Fig 7.

Two important observations can be made from the result in Fig 7. 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 structure is the essential necessary condition that makes the proposed VP technique applicable to this example.

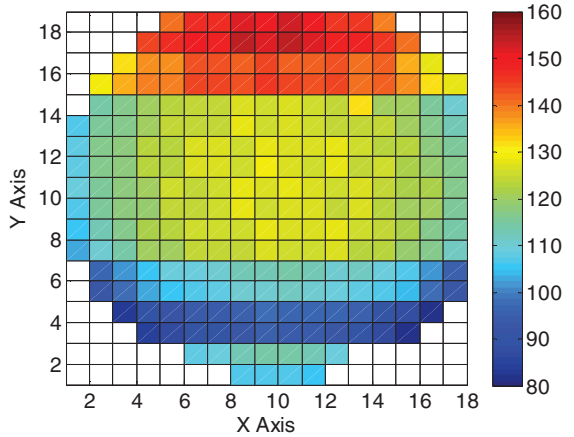

Fig 6. Measured flush-delay values (normalized by a randomly selected constant) of 269 industrial chips from the same wafer.

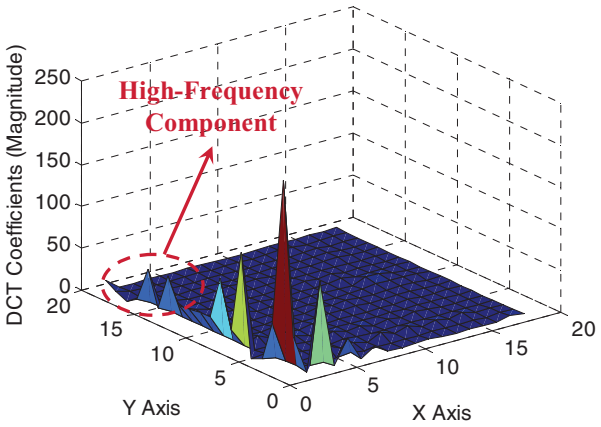

Fig 7. Discrete cosine transform (DCT) of the normalized flush-delay measurement.

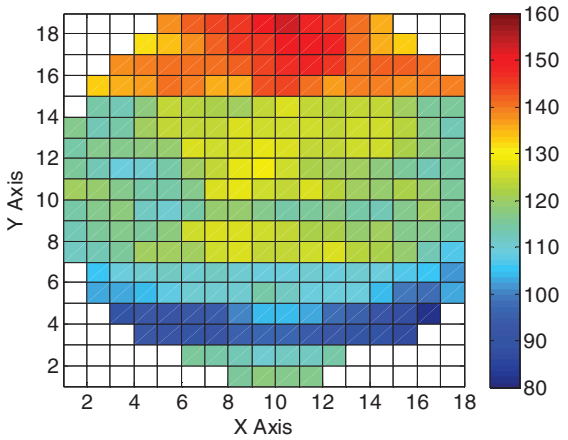

Fig 8. Recovered flush-delay values from 50 tested chips by using VP.

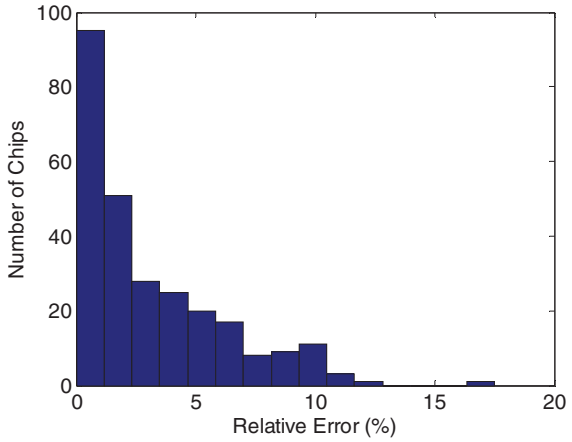

Fig 9. Histogram of the relative error calculated by Eq. (17) for all chips on the same wafer.

We apply Algorithm 1 to recover $g(x, y)$ based on a small number of (i.e., M) sampling points $\{g(x_m, y_m); m = 1,2,...,M\}$. The linear optimization in (16) is solved by the commercial optimization software MOSEK (www.mosek.com). Fig 8 shows the recovered flush-delay values from 50 tested chips (i.e., M = 50). In this case, the total runtime of Algorithm 1 is less than 1

second.

