A Framework for Scalable Post-Silicon Statistical Delay Prediction under Spatial Variations

Qunzeng Liu and Sachin S. Sapatnekar,

***Abstract*— Due to increased variability trends in nanoscale integrated circuits, statistical circuit analysis and optimization has become essential. While statistical timing analysis has an important role to play in this process, it is equally important to develop die-specific delay prediction techniques using post-silicon measurements. We present a novel method for post-silicon delay analysis. We gather data from a small number of on-chip test structures, and combine this information with presilicon statistical timing analysis to obtain narrow, die-specific, timing probability density function (PDF). Experimental results show that for the benchmark suite being considered, taking all parameter variations into consideration, our approach can obtain a PDF whose standard deviation is 79.0% smaller, on average, than the statistical timing analysis result. The accuracy of the method defined by our metric is 99.6% compared to Monte-Carlo simulation. The approach is scalable to smaller test structure overheads and can still produce acceptable results.**

I. Introduction

Feature sizes in VLSI design have been shrinking for several decades, and are currently in the tens of nanometers. In this regime, process variations play a critical role in determining circuit performance, and it is widely accepted that they must be taken into consideration during the design process in order to ensure that a manufactured circuit meets its specifications. Generally speaking, process variations can be classified as inter-die variations and intra-die variations. Inter-die variations are fluctuations in process parameters from chip to chip, while intra-die variations are the variations among different elements within a single die. Some, but not all, intra-die variations may show the property of spatial correlation, which implies that the process parameters associated with transistors or wires that are close to each other are more likely to vary in a similar way than those of transistors or wires that are far away from each other. The variation in the effective channel length L and transistor width W are observed to show a spatial correlation structure, while the dopant concentration N_A and the oxide thickness T_{ox} are generally considered not to be spatially correlated.

These variations pose great challenges to analyzing the timing behavior of a circuit, as traditional corner-based static timing analysis (STA) may be overly pessimistic [1]. To overcome this problem, statistical static timing analysis (SSTA) has been proposed as an alternative that replaces the deterministic delay values from STA with probability density functions (PDFs) that capture the mean as well as the spread of the delay in manufactured parts. Many techniques aimed at developing accurate and efficient SSTA algorithms, such as [2]–[8], have been developed, among which parameterized block-based SSTA methods [2]–[6] have distinguished themselves by easily taking into consideration the spatial and structural correlations of the parameter variations in the circuit to be analyzed. The computational efficiency of these methods is made practical through a preprocessing step, proposed in [2], [3], which has shown that Gaussian-distributed correlated variations can be orthogonalized using principal component analysis (PCA). Much of the work in this area assumes that all parameter variations are Gaussian and that a linear delay model obtained by Taylor expansion for each circuit component is sufficient to capture the impact of the variations. More recently, non-Gaussian parameter variations as well as nonlinear delay models have been addressed in, for example, [5], [6], [9].

With the aid of SSTA tools, designers can optimize a circuit before it is fabricated, in the expectation that it will meet the delay and power requirements after being manufactured. In other words, SSTA is a presilicon analysis technique used to determine the range of performance (delay or power) variations over a large population of dies. A complementary role, after the chip is manufactured, is played by post-silicon diagnosis, which is typically directed toward determining the performance of an individual fabricated chip based on measurements on that specific chip. This procedure provides particular information that can be used to perform post-silicon optimizations to make a fabricated part meet its specifications. Because presilicon analysis has to be generally applicable to the entire population of manufactured chips, the statistical analysis that it provides shows a relatively large standard deviation for the delay. On the other hand, post-silicon procedures, which are tailored to individual chips, can be expected to provide more specific information. Since tester time is generally prohibitively expensive, it is necessary to derive the maximum possible information through the fewest post-silicon measurements.

In the past, the interaction between presilicon analysis and post-silicon measurements has been addressed in several ways. In [10], post-silicon measurements are used to learn a more accurate spatial correlation model, which is fed back to the analysis stage to refine the statistical timing analysis framework. In [11], a path-based methodology is used for correlating post-silicon test data to presilicon timing analysis. In [12], a statistical gate sizing approach is studied to optimize the binning yield. Post-silicon debug methods and their interaction with circuit design are discussed in [13].

The authors are with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455. This research was supported in part by the NSF under award under award CCF-0205227 and by the SRC under contract 2007-TJ-1572.

The method that we present in this paper differs from these in terms of its goals. Our approach forms a framework for post-silicon statistical delay prediction: the role of this step is seated between presilicon SSTA and post-silicon full chip testing. We combine the results of presilicon SSTA for the circuit with the result of a small number of post-silicon measurements on an individual manufactured die to estimate the delay of that particular die.

Given the *original circuit* whose delay is to be estimated, the primary idea is to determine information from specific on-chip *test structures* to narrow the range of the performance distribution substantially; for purposes of illustration, we will consider delay to be the performance metric in this work. In particular, we gather information from a small set of test structures such as ring oscillators (ROs), distributed over the area of the chip, to capture the variations of spatially correlated parameters over the die. The physical sizes of the test structures are small enough that it is safe to assume that they can be incorporated into the circuit using reserved space that may be left for buffer insertion, decap insertion, etc. without significantly perturbing the layout. To illustrate the idea, we show a die in Figure 1, whose area is gridded into spatial correlation regions[1]. Figure 1(a) and 1(b) show two cases where test structures are inserted on the die: the two differ only in the number and the locations of these test structures. Figure 2 shows a sample test structure consisting of a 3-stage RO; however, in practice, the number of stages in this structure may be larger, and these trade-offs are explored in Section VI. The data gathered from the test structures in Figures 1(a) and 1(b) are used in this paper to determine a new PDF for the delay of the original circuit, conditioned on this data. This PDF has a significantly smaller variance than that obtained from SSTA, as is illustrated in Figure 3; detailed experimental results are available in Section VII.

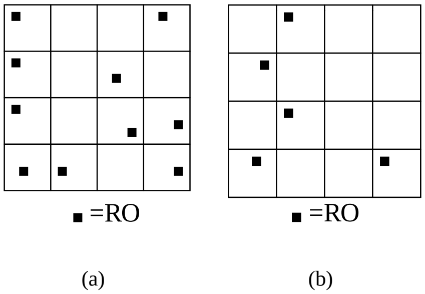

Fig. 1. Two different placements of test structures under the grid spatial correlation model.

The plots in Figure 3 may be interpreted as follows. When no test structures are used and no post-silicon measurements are performed, the PDF of the original circuit is the same as that computed by SSTA. When 5 ROs are used, a tighter spread is seen for the PDF, and the mean shifts towards the actual frequency for the die. This spread becomes tighter still when 10 ROs are used. In other words, as the number of test structures is increased, more information can be derived about variations on the die, and its delay PDF can be predicted with greater confidence: the standard deviation of the PDF from SSTA is always an upper bound on the standard deviation of this new delay PDF. In other words, by using more or fewer test structures, the approach is *scalable* in terms of statistical confidence.

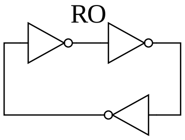

Fig. 2. An example of a test structure: A three-stage ring oscillator.

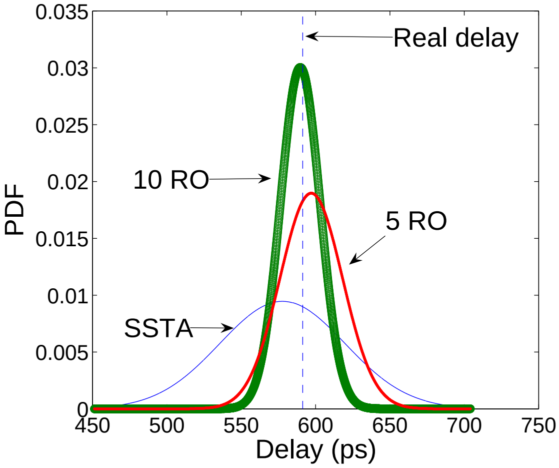

Fig. 3. Reduced-variance PDFs, obtained from statistical delay prediction, using data gathered from the test structures in Figure 1.

The focus of our approach is on post-silicon delay analysis, but we will outline a use case scenario for this analysis in the realm of post-silicon tuning. Adaptive Body Bias (ABB) [14]–[16] is a post-silicon method that determines the appropriate level of body bias to be applied to a die to influence its performance characteristics. ABB is typically a coarse-grained optimization, both in terms of the granularity at which it can be applied (typically on a per-well basis) as well as in terms of the granularity of the voltage levels that may be applied (typically, the separation between ABB levels is 50 to 100 mV). Current ABB techniques use a critical path replica to predict the delay of the fabricated chip, and use this to feed a phase detector and a counter, whose output is then used to generate the requisite body bias value. Such an approach assumes that one critical path on a chip is an adequate reflection of on-chip variations. In general, there will be multiple potential critical paths even within a single combinational block, and there will be a large number of combinational blocks in a within-die region. Choosing a single critical path as representative of all of these variations is impractical and inaccurate. In contrast, our approach implicitly considers the effects of all paths in a circuit (without enumerating them, of course), and provides a PDF that concretely takes spatially correlated and uncorrelated

[1] For simplicity, we will assume in this example that the spatial correlation regions for all parameters are the same, although the idea is valid, albeit with an uglier picture, if this is not the case.

parameters into account to narrow the variance of the sample, and has no preconceived notions, prior to fabrication, as to which path will be critical. The 3σ or 6σ point of this PDF may be used to determine the correct body bias value that compensates for process variations. Temperature variations may be compensated for separately using temperature sensors, for example, as in [17].