To assess the accuracy of VP, we calculate the following relative error for each chip:

$$Error_{REL}(x,y)=\left|\frac{g(x,y)-\tilde{g}(x,y)}{g(x,y)}\right| \quad (17)$$

where $g(x, y)$ and $\tilde{g}(x, y)$ denote the exact value and the estimated value of the performance function, respectively. The error metric defined in (17) quantitatively measures the difference between the measurement data (i.e., Fig 6) and the prediction result (i.e., Fig 8). Fig 9 shows the histogram of the relative error calculated for all chips on the wafer. Note that the relative error of VP is less than 10% for most chips in this example.

For testing and comparison, we also study this example from Nyquist point of view. As shown in Fig 7, if we want to completely recover $g(x, y)$, we cannot decrease the spatial sampling frequency at all; otherwise, we will not be able to capture the high-frequency components due to aliasing. To quantitatively study the impact of down-sampling, we sample $g(x, y)$ by a uniform two-dimensional grid and recover $g(x, y)$ using the traditional two-dimensional interpolation [23]. The average error is then calculated as:

$$Error_{AVG}=\sqrt{\frac{\sum_{x=1}^{18}\sum_{y=1}^{19}[g(x,y)-\tilde{g}(x,y)]^2}{\sum_{x=1}^{18}\sum_{y=1}^{19}[g(x,y)]^2}}. \quad (18)$$

Fig 10 shows the average error as a function of the number of samples (i.e., M) for both VP and the two-dimensional interpolation method. Note that VP achieves up to 10× error reduction in this example. The error of VP is around 6%, when 50 chips (out of 269 chips in total) are tested. To achieve the same accuracy, the traditional two-dimensional interpolation has to measure 225 chips (4× more).

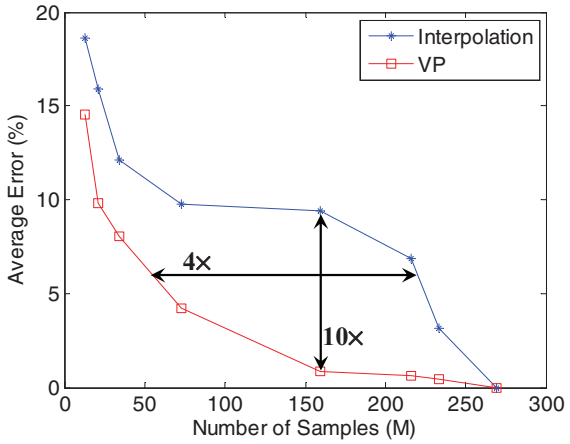

Fig 10. Average error decreases as the number of samples (i.e., M) increases.

4.2 Leakage Current Measurement Data

We consider the leakage-current measurement collected by IDDQ test for the same industrial circuit design. Fig 11 shows the normalized leakage-current values $\log_{10}(I_{LEAK})$ (after logarithmic transform) as a function of the location (x, y). Fig 12 further shows the frequency-domain components after DCT. Similar to the previous example, the DCT coefficients contain important high-frequency components; yet they are quite sparse.

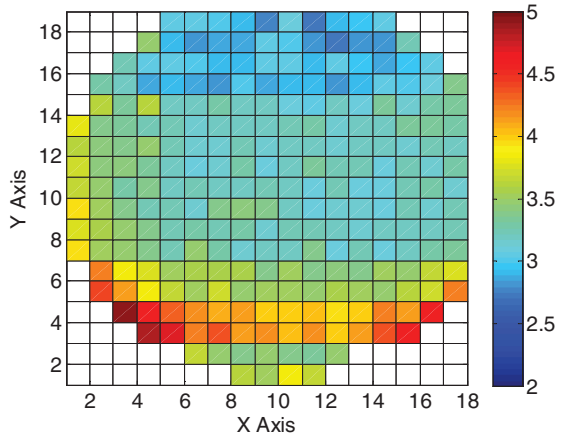

Fig 11. Measured leakage-current values $\log_{10}(I_{LEAK})$ (normalized by a randomly selected constant) of 269 industrial chips from the same wafer.