The remainder of this paper is organized as follows. Section II abstracts the physical problem as a mathematical abstraction. Next, Sections III through V introduce our approach in detail and outline its limitations. Section VI then discusses the impact of changing the number of stages in the RO test structures on the quality of the results. Experimental results are shown in Section VII, followed by concluding remarks in Section VIII.

II. Background and Problem Formulation

A. Spatial Correlations

Spatial correlations of parameter variations were considered as a challenge in SSTA until the arrival of parameterized methods. We use the grid-based spatial correlation model [2] in this paper. Under this model, we assume that variations of the same process parameter inside each grid are fully correlated, variations of the same process parameter between grids that are physically close to each other are more correlated than is the case for grids that are far away.

From the post-silicon analysis perspective, these spatial correlations may be exploited to generate information based on a limited number of test structures. More specifically, the parameter variations for the test structures in a chip are correlated with those of the gates near them. For the specific case where only inter-die variations are seen, and no intra-die variations exist (a special case of spatial correlations), the parameters of a test structure anywhere on a chip are identical to those of the original circuit to be tested. The presence of intra-die variations creates some challenges: the parameters of the test structure may now be correlated with, but not identical to, those in the original circuit. In such a case, the parameter variation of a test structure cannot reveal information for all the parameters of the original circuit, but can reveal some characteristics for the devices nearby. Therefore, our proposed post-silicon statistical delay prediction approach uses a number of test structures, placed at different locations on chip, to provide diverse test data. The presence of uncorrelated variations creates further challenges, which are examined in this paper.

B. Problem Formulation

We assume that the circuit undergoes SSTA prior to manufacturing, and that the random variable that represents the maximum delay of the original circuit is d. Further, if the number of test structures placed on the chip is n, we define a *delay vector* $\mathbf{d}_t = \begin{bmatrix} d_{t,1} & d_{t,2} & \cdots & d_{t,n} \end{bmatrix}^T$ for the test structures, where $d_{t,i}$ is the random variable (over all manufactured chips) corresponding to the delay of the i^{th} test structure.

For a particular fabricated die, the delay of the original circuit and the test structures correspond, respectively, to one sample of the underlying process parameters, which results in a specific value of d and of $\mathbf{d}_t$. After manufacturing, measurements are performed on the test structures to determine the sample of $\mathbf{d}_t$, which we call the *result vector* $\mathbf{d}_r = \begin{bmatrix} d_{r,1} & d_{r,2} & \cdots & d_{r,n} \end{bmatrix}^T$. This corresponds to a small set of measurements that can be performed rapidly. The objective of our work is to develop techniques that permit these measurements to be used to predict the corresponding sample of d on the same die. In other words, we define the problem of post-silicon statistical delay prediction as finding the conditional PDF given by $f(d|\mathbf{d}_t = \mathbf{d}_r)$.

In the ideal case, given enough test structures, we can estimate the delay of the original circuit with very little variance by measuring these test structures. However, practical constraints limit the overhead of the added test structures (such as area, power, and test time) so that the number of these structures cannot be arbitrarily large. Moreover, as stated in Section II-A, our method is made possible by spatial correlations of parameter variations at different locations. However, the variations in some parameters, such as T_{ox} and N_A, are widely believed to show no spatial correlation structure at all. Test structures are inherently not capable of capturing any such variations in the original circuit (beyond the overall statistics that are available to the SSTA engine): these parameters can vary from one device to the next, and thus, variations in the test circuit are totally independent of any variations in the original circuit, but even under these limitations, any method that can narrow down the variational range of the original circuit through a few test measurements is of immense practical use.

We develop a method that robustly accounts for the aforementioned limitations by providing a conditional PDF of the delay of the original circuit with insufficient number of test structures and/or purely random variations. In the case when the original circuit delay can actually be computed as a fixed value, the conditional PDF is an impulse function with mean equal to the delay of the original circuit and zero variance. The variance becomes larger with fewer test structures, and shows a graceful degradation in this regard. We include all of these in a single generalized framework and automatically take each case into consideration.

III. Statistical Delay Prediction

A. The SSTA Framework

SSTA provides a PDF of the delay distribution of the circuit, rather than predicting a specific delay value at a process corner, as is the case for STA methods. The parameterized approach to SSTA propagates a canonical form of the delay PDF, typically including the nominal value, a set of normalized underlying independent sources of variation (for spatially correlated variations, these should be the principal components (PCs) [2], computed by applying PCA to the underlying covariance matrix of the correlated variations; uncorrelated variations are typically captured by a single independent random variable).

In this work, we assume that the process parameters, which affect both the original circuit and test structures, are Gaussian-distributed. The m PCs affect the statistical distribution of both the original circuit and the test structures on the same chip, and the canonical form for the delay is represented as:

$$d = \mu + \sum_{i=1}^{m} a_i p_i + R = \mu + \mathbf{a}^T \mathbf{p} + R, \tag{1}$$

where d is defined in Section II-B, and μ is the mean of the delay distribution. The value of μ is also an approximation of its nominal value[2]. The random variable p_i corresponds to the ith principal component, and is normally distributed, with zero mean and unit variance; note that p_i and p_j for $i \neq j$ are uncorrelated by definition, stemming from a property of PCA. The parameter a_i is the first order coefficient of the delay with respect to p_i. Finally, R corresponds to a variable that captures the effects of all the spatially uncorrelated variations. It is a placeholder to indicate the additional variations of the delay caused by the spatially uncorrelated variations, and cannot be regarded as a principal component. For simplicity, we refer to $\mathbf{p} = \begin{bmatrix} p_1 & p_2 & \cdots & p_m \end{bmatrix}^T \in \mathbf{R}^m$ as the *PC vector* and $\mathbf{a} = \begin{bmatrix} a_1 & a_2 & \cdots & a_m \end{bmatrix}^T \in \mathbf{R}^m$ as the *coefficient vector* for the original circuit.

Equation (1) is general enough to incorporate both inter-die and intra-die variations. It is well known that for a spatially correlated parameter, the inter-die variation can be taken into account by adding a value σ^2_{inter}, the variance of inter-die parameter variation, to all entries of the covariance matrix of the intra-die variation of that parameter before performing PCA. The uncorrelated component R accounts for contributions from both the inter-die and intra-die variations. Systematic variations affect only the nominal values and the PC coefficients in SSTA. Therefore, they can be accounted for by determining the shifted nominal values and sensitivities prior to SSTA, and computing the nominal values and PC coefficients in SSTA based on these shifted values. While our theory is general enough to capture this, for simplicity, our experimental results do not consider this effect.

In a similar manner, the delay of the i^{th} of the n test structures can also be represented in the canonical form as:

$$d_{t,i} = \mu_{t,i} + \mathbf{a}^T_{t,i}\mathbf{p} + R_{t,i}. \tag{2}$$

The meanings of all variables are inherited from Equation (1).

We define $\boldsymbol{\mu}_t = \begin{bmatrix} \mu_{t,1} & \mu_{t,2} & \cdots & \mu_{t,n} \end{bmatrix}^T \in \mathbf{R}^n$ as the *mean vector*, $\mathbf{R}_t = \begin{bmatrix} R_{t,1} & R_{t,2} & \cdots & R_{t,n} \end{bmatrix}^T \in \mathbf{R}^n$ as the *independent parameter vector*, and $\mathbf{A}_t \in \mathbf{R}^{m \times n}$ as the *coefficient matrix* of the test structures, respectively, where $\mathbf{A}_t = \begin{bmatrix} \mathbf{a}_{t,1} & \mathbf{a}_{t,2} & \cdots & \mathbf{a}_{t,n} \end{bmatrix}$. We can then stack the delay equations of all of the test structures into a matrix form.

$$\mathbf{d}_t = \boldsymbol{\mu}_t + \mathbf{A}^T_t \mathbf{p} + \mathbf{R}_t \tag{3}$$

where $\mathbf{d}_t$ is defined in Section II.

To illustrate the procedure more clearly and in an easier way, we will first assume, in the remainder of this section and in Section IV, that the spatially uncorrelated parameters can be ignored, i.e., $R = 0$ and $\mathbf{R_t} = \mathbf{0}$. We will relax this assumption later in Section V, and introduce the extension of the method to include those parameters.

The variance of the Gaussian variable d and the covariance matrix of the multivariate normal variable $\mathbf{d}_t$ can be conveniently calculated as:

$$\sigma^2 = \mathbf{a}^T \mathbf{a} \tag{4a}$$

$$\boldsymbol{\Sigma}_t = \mathbf{A}^T_t \mathbf{A}_t. \tag{4b}$$

B. Conditional PDF Evaluation

The objective of our approach is to find the conditional PDF of the delay, d, of the original circuit, given the vector of delay values, $\mathbf{d}_r$. The values of $\mathbf{d}_r$ are measured from the test structures after the circuit is manufactured, corresponding to one set of samples of $\mathbf{d}_t$. We first introduce a theorem below; a sketch of the proof of the theorem can be found in [18].

Theorem 3.1: Consider a Gaussian-distributed vector $\begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$ with mean $\boldsymbol{\mu}$ and a nonsingular covariance matrix $\boldsymbol{\Sigma}$. Let us define $\mathbf{X}_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$, $\mathbf{X}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$. If $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are partitioned as follows,

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}, \tag{5}$$

then the distribution of $\mathbf{X}_1$ conditional on $\mathbf{X}_2 = \mathbf{x}$ is multivariate normal, and its mean and covariance matrix are given by

$$\mathbf{X}_1 | (\mathbf{X}_2 = \mathbf{x}) \sim N(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}}) \tag{6a}$$

$$\bar{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}^{-1}_{22}(\mathbf{x} - \boldsymbol{\mu}_2) \tag{6b}$$

$$\bar{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}^{-1}_{22}\boldsymbol{\Sigma}_{21}. \tag{6c}$$

It can be shown that our problem can be mapped directly to the theorem. To show this correspondence, we define $\mathbf{X}_1$ as the *original subspace*, and $\mathbf{X}_2$ as the *test subspace*. By stacking d and $\mathbf{d}_t$ together, a new vector $\mathbf{d}_{all} = \begin{bmatrix} d \; \mathbf{d}^T_t \end{bmatrix}^T$ is formed, with the original subspace containing only one variable d and the test subspace containing the vector $\mathbf{d}_t$. The random vector $\mathbf{d}_{all}$ is multivariate Gaussian-distributed, with its mean and covariance matrix given by:

$$\boldsymbol{\mu}_{all} = \begin{bmatrix} \mu \\ \boldsymbol{\mu}_t \end{bmatrix} \text{ and } \boldsymbol{\Sigma}_{all} = \begin{bmatrix} \sigma^2 & \mathbf{a}^T\mathbf{A}_t \\ \mathbf{A}^T_t\mathbf{a} & \boldsymbol{\Sigma}_t \end{bmatrix}. \tag{7}$$

We may then apply the result of Theorem 3.1 to obtain the conditional PDF of d, given the delay information from the test structures. We know the conditional distribution of d is Gaussian, and its mean and variance can be obtained as:

$$\text{PDF}(d_{cond}) = \text{PDF}\left(d | (\mathbf{d}_t = \mathbf{d}_r)\right) \sim N(\bar{\mu}, \bar{\sigma}^2) \tag{8a}$$

$$\bar{\mu} = \mu + \mathbf{a}^T\mathbf{A}_t\boldsymbol{\Sigma}^{-1}_t(\mathbf{d}_r - \boldsymbol{\mu}_t) \tag{8b}$$

$$\bar{\sigma}^2 = \sigma^2 - \mathbf{a}^T\mathbf{A}_t\boldsymbol{\Sigma}^{-1}_t\mathbf{A}^T_t\mathbf{a}. \tag{8c}$$

C. Interpretation of the Conditional PDF

In this section, we analyze the information provided by the equations that represent the conditional PDF. From equations (8b) and (8c), we conclude that while the conditional mean of the original circuit is adjusted making use of the result vector,

[2] The nominal value of the delay of the circuit is the delay value when no parameter variations are present. This can be computed exactly by a conventional static timing analysis with all parameters at their nominal values. However, because of the approximation of the max operation in the statistical timer, the mean value we computed from the pert-like traversal is more compatible with the rest of the canonical form.

$\mathbf{d}_r$, the conditional variance is *independent* of the measured delay values, $\mathbf{d}_r$.

Examining Equation (8c) more closely, we see that for a given circuit, the variance of its delay before measuring the test structures, σ^2, and the coefficient vector, $\mathbf{a}$, are fixed and can be obtained from SSTA. The only variable that is affected by the test mechanism is the coefficient matrix of the test structures, $\mathbf{A}_t$, which also impacts $\mathbf{\Sigma}_t$. Therefore, the value of the conditional variance can be modified by adjusting the matrix $\mathbf{A}_t$. We know that $\mathbf{A}_t$ is the coefficient matrix formed by the sensitivities with respect to the principal components of the test structures. The size of $\mathbf{A}_t$ is determined by the number of test structures on the chip, and the entry values of $\mathbf{A}_t$ is related to the type of the test structures and their locations on the chip. Therefore if we use the same type of test structures on the circuit, then by varying their number and locations, we can modify the matrix $\mathbf{A}_t$, hence adjust the value of the conditional variance. Intuitively, this implies that the value of the conditional variance depends on how many test structures we have, and how well the test structures are distributed, in the sense of capturing spatial correlations between variables.

In our problem, $\mathbf{A}_t^T \in \mathbf{R}^{n \times m}$, where n is the number of test structures on chip, and m is the number of principal components. In the grid-based spatial correlation model, a large circuit is usually tessellated into numerous grids, and hence is affected by numerous principal components, whereas the number of test structures we can place on-chip is limited by several factors mentioned in Section II. Therefore n is usually less than m. Theorem 3.1 assumes that $\mathbf{\Sigma}_t = \mathbf{A}_t^T \mathbf{A}_t$ is of full rank and has an inverse, which means $\mathbf{A}_t^T$ must have full row rank. A detailed discussion about the ranks of $\mathbf{A}_t^T$ and $\mathbf{\Sigma}_t$ can be found in Section IV. For the present, we will assume that $\mathbf{A}_t^T$ is of full row rank.

Based on this assumption, consider the special case when $m = n$; in other words, that the number of test structures is identical to the number of PCA components. Intuitively, this means that we have independent data points that can predict the value of each of these components. In this case, $\mathbf{A}_t$ is a square matrix with full rank and has an inverse $\mathbf{A}_t^{-1}$. Substituting $\mathbf{\Sigma}_t^{-1} = (\mathbf{A}_t^T \mathbf{A}_t)^{-1} = \mathbf{A}_t^{-1}(\mathbf{A}_t^T)^{-1}$ into Equation (8b),

$$\begin{aligned} \bar{\mu} &= \mu + \mathbf{a}^T \mathbf{A}_t \mathbf{\Sigma}_t^{-1}(\mathbf{d}_r - \boldsymbol{\mu}_t) \\ &= \mu + \mathbf{a}^T (\mathbf{A}_t^T)^{-1}(\mathbf{d}_r - \boldsymbol{\mu}_t). \end{aligned} \tag{9}$$

It is interesting to note that the term $(\mathbf{A}_t^T)^{-1}(\mathbf{d}_r - \boldsymbol{\mu}_t)$ is the solution of the linear equations

$$\mathbf{d}_t = \boldsymbol{\mu}_t + \mathbf{A}_t^T \mathbf{p} = \mathbf{d}_r \tag{10}$$

with $\mathbf{p}$ as the set of unknowns. Therefore, Equation (9) is equivalent to first solving $\mathbf{p}$ from linear equations (10), then substituting its value into Equation (1) (with uncorrelated parameters disregarded for now) to find d. We can see that in this case,

$$\begin{aligned} \bar{\sigma}^2 &= \sigma^2 - \mathbf{a}^T \mathbf{A}_t \mathbf{\Sigma}_t^{-1} \mathbf{A}_t^T \mathbf{a} \\ &= \sigma^2 - \mathbf{a}^T \mathbf{A}_t \mathbf{A}_t^{-1} (\mathbf{A}_t^T)^{-1} \mathbf{A}_t^T \mathbf{a} \\ &= \sigma^2 - \mathbf{a}^T \mathbf{a} \\ &= 0. \end{aligned} \tag{11}$$

Thus the derived PDF is an impulse function with the mean equal to the original circuit delay and the variance equal to zero, and Equation (8) automatically takes the special case of $m = n$ into consideration.

We end this section by pointing out that an equivalent way of looking at the problem is to first stack the PC vector $\mathbf{p}$ and the delay vector $\mathbf{d}_t$ together, where $\mathbf{p}$ is the original subspace, and $\mathbf{d}_t$ is the test subspace. From this, we obtain the conditional distribution of $\mathbf{p}$, using Theorem 3.1, as:

$$\text{PDF}(\mathbf{p}_{cond}) = \text{PDF}\left(\mathbf{p} | (\mathbf{d}_t = \mathbf{d}_r)\right) \sim N(\bar{\boldsymbol{\mu}}_{\mathbf{p}}, \bar{\mathbf{\Sigma}}_{\mathbf{p}}) \tag{12a}$$

$$\bar{\boldsymbol{\mu}}_{\mathbf{p}} = \mathbf{A}_t \mathbf{\Sigma}_t^{-1}(\mathbf{d}_r - \boldsymbol{\mu}_t) \tag{12b}$$

$$\bar{\mathbf{\Sigma}}_{\mathbf{p}} = \mathbf{I} - \mathbf{A}_t \mathbf{\Sigma}_t^{-1} \mathbf{A}_t^T \tag{12c}$$

where $\mathbf{I}$ represents the identity matrix, which is the unconditional covariance matrix of $\mathbf{p}$. The result (12) tells us that given the condition $\mathbf{d}_t = \mathbf{d}_r$, the mean and covariance matrix of $\mathbf{p}_{cond}$ are no longer $\mathbf{0}$ and $\mathbf{I}$. In other words, the entries in $\mathbf{p}_{cond}$ can no longer be perceived as principal components. Due to the linear relationship between $\mathbf{p}_{cond}$ and the process parameter variations, we are in fact gaining information on the parameter variations inside each grid.