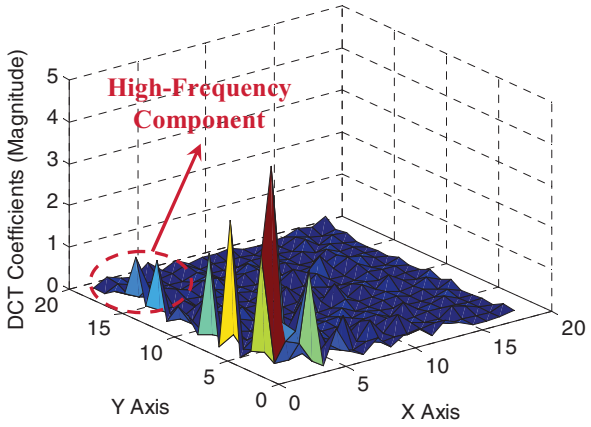

Fig 12. Discrete cosine transform (DCT) of the normalized leakage-current measurement $\log_{10}(I_{LEAK})$.

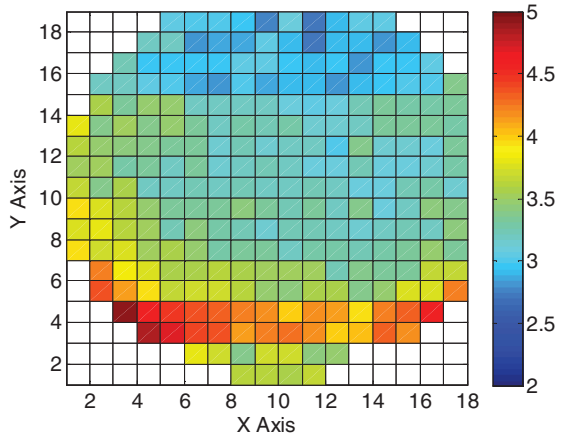

Fig 13. Recovered leakage-current values $\log_{10}(I_{LEAK})$ from 100 tested chips by using VP.

Next, we apply Algorithm 1 to recover the spatial variation based on a few (i.e., M) sampling points. Fig 13 shows the recovered leakage-current values $\log_{10}(I_{LEAK})$ (after logarithmic

transform) from 100 tested chips (i.e., $M = 100$). In this case, the total runtime of Algorithm 1 is less than 1 second. Fig 9 further shows the histogram of the relative error calculated for all chips using (17). Note that the relative error of VP is less than 10% for most chips in this example.

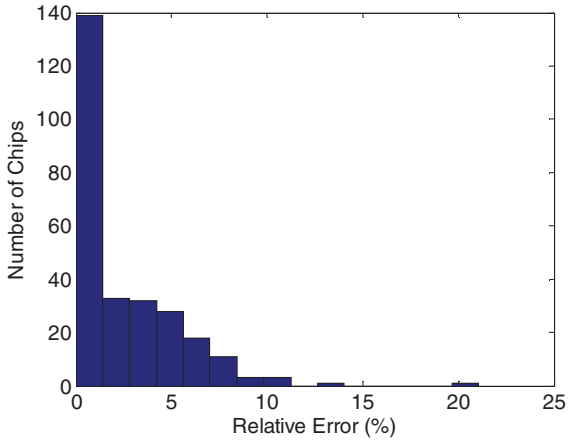

Fig 14. Histogram of the relative error calculated by Eq. (17) for all chips on the same wafer.

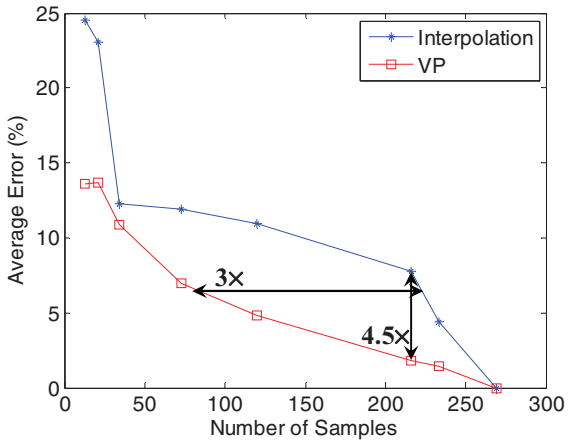

Fig 15. Average error decreases as the number of samples (i.e., M) increases.