According to Theorem 3.1, $\mathbf{p}_{cond}$ remains Gaussian distributed. Because d_{cond} has a linear relationship with $\mathbf{p}_{cond}$, d_{cond} is also Gaussian-distributed. Since $\mathbf{a}$ is fixed for a given circuit, the conditional mean and variance of d can be calculated as:

$$\begin{aligned} \bar{\mu} &= \mu + \mathbf{a}^T E(\mathbf{p}_{cond}) = \mu + \mathbf{a}^T \mathbf{A}_t \mathbf{\Sigma}_t^{-1}(\mathbf{d}_r - \boldsymbol{\mu}_t) \\ \bar{\sigma}^2 &= E(\mu + \mathbf{a}^T \mathbf{p}_{cond} - (\mu + \mathbf{a}^T \bar{\boldsymbol{\mu}}_{\mathbf{p}}))^2 \\ &= \mathbf{a}^T E((\mathbf{p}_{cond} - \bar{\boldsymbol{\mu}}_{\mathbf{p}})(\mathbf{p}_{cond} - \bar{\boldsymbol{\mu}}_{\mathbf{p}})^T)\mathbf{a} \\ &= \mathbf{a}^T (\mathbf{I} - \mathbf{A}_t \mathbf{\Sigma}_t^{-1} \mathbf{A}_t^T)\mathbf{a} \\ &= \sigma^2 - \mathbf{a}^T \mathbf{A}_t \mathbf{\Sigma}_t^{-1} \mathbf{A}_t^T \mathbf{a} \end{aligned} \tag{13}$$

Not surprisingly, this end result is exactly the same as (8). However, dividing the derivation into two steps, as we have done here, provides additional insight into the problem.

IV. Locally Redundant but Globally Insufficient Test Structures

In practice, correlation matrices tend to be sparse since the spatial density of correlation goes up to a limited radius. As a consequence, it is found that a number of entries of each row of A_t^T are zero for typical correlation matrices. For such a scenario, it is possible that we place too many test structures that collectively capture only a small portion of PCs, with the coefficients of other PCs being all zeros. In other words, in some portion of the chip, the number of test structures may exceed the number of PCs with nonzero coefficients, but overall there are not enough test structures to actually compute the delay of the original circuit. We refer to this as a *locally redundant but globally insufficient* problem.

We show below that in such a scenario $\mathbf{\Sigma}_t$ would be rank deficient. While this problem can be overcome by appropriate placement of the test structures, the placement of these structures is beyond the scope of this paper: we assume that this has been done by the designer, and that it is provided as an input to our problem. Instead, we provide a general solution to take the locally redundant but globally insufficient problem into consideration during the evaluation of the conditional distribution. Our approach groups the redundant equations together and use a least-squares approach to capture the information.

With locally redundant but globally insufficient test structures, the matrix $\mathbf{A}_t^T$ has the following structure after grouping the all-zero coefficients for a group of test structures together:

$$\mathbf{A}_t^T = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{0} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix} \tag{14}$$

where $\mathbf{B}_{11} \in \mathbf{R}^{s \times q}$, with s being the number of test structures that have all-zero coefficients for the last $n-q$ principal components, and $s > q$, which means we have locally redundant test structures for these q principal components. Since we have prohibited two test structures with the same configurations from being placed in one grid, $\mathbf{B}_{11}$ must be of full column rank with rank q. Therefore, the maximum rank of $\mathbf{A}_t^T$ is $q+n-s$, less than n, so $\mathbf{\Sigma}_t$ also has a rank less than n and is singular. In this case, Equation (10) can be divided into two sets of equations:

$$\mathbf{B}_{11}\mathbf{p}_u = \mathbf{d}_{r,u} - \mu_{r,u} \tag{15}$$

$$\mathbf{B}_{21}\mathbf{p}_u + \mathbf{B}_{22}\mathbf{p}_v = \mathbf{d}_{r,v} - \mu_{r,v} \tag{16}$$

where $\mathbf{p}_u$, $\mathbf{p}_v$, $\mathbf{d}_{r,u}$, $\mathbf{d}_{r,v}$, $\mu_{r,u}$, $\mu_{r,v}$ are sub-vectors of the PC vector $\mathbf{p}$, the result vector $\mathbf{d}_r$, and the mean vector μ_t, correspondingly. Note that $\mathbf{B}_{11}$ is not square, and Equation (15) is an over-determined system. This can be solved in several ways, and we take the least-squares solution as its equivalence.

$$\bar{\mathbf{p}}_u = (\mathbf{B}_{11}^T \mathbf{B}_{11})^{-1} \mathbf{B}_{11}^T (\mathbf{d}_{r,u} - \mu_{r,u}) \tag{17}$$

Under conditions (17) as well as (16), the conditional PDF of d can be computed as follows.

$$\begin{aligned} & \text{PDF}(d_{cond}) = \text{PDF}(d|\mathbf{d}_t = \mathbf{d}_r) \\ = \; & \text{PDF}(d|\mathbf{p}_u = \bar{\mathbf{p}}_u, \mathbf{B}_{21}\bar{\mathbf{p}}_u + \mathbf{B}_{22}\mathbf{p}_v = \mathbf{d}_{r,v} - \mu_{r,v}) \end{aligned} \tag{18}$$

This step is safe because Equation (15) does not provide any information for $\mathbf{p}_v$. The statistical properties of $\mathbf{p}_v$ have not been changed, meaning they can still act as PCs. Assume $\mathbf{a}_u$ is the sub-vector of $\mathbf{a}$ corresponding to $\mathbf{p}_u$, and $\mathbf{a}_v$ is the sub-vector corresponding to $\mathbf{p}_v$, then $d = \mu + \mathbf{a}_u^T \bar{\mathbf{p}}_u + \mathbf{a}_v^T \mathbf{p}_v$. The mean, variance of d and $\mathbf{B}_{21}\bar{\mathbf{p}}_u + \mathbf{B}_{22}\mathbf{p}_v$, and their covariance can be easily updated. The same technique introduced in Section III can be applied to calculate the final conditional PDF of d.

Special cases include when $q = m$, in which case we can compute all the PCs and the delay of the original circuit by applying least-squares approach to the whole system, and when $s = n$, in which case we cannot obtain any information on $\mathbf{p}_v$, and the PCs will still be uncorrelated Gaussians with zero mean and unit variance.

V. Spatially Uncorrelated Parameters

In Section III, we had developed a theory for determining the conditional distribution of the delay, d, of the original circuit, under the data vector, $\mathbf{d}_r$, provided by the test structures. This derivation neglected the random variables R and $\mathbf{R}_t$ in the canonical form of Equation (1) and (3), corresponding to spatially uncorrelated variations.

We now extend this theory to include such effects, which may arise due to parameters such as T_{ox} and N_A that can take on a different and spatially uncorrelated value for each transistor in the layout. While these parameters can show both inter-die and intra-die variations, because the inter-die variation of each such parameter can be regarded as a PC and easily incorporated in the procedure of Section III, we hereby focus on the intra-die variations of these parameters, i.e., the purely random part. Thus, R is the random variable generated by merging the intra-die variations for each gate during traversal of the whole circuit [3], with mean 0 and variance $\sigma_R^2 \neq 0$. Considering this effect, the variance of the original circuit is adjusted to be

$$\sigma'^2 = \mathbf{a}^T\mathbf{a} + \sigma_R^2. \tag{19}$$

The covariance matrix of the test structures must also be updated as follows:

$$\mathbf{\Sigma}_t' = \mathbf{A}_t^T\mathbf{A}_t + diag[\sigma_{R_{t,1}}^2, \sigma_{R_{t,2}}^2, \cdots, \sigma_{R_{t,n}}^2]. \tag{20}$$

The same kind of technique from Section III can still be applied. However, in this case, due to the nonzero diagonal matrix added to $\mathbf{\Sigma}_t$, $\bar{\sigma}$ is never equal to zero, meaning that we can never compute the actual delay of the original circuit, which is a fundamental limitation of any testing-based diagnosis method. Any such strategy is naturally limited to spatially correlated parameters. The values of uncorrelated parameters in the original circuit cannot be accurately replicated in the test structures: these values may change from one device to the next, and therefore, their values in a test structure cannot perfectly capture their values in the original circuit.

VI. Changing the Number of Stages in the ROs

In Section V, it was shown that spatially uncorrelated parameter variations impose a challenge for our method, since it is physically impossible for a test structure to capture uncorrelated variations. However, it is possible to dilute the effects of uncorrelated variations, and to overcome this problem, an intuitive idea is to increase the number of stages of the RO test structures.

The essential idea of increasing the number of stages is that it leaves the spatially correlated variations unchanged: since each RO is small and lies within a spatial correlation grid, all spatially correlated parameters that affect its delay show identical variations. However, variations for spatially uncorrelated parameters may be in opposite directions and thus increasing the number of stages increases the likelihood of cancellations, implying that spatially uncorrelated parameters are likely to become relatively less important. In other words, this implies that the delay of each RO as a variable will be more correlated to the delay of the original circuit.

On the other hand, while increasing the number of stages of the ROs increases the correlation coefficient between the delays of the RO and the original circuit, it also makes the delays of the ROs more correlated with each other. This suggests that the RO test structures may collectively yield less independent information about the variations.

There is a clear trade-off here, and in this section, we illustrate the above qualitative argument from a more rigid, mathematical perspective, and present it in a quantitative way. We will show in Section VII that for our implementation, increasing the number of stages does indeed yield better estimations of the post-silicon delay.

As stated in Section III, the delay of the original circuit can be written in the canonical form of Equation (1). We rewrite the equation below.

$$d = \mu + \sum_{i=1}^{m} a_i p_i + R = \mu + \mathbf{a}^T \mathbf{p} + R. \tag{21}$$

Similarly, the delay of RO i can be written in the form of Equation (2), which is

$$d_{t,i} = \mu_{t,i} + \mathbf{a}_{t,i}^T \mathbf{p} + R_{t,i}. \tag{22}$$

First, if we assume that there is only one RO i on the chip, Equation (8c) becomes

$$\bar{\sigma}^2 = \sigma^2 - \frac{\mathbf{a}^T \mathbf{a}_{t,i} \mathbf{a}_{t,i}^T \mathbf{a}}{\sigma_{t,i}^2} = \sigma^2 \left(1 - \rho_i^2\right). \tag{23}$$

where ρ_i is the correlation coefficient between the delay of the original circuit and the delay of RO i. It is obvious that in this case, the result only depends on ρ_i.

Second, we explain how the number of stages affects the value of ρ_i, so that we can observe clearly how the number of stages affects our results. Let us assume that RO i has k stages, and for purposes of illustration, we will assume that each stage of the RO is identical, with a canonical delay of the form $\alpha_i + \sum_{j=1}^{m} \gamma_{ij} p_j + \zeta_i = \alpha_i + \boldsymbol{\Gamma}_i \mathbf{p} + \zeta_i$. The half-period of RO i, which is a surrogate for its delay, is therefore given by

$$d_{t,i} = k\alpha_i + k\boldsymbol{\Gamma}_i \mathbf{p} + \sqrt{k}\zeta_i \tag{24}$$

From Equation (20) in Section V, the variance of the delay of RO i can be written as

$$\sigma_{t,i}^2 = k^2 \boldsymbol{\Gamma}_i^T \boldsymbol{\Gamma}_i + k\zeta_i^2. \tag{25}$$

The correlation coefficient between RO i and the original circuit can thus can be calculated from the relation:

$$\rho_i^2 = \frac{k^2 \mathbf{a}^T \boldsymbol{\Gamma}_i \boldsymbol{\Gamma}_i^T \mathbf{a}}{\sigma^2 \left(k^2 \boldsymbol{\Gamma}_i^T \boldsymbol{\Gamma}_i + k\zeta_i^2\right)} = \frac{\mathbf{a}^T \boldsymbol{\Gamma}_i \boldsymbol{\Gamma}_i^T \mathbf{a}}{\sigma^2 \left(\boldsymbol{\Gamma}_i^T \boldsymbol{\Gamma}_i + \frac{1}{k}\sigma_r^2\right)}. \tag{26}$$

It is easy to see that as k increases, the correlation coefficient between RO i and the original circuit increases, implying that the conditional variance of the delay of the original circuit decreases. Therefore, we have more specific information about the delay of the original circuit. This is in accordance with the intuition that increasing the number of stages in the RO helps in reducing the effect of the spatially uncorrelated parameters.

Third, we illustrate the fact that as the number of stages increases, the ROs can become more correlated with each other and might not give as much information collectively. To see this, we consider the delays of two ROs $d_{t,1}$ and $d_{t,2}$. If we assume that each has k stages, then

$$d_{t,1} = k\alpha_1 + k\boldsymbol{\Gamma}_1^T \mathbf{p} + \sqrt{k}\zeta_1 \tag{27}$$

$$d_{t,2} = k\alpha_2 + k\boldsymbol{\Gamma}_2^T \mathbf{p} + \sqrt{k}\zeta_2. \tag{28}$$

The correlation coefficient between the two can be calculated as

$$\begin{aligned} \rho_{1,2} &= \frac{k^2 \boldsymbol{\Gamma}_1^T \boldsymbol{\Gamma}_2}{\left(k^2 \boldsymbol{\Gamma}_1^T \boldsymbol{\Gamma}_1 + k\zeta_1^2\right)\left(k^2 \boldsymbol{\Gamma}_2^T \boldsymbol{\Gamma}_2 + k\zeta_2^2\right)} \\ &= \frac{\boldsymbol{\Gamma}_1^T \boldsymbol{\Gamma}_2}{\left(\boldsymbol{\Gamma}_1^T \boldsymbol{\Gamma}_1 + \frac{1}{k}\zeta_1^2\right)\left(\boldsymbol{\Gamma}_2^T \boldsymbol{\Gamma}_2 + \frac{1}{k}\zeta_2^2\right)}. \end{aligned} \tag{29}$$

It is easily observed that as k increases, the correlation coefficient between the delays of the two ROs increases.

The conditional variance of the delay of the original circuit can be calculated based on the testing results of the delays of these two ROs, using Equation (8c), as

$$\begin{aligned} \bar{\sigma}^2 &= \sigma^2 - \mathbf{a}^T \begin{bmatrix} \mathbf{a}_{t,1} & \mathbf{a}_{t,2} \end{bmatrix} \begin{bmatrix} \sigma_{t,1}^2 & \mathbf{a}_{t,1}^T \mathbf{a}_{t,2} \\ \mathbf{a}_{t,2}^T \mathbf{a}_{t,1} & \sigma_{t,2}^2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{a}_{t,1}^T \\ \mathbf{a}_{t,2}^T \end{bmatrix} \mathbf{a} \\ &= \sigma^2 \left(1 - \frac{\rho_1^2 + \rho_2^2 - 2\rho_1\rho_2\rho_{1,2}}{1 - \rho_{1,2}^2}\right) \end{aligned} \tag{30}$$

This result confirms our intuition that the conditional variance of the delay of the original circuit is not only dependent upon the correlation coefficient between the delay of the original circuit and the delay of each RO (ρ_1,ρ_2), but also dependent upon the correlation coefficient between the two ROs ($\rho_{1,2}$).

To see the effect of k on the conditional variance more clearly, we write the above equation as

$$\bar{\sigma}^2 = \sigma^2 - \frac{C_1^2\left(V_2 + \frac{1}{k}\zeta_2^2\right) - C_2^2\left(V_1 + \frac{1}{k}\zeta_1^2\right) + 2C_1C_2\boldsymbol{\Gamma}_1^T \boldsymbol{\Gamma}_2}{\left(V_2 + \frac{1}{k}\zeta_2^2\right)\left(V_1 + \frac{1}{k}\zeta_1^2\right) - \left(\boldsymbol{\Gamma}_1^T \boldsymbol{\Gamma}_2\right)^2} \tag{31}$$

where $C_i = \mathbf{a}^T \boldsymbol{\Gamma}_i$ and $V_i = \boldsymbol{\Gamma}_i^T \boldsymbol{\Gamma}_i$ are not dependent on k. As k increases, both the numerator and the denominator decrease, the function is not guaranteed to be monotonic with respect to k. Therefore theoretically increasing the number of stages doesn't necessarily reduce the conditional variance of the delay of the original circuit we can get. We show in Section VII that for the practical results that we show, we lie within a monotone decreasing region with respect to k.

VII. Experimental Results

We summarize the proposed post-silicon statistical delay prediction approach as follows:

We use the software package *MinnSSTA* [2] to perform SSTA, and use Monte-Carlo methods to test our approach. The original circuits correspond to the ISCAS89 benchmark suite, and each test structure is assumed to be a RO. Specifically,

Algorithm 1 Post-silicon statistical delay prediction.

```
Perform SSTA on both the original circuit and the test structures to determine $\mu$, $\mathbf{a}$, $\boldsymbol{\mu}_t$, $\mathbf{A}_t$, and $\sigma_R$, $\sigma_{R_{t,1}}, \cdots, \sigma_{R_{t,n}}$.
After fabrication, test the delay of the test structures on-chip to obtain $\mathbf{d}_r$.
Compute the conditional mean $\bar{\mu}$ and variance $\bar{\sigma}^2$ for the original circuit using the expressions in Equation (8).
```

the RO used in our experiments has five stages. Section VI, combined with simulation later in this section, shows that increasing the number of stages can compensate for the effects of spatially uncorrelated parameter variations in practice.