For testing and comparison, we sample $g(x, y)$ by a uniform two-dimensional grid and recover $g(x, y)$ using the traditional two-dimensional interpolation [23]. Fig 15 shows the average error calculated by (18) for both methods. Note that VP achieves up to 4.5× error reduction compared to the interpolation approach. The error of VP is around 6%, when 75 chips (out of 269 chips in total) are tested. To achieve the same accuracy, the traditional two-dimensional interpolation has to measure 225 chips (3× more).

5. CONCLUSIONS

In this paper, we propose a novel *virtual probe* (VP) technique to efficiently and accurately recover full-wafer/chip spatial variation from an extremely small set of measurement data, thereby reducing the cost of silicon characterization and testing. VP exploits recent breakthroughs in compressed sensing [15]-[17]. It is formulated as a maximum posterior estimation (MAP) problem [21] and is solved via efficient linear programming algorithm [20]. Our numerical examples based on industrial measurement data demonstrate that VP reduces the estimation error by up to 10× compared to other traditional techniques.

6. ACKNOWLEDGEMENTS

This work was supported in part by the National Science Foundation under contract CCF–0915912 and by the Center for Circuit & System Solutions (C2S2).

7. REFERENCES

[1] S. Nassif, "Delay variability: sources, impacts and trends," *IEEE ISSCC*, pp. 368-369, 2000.

[2] Semiconductor Industry Associate, *International Technology Roadmap for Semiconductors*, 2007.

[3] H. Chang and S. Sapatnekar, "Statistical timing analysis under spatial correlations," *IEEE Trans. CAD*, vol. 24, no. 9, pp. 1467-1482, 2005.

[4] C. Visweswariah, K. Ravindran, K. Kalafala, S. Walker and S. Narayan, "First-order incremental block-based statistical timing analysis," *IEEE DAC*, pp. 331-336, 2004

[5] Y. Zhan, A. Strojwas, X. Li, L. Pileggi, D. Newmark and M. Sharma, "Correlation aware statistical timing analysis with non-Gaussian delay distributions," *IEEE DAC*, pp. 77-82, 2005.

[6] K. Heloue and F. Najm, "Statistical timing analysis with two-sided constraints," *IEEE ICCAD*, pp. 829-836, 2005.

[7] M. Mani, A. Singh, and M. Orshansky, "Joint design-time and post-silicon minimization of parametric yield loss using adjustable robust optimization," *IEEE ICCAD*, pp. 19-26, 2006.

[8] S. Kulkarni, D. Sylvester and D. Blaauw, "A statistical framework for post-silicon tuning through body bias clustering," *IEEE ICCAD*, pp. 39-46, 2006.

[9] Q. Liu and S. Sapatnekar, "Synthesizing a representative critical path for post-silicon delay prediction," *ACM ISPD*, pp. 183-190, 2009.

[10] M. Ketchen, M. Bhushan and D. Pearson, "High speed test structures for in-line process monitoring and model calibration," *IEEE ICMTS*, pp. 33-38, 2005.

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

[12] W. Mann, F. Taber, P. Seitzer and J. Broz, "The leading edge of production wafer probe test technology," *IEEE ITC*, pp. 1168-1195, 2004.

[13] F. Koushanfar, P. Boufounos and D. Shamsi, "Post-silicon timing characterization by compressed sensing," *IEEE ICCAD*, pp. 185-189, 2008.

[14] S. Reda and S. Nassif, "Analyzing the impact of process variations on parametric measurements: novel models and applications," *IEEE DATE*, pp. pp. 375-380, 2009.

[15] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of Royal Statistical Society*, vol. 58, no. 1, pp. 267-288, 1996.

[16] E. Candes, "Compressive sampling," *International Congress of Mathematicians*, 2006.

[17] D. Donoho, "Compressed sensing," *IEEE Trans. Information Theory*, vol. 52, no. 4, pp. 1289-1306, Apr. 2006.

[18] A. Oppenheim, *Signals and Systems*, Prentice Hall, 1996.

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

[20] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[21] C. Bishop, *Pattern Recognition and Machine Learning*, Prentice Hall, 2007.

[22] R. Gonzalez and R. Woods, *Digital Image Processing*, Prentice Hall, 2007.

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