A grid-based spatial correlation model [19] is used to compute the covariance matrix for each spatially correlated parameter. Under this model, if the number of grids is G, and the number of spatially correlated parameters being considered is P, then the total number of principal components is no more than $P \cdot G$. Because we only use one type of test structure in the experiments, we place at most one RO inside each grid. The parameters that are considered as sources of spatially correlated variations include the effective channel length L, the transistor width W, the interconnect width W_{int}, the interconnect thickness T_{int} and the inter-layer dielectric H_{ILD}. The dopant concentration, N_A, is regarded as the source of spatially uncorrelated variations. For interconnects, instead of two metal tiers used in [20], we use four metal tiers (corresponding to two horizontal and two vertical layers). Parameters of different metal tiers are assumed to be uncorrelated. Table I lists the level of parameter variations assumed in this work. The process parameters are Gaussian-distributed, and their mean and 3σ values are shown in the table. For each parameter, half of the variational contribution is assumed to be from inter-die variations and half from intra-die variations. Our experiments ignore the effects of systematic variations, but if available, this information may be used to alter the nominal values and sensitivities of the gate delays. We assume this variation model is accurate in our simulation. In practice, the model should be tailored according to manufacturing data.

TABLE I
PARAMETERS USED IN THE EXPERIMENTS.

	L	W	W_{int}	T_{int}	H_{ILD}	N_A
						nmos/pmos
	(nm)	(nm)	(nm)	(nm)	(nm)	(10^{17}cm^{-3})
μ	60.0	150.0	150.0	500.0	300.0	9.7/10.04
3σ	12.0	22.5	30.0	75.0	45.0	1.45

In the *first set* of experiments, only one variation is taken into consideration in the Monte Carlo analysis: in this case, we consider the effective channel length L, which we observe to be the dominant component of intra-die variations. Under the grid-based correlation model, there will only be G independent variation sources in this case, and by providing G test structures, we can use the techniques in Section III to calculate the delay of the original circuit.

The result is shown as a scatter plot in Figure 4. The method is applied to 1000 chips: we simulate this by performing 1000 Monte-Carlo simulations on each benchmark, each corresponding to a different set of parameter values. For each of these values, we compute the deterministic delays of the test structures[3] and the original circuit: we use the former as inputs to our approach, and compare the delay from our statistical delay prediction method with the latter. The fact that all of the points lie closely around the $y = x$ line indicates that the circuit delays predicted by our approach matches very well with the Monte-Carlo simulation results.

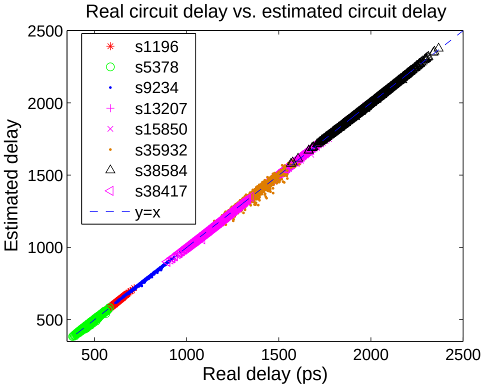

Fig. 4. The scatter plot: real circuit delay vs. predicted circuit delay.

The precise testing error for each benchmark is listed in Table II. If we denote the delay of the original circuit at a sample point as d_{orig} and the delay of the original circuit, as predicted by our statistical delay prediction approach, as d_{pred}, the test error for each simulation is defined as $\frac{|d_{orig} - d_{pred}|}{d_{orig}} \times 100\%$. The second column of the table shows the average test error, based on all 1000 sample points, which indicates the overall aggregate accuracy: this is seen to be well below 1% in almost all cases. The third column shows the maximum deviation from the mean value of statistical timing over all 1000 sample points, as a fraction of the mean. The test error at this point is shown in the fourth column of the table. These two columns indicate that the results are accurate even when the sampled delay is very different from the mean value.

Note that in theory, according to the discussion in Section III, when one test structure is placed in each variational grid, the prediction should be perfect. However, some inaccuracies creep in during SSTA, primarily due to the error in approximating the *max* operation in SSTA, during which the the distribution of the maximum of two Gaussians, which is a non-Gaussian, is approximated as a Gaussian to maintain the invariant. For circuits like s35932, which show the largest average error among this set, of under 2%, the canonical form (1) is not perfectly accurate in modeling the circuit delay. Note that our experimental setup is based on simulation, and does not include any measurement noise.

[3] Because of the way in which these values are computed in our experimental setup, variations in the test structure delays are only caused by random variations. In practice, the measured test structure delays will consist of deterministic variations, random variations, and measurement noise. It is assumed here that standard methods can be used to filter out the effects of the first and the third factor.

TABLE II

TEST ERRORS CONSIDERING ONLY VARIATIONS IN L.

Benchmark	Average Error	Maximum Deviation (% of mean)	Error at Maximum Deviation
s1196	0.18%	24.2%	0.20%
s5378	0.58%	25.7%	0.02%
s9234	0.35%	22.7%	0.50%
s13207	0.09%	25.2%	0.51%
s15850	0.25%	26.1%	0.47%
s35932	1.31%	22.4%	1.01%
s38584	0.10%	27.5%	0.69%
s38417	0.09%	27.4%	0.58%

For the unoptimized ISCAS89 benchmark suite, one or a small number of critical paths tend to dominate the circuit, which is unrealistic. However, s35932 is an exception and thus is used to compare our approach with the critical path replica approach currently used in ABB. We assume that in the critical path approach, the entire critical path for the nominal design can be perfectly replicated, and compare the delay of that path and the delay of the whole circuit during the Monte-Carlo simulation. It is observed that the critical path replica can show a maximum error of 15.5%, while our approach has a maximum error of 6.92%, an improvement of more than 50%. The average error of critical path replica for this circuit is 1.92%, also significantly larger than our result of 1.31%.

To show the confidence scalability of our approach, in the *second set* of experiments, we consider cases in which the number of test structures is insufficient to completely predict the delay of the original circuit. In this experiment, different numbers of test structures are implanted on the die. Specifically, for circuits divided into 16 grids, we investigate Case 1, when 10 test structures and Case 2, when 5 test structures are available.

For circuits where the die is divided into 256 grids, Case 1 corresponds to a die with 150 test structures, and Case 2 to 60 test structures. To show how much more information than SSTA we can obtain from the test structures, we define $\sigma_{reduction}$ as $\frac{\sigma-\bar{\sigma}}{\sigma} \times 100\%$ which is independent of the test results but is dependent on how the available test structures are placed on the chip. To be as general as possible, we perform 1000 random selections of the grids to place test structures in. The μ, σ of the original circuit, obtained from SSTA, and the average $\bar{\sigma}$, $\sigma_{reduction}$ of the statistical delay prediction approach for both cases, over the 1000 selections, are listed in Table III for each benchmark circuit. It is observed that there is a trade-off between test structure overhead and $\sigma_{reduction}$. In order to understand what the result would be like if a really bad set of grids are selected to place test structures in, in this table we also show the minimum (Min.) $\sigma_{reduction}$ over the 1000 random selections for each circuit in both cases.

Figure 5 shows the predicted delay distribution for a typical sample of the circuit s38417, the largest circuit in the benchmark suite. Each curve in the circuit corresponds to a different number of test structures, and it is clearly seen that even when the number of test structures is less than G, a sharp PDF of the original circuit delay can still be obtained using our method, with a variance much smaller than provided by SSTA. The trade-off between the number of test structures and the reduction in the standard deviation can also be observed clearly. For this particular die, while SSTA can only assert that it can meet a 1400 ps delay requirement, using 150 test structures we can say with more than 99.7% confidence that the fabricated chip meets a 1040 ps delay requirement, and using 60 test structures we can say with such confidence that it can meet a 1080 ps delay requirement.

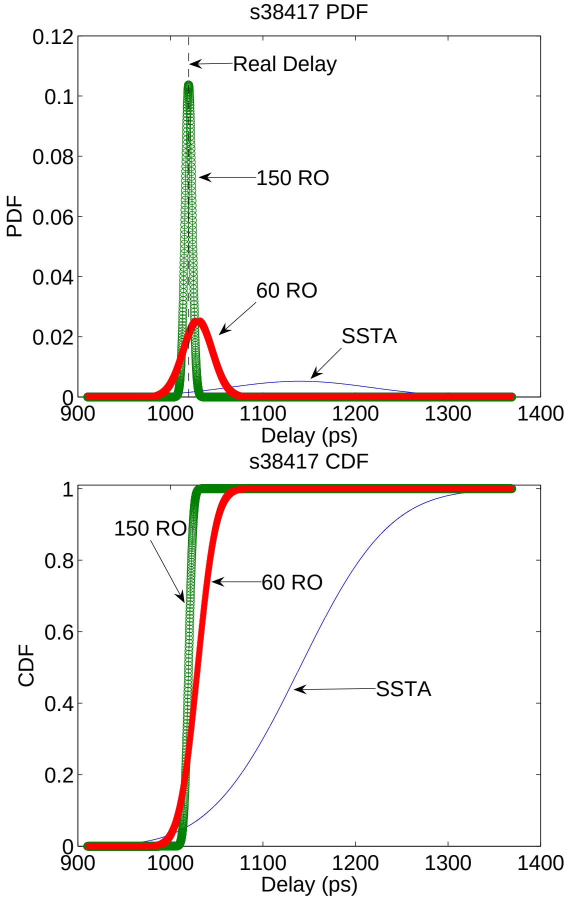

Fig. 5. PDF and CDF with insufficient number of test structures for circuit s38417 (considering L).

In our *third set* of experiments, we consider the most general case in which all parameter variations are included. While the first two sets of experiments provided general insight into our method, this third set shows the result of applying it to real circuits under the full set of parameter variations listed in Table I. In Case I of this set of experiments, the number of test structures is equal to the number of grids. The values of $\bar{\sigma}$ and $\sigma_{reduction}$ are fixed in this case. Case II and Case III are set up the same way as in Case 1 and Case 2, respectively, of the second set of experiments described earlier. The μ, σ of each benchmark circuit obtained by SSTA, the $\bar{\sigma}$, $\sigma_{reduction}$ for Case I, the average $\bar{\sigma}$, the average and minimum $\sigma_{reduction}$ for Case II and Case III obtained from the post-silicon statistical delay prediction are listed in Table IV. The distribution plot

TABLE III

PREDICTION RESULTS WITH INSUFFICIENT NUMBER OF TEST STRUCTURES (CONSIDERING L): CASE 1 AND CASE 2 ARE DISTINGUISHED BY THE NUMBER OF ROS AVAILABLE FOR EACH CIRCUIT.

Benchmark			SSTA Results		Case 1				Case 2			
Name	#Cells	#Grids	μ(ps)	σ(ps)	#RO	Avg. $\bar{\sigma}$(ps)	$\sigma_{reduction}$ Avg.	$\sigma_{reduction}$ Min.	#RO	Avg. $\bar{\sigma}$(ps)	$\sigma_{reduction}$ Avg.	$\sigma_{reduction}$ Min.
s1196	547	16	577.06	35.32	10	6.48	81.64%	73.6%	5	11.97	66.1%	64.1%
s5378	2958	16	475.97	29.84	10	5.96	80.02%	75.5%	5	10.77	63.9%	61.7%
s9234	5825	16	775.36	51.51	10	9.50	81.55%	68.1%	5	18.85	63.4%	56.5%
s13207	8260	256	1399.8	92.81	150	9.63	89.62%	81.9%	60	18.56	80.0%	70.4%
s15850	10369	256	1573.7	100.48	150	8.25	91.79%	86.7%	60	16.88	83.2%	78.0%
s35932	17793	256	1359.5	82.17	150	11.08	86.52%	76.8%	60	27.69	76.3%	70.7%
s38584	20705	256	1994.0	120.83	150	16.54	86.31%	74.4%	60	29.96	75.2%	68.3%
s38417	23815	256	1139.8	76.38	150	9.40	87.69%	76.2%	60	17.87	76.6%	61.8%

TABLE IV

PREDICTION RESULTS CONSIDERING ALL PARAMETER VARIATIONS: CASE I, CASE II AND CASE III ARE DISTINGUISHED BY THE NUMBER OF ROS.

Benchmark	SSTA Results		Case I			Case II				Case III			
	μ (ps)	σ (ps)	#RO	$\bar{\sigma}$ (ps)	$\sigma_{reduction}$	#RO	Avg. $\bar{\sigma}$ (ps)	$\sigma_{reduction}$ Avg.	$\sigma_{reduction}$ Min.	#RO	Avg. $\bar{\sigma}$ (ps)	$\sigma_{reduction}$ Avg.	$\sigma_{reduction}$ Min.
s1196	577.42	45.61	16	11.32	75.2%	10	12.67	72.2%	65.3%	5	15.20	66.7%	58.4%
s5378	475.65	37.24	16	6.35	82.9%	10	7.69	79.4%	71.4%	5	10.28	72.4%	59.8%
s9234	776.79	62.63	16	9.17	85.4%	10	12.20	80.5%	66.7%	5	17.21	72.5%	56.2%
s13207	1404.25	109.41	256	20.90	80.9%	150	22.97	79.0%	74.6%	60	27.13	75.2%	66.5%
s15850	1579.73	119.45	256	19.59	83.6%	150	21.09	82.3%	79.4%	60	24.69	79.3%	73.7%
s35932	1371.55	98.45	256	24.75	74.9%	150	27.11	72.5%	67.7%	60	30.69	68.8%	63.9%
s38584	2011.62	147.46	256	39.47	73.2%	150	43.16	70.7%	64.7%	60	48.77	66.9%	60.8%
s38417	1146.56	89.84	256	22.01	75.5%	150	24.09	73.2%	67.2%	60	28.17	68.6%	57.3%

for this set of experiment is similar to that in Figure 5, and the conditional PDFs of one particular sample of the circuit s1196 for Case II and Case III are shown in Section I as Figure 3, with the SSTA PDF as a comparison. Note that the conditional PDF obtained by our approach would be even sharper for Case I.

The reduction in the standard deviation is only able to demonstrate that our predicted delay is within a certain range. To see whether the prediction is reasonable and accurate, in our third set of experiments, we also perform the following Monte-Carlo simulations. In Case I of this experiment, because in each grid we have one RO, we just perform one thousand Monte-Carlo simulations based on this structure. In Case II and Case III, however, the number of ROs is smaller than the number of grids. Therefore we use five randomly selected sets of grids to place ROs in, and for each set of grids, we perform 1000 Monte-Carlo simulations, which means totally we have 5000 Monte-Carlo simulations for each circuit of Case II and Case III. While each Monte-Carlo simulation generates a specific delay number, our prediction result is a conditional distribution of the delay. Therefore if the Monte-Carlo result falls within $\pm 3\bar{\sigma}$ of the predicted distribution, then we call the result a *hit*. Otherwise, we call it a *miss*. The *hit rate* of our prediction for a circuit is then defined as the number of hits divided by the total number of Monte-Carlo simulations. We show the hit rates for each circuit in Table VI. It is observed that most hit rates are above 99.9%.

TABLE VI

HIT RATES CONSIDERING ALL PARAMETER VARIATIONS: CASE I, CASE II AND CASE III ARE DISTINGUISHED BY DIFFERENT NUMBER OF ROS AVAILABLE FOR EACH CIRCUIT.

Benchmark	Hit Rate		
	Case I	Case II	Case III
s1196	100.0%	99.9%	99.9%
s5378	99.8%	99.7%	99.9%
s9234	100.0%	99.9%	99.9%
s13207	99.9%	100.0%	100.0%
s15850	99.9%	99.9%	99.9%
s35932	97.2%	97.7%	98.7%
s38584	100.0%	99.9%	99.8%
s38417	99.9%	100.0%	99.9%

Now we show that for the ISCAS89 benchmark circuits and our experimental setup, increasing the number of stages can compensate for the effect of spatially uncorrelated parameter variations and give us more specific information about the circuit delay after fabrication. We assume that each grid contains an RO, and for each RO, every stage has the same timing characteristics. Therefore we can use the coefficients and the spatially uncorrelated variable calculated for a 5-stage RO to derive the corresponding coefficients and spatially uncorrelated variable for a unit stage of that RO. Based on these timing characteristics of one unit stage, the timing characteristics of a RO with any number of stages can be

TABLE V

RUNTIME RESULTS.

Circuit	s1196	s5378	s9234	s13207	s15850	s35932	s38584	s38417
Runtime (sec)	5.68×10^{-4}	5.70×10^{-4}	5.96×10^{-4}	0.39	0.39	0.35	0.37	0.68

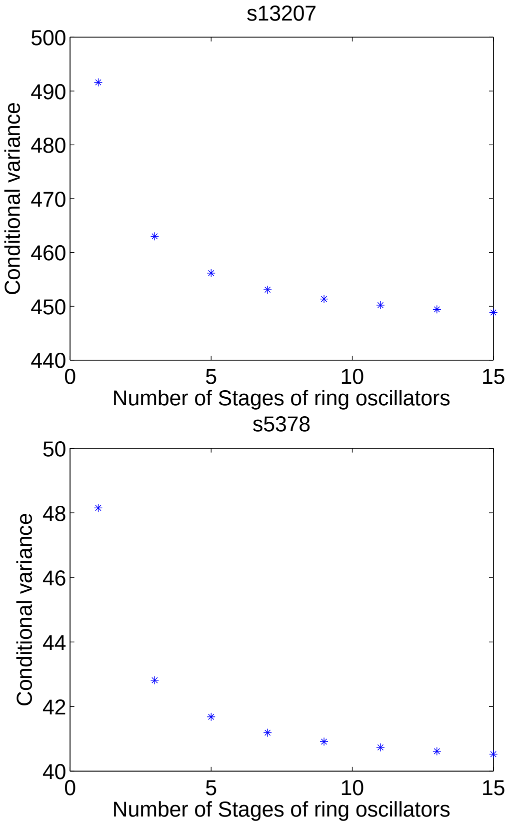

Fig. 6. Conditional variance of the delay of the original circuit with respect to the number of stages of ROs.

calculated. This procedure is repeated for each of the ROs on chip. For each circuit we draw a curve, with the y axis being the conditional variance of the circuit delay computed by our approach, and the x axis being the number of unit stages we have for every ring oscillator built on this circuit. This plot shows that for our set of benchmarks, as the number of stages increases, the conditional variance we obtain becomes progressively smaller. Sample results of the circuits s13207 and s5378 are shown in Figure 6. It is easily observed that the curves are monotonically decreasing. The results are similar for all other circuits in the benchmark set.

Finally we provide runtime results for our approach. It is easily observed that our algorithm can be divided into two parts, separated by the physical measurements of the delays of the ring oscillators. The first part corresponds to SSTA, and because the framework is similar to [2], the readers are referred to that paper for a runtime estimate. The runtime for the second part, which is conditional PDF evaluation, is listed in Table V. The experiments are run on a Linux PC with a 2.0GHz CPU and 256MB memory. The results we show here are for Case I of Table IV, where the matrix Σ_t in Section II is the largest of all three cases. It is shown that for all the benchmark circuits, the runtime is less than one second.

VIII. CONCLUSION

In this paper, a general framework for the post-silicon statical delay prediction approach is proposed, using SSTA and a conditional PDF evaluation method, making use of test data from RO test structures. Future directions include the development of methods for placing these structures optimally and designing appropriate structures that are better at delay prediction than ring oscillators.

In cases where the circuit is dominated by a single critical path (this is not often the case, since most circuits are timing-optimized, which implies that there are numerous near-critical paths), it may be beneficial to use a critical path replica instead of our ring oscillator based scheme. The critical path replica can also be viewed as a type of test structure, which means that after determining the nominal critical path, we can replicate it, perform SSTA on this path, and calculate the conditional variance of the original circuit delay, given that the delay of this path is known. If a circuit is highly dominated by this path, then the conditional variance would be small. We then can compare the conditional variance calculated in this way with the conditional variance calculated by our approach.

Depending on which variance is smaller, we can choose the appropriate approach and start building the circuit embedded with the proper test structure. This choice can be made entirely through presilicon analysis. The variances of the conditional PDFs for the two possible test structures (a set of RO measurements, or a critical path replica) may be computed using Equation (8c). Note that (8c) provides results that are independent of measurement data, and hence depending on which structure has the smaller covariance, we can choose an appropriate test structure.

IX. ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers, whose comments provided excellent feedback, resulting in an improvement in the quality of our paper.

REFERENCES

[1] S. S. Sapatnekar, *Timing.* Kluwer Academic Publishers, 2004.

[2] H. Chang and S. S. Sapatnekar, "Statistical Timing Analysis Considering Spatial Correlations using a Single PERT-Like Traversal," in *Proceedings of the IEEE/ACM International Conference on Computer Aided Design*, pp. 621–625, Nov. 2003.

[3] H. Chang and S. S. Sapatnekar, "Statistical Timing Analysis Considering Spatial Correlations," *IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems*, vol. 24, pp. 1467–1482, Sept. 2005.

[4] C. Visweswariah, K. Ravindran, K. Kalafala, S. G. Walker, and S. Narayan, "First-order Incremental Block-Based Statistical Timing Analysis," in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 331–336, June 2004.

[5] H. Chang, V. Zolotov, S. Narayan, and C. Visweswariah, “Parameterized Block-Based Statistical Timing Analysis with Non-Gaussian Parameters, Nonlinear Delay Functions,” in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 71–76, June 2005.

[6] J. Singh and S. S. Sapatnekar, “Statistical Timing Analysis with Correlated Non-Gaussian Parameters using Independent Component Analysis,” in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 155–160, July 2006.

[7] A. Agarwal, D. Blaauw, V. Zolotov, and S. Vrudhula, “Statistical Timing Analysis Using Bounds and Selective Enumeration,” in *Proceedings of the ACM/IEEE International Workshop on Timing Issues in the Specification and Synthesis of Digital Systems*, pp. 29–36, Dec. 2002.

[8] A. Devgan and C. Kashyap, “Block-based Static Timing Analysis with Uncertainty,” in *Proceedings of the IEEE/ACM International Conference on Computer Aided Design*, pp. 607–614, Nov. 2003.

[9] Y. Zhan, A. J. Strojwas, X. Li, L. Pileggi, D. Newmark, and M. Sharma, “Correlation-Aware Statistical Timing Analysis with Non-Gaussian Delay Distributions,” in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 77–82, June 2005.

[10] B. Lee, L. Wang, and M. S. Abadir, “Refined Statistical Static Timing Analysis Through Learning Spatial Delay Correlations,” in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 149–154, July 2006.

[11] L. Wang, P. Bastani, and M. S. Abadir, “Design-Silicon Timing Correlation–A Data Mining Perspective,” in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 385–389, June 2007.

[12] A. Davoodi and A. Srivastava, “Variability Driven Gate Sizing for Binning Yield Optimization,” in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 956–964, July 2006.

[13] M. Abranmovici, P. Bradley, K. Dwarakanath, P. Levin, G. Memmi, and D. Miller, “A Reconfigurable Design-for-Debug Infrastructure for SoCs,” in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 7–12, July 2006.

[14] J. W. Tschanz, J. T. Kao, S. G. Narendra, R. Nair, D. A. Antoniadis, A. P. Chandrakasan, and V. De, “Adaptive Body Bias for Reducing Impacts of Die-to-Die and Within-Die Parameter Variations on Microprocessor Frequency and Leakage,” *IEEE Journal of Solid-State Circuits*, vol. 37, pp. 1396–1402, Nov. 2002.

[15] J. W. Tschanz, S. Narendra, R. Nair, and V. De, “Effectiveness of Adaptive Supply Voltage and Body Bias for Reducing the Impact of Parameter Variations in Low Power and High Performance Microprocessors,” *IEEE Journal of Solid-State Circuits*, vol. 38, pp. 826–829, May 2003.

[16] J. W. Tschanz, S. Narendra, A. Keshavarzi, and V. De, “Adaptive Circuit Techniques to Minimize Variation Impacts on Microprocessor Performance and Power,” in *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 23–26, May 2005.

[17] S. V. Kumar, C. H. Kim, and S. S. Sapatnekar, “Mathematically Assisted Adaptive Body Bias (ABB) for Temperature Compensation in Gigascale LSI Systems,” in *Proceedings of the Asia and South Pacific Design Automation Conference*, pp. 559–564, Jan. 2006.

[18] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis (3rd ed.)*. Prentice Hall, 1992.

[19] A. Agarwal, D. Blaauw, V. Zolotov, S. Sundareswaran, M. Zhao, K. Gala, and R. Panda, “Path-Based Statistical Timing Analysis Considering Inter- and Intra-die Correlations,” in *Proceedings of the ACM/IEEE International Workshop on Timing Issues in the Specification and Synthesis of Digital Systems*, pp. 16–21, Dec. 2002.

[20] Q. Liu and S. S. Sapatnekar, “Confidence Scalable Post-silicon Statistical Delay Prediction under Process Variations,” in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 497–502, June 2007.

PLACE
PHOTO
HERE

Qunzeng Liu received the B.S. degree from Zhejiang University, Hangzhou, China, in Electrical Engineering in 2005. He is currently pursuing the Ph.D. degree in the Department of Electrical and Computer Engineering at the University of Minnesota. His research interests include statistical analysis and optimization of electronic circuits. He received a best paper award at ISPD2009.

PLACE
PHOTO
HERE

Sachin S. Sapatnekar received the B.Tech. degree from the Indian Institute of Technology, Bombay in 1987, the M.S. degree from Syracuse University in 1989, and the Ph.D. degree from the University of Illinois at Urbana-Champaign in 1992. From 1992 to 1997, he was an assistant professor in the Department of Electrical and Computer Engineering at Iowa State University. He is currently the Robert and Marjorie Henle Chair and a Distinguished McKnight University Professor in the Department of Electrical and Computer Engineering at the University of Minnesota.

He is an author of four books and a coeditor of one volume, and has published mostly in the areas of timing and layout. He has held positions on the editorial board of the IEEE Transactions on VLSI Systems, and the IEEE Transactions on Circuits and Systems II, the IEEE Transactions on CAD, and has been a Guest Editor for the latter. He has served on the Technical Program Committee for various conferences, and as Technical Program and General Chair for the Tau workshop and ISPD. He is currently the Vice Chair for DAC. He has been a Distinguished Visitor for the IEEE Computer Society and a Distinguished Lecturer for the IEEE Circuits and Systems Society. He is a recipient of the NSF Career Award, three best paper awards at DAC, one at ICCD, one at ISPD, and the SRC Technical Excellence award. He is a fellow of the IEEE.