Confidence Scalable Post-Silicon Statistical Delay Prediction under Process Variations*

Qunzeng Liu
University of Minnesota
liuxx575@umn.edu

Sachin S. Sapatnekar
University of Minnesota
sachin@umn.edu

ABSTRACT

Due to increased variability trends in nanoscale integrated circuits, statistical circuit analysis has become essential. We present a novel method for post-silicon analysis that gathers data from a small number of on-chip test structures, and combines this information with pre-silicon statistical timing analysis to obtain narrow, die-specific, timing PDFs. Experimental results show that for the benchmark suite being considered, taking all parameter variations into consideration, our approach can get a PDF with the standard deviation 83.5% smaller on average than the SSTA result. The approach is scalable to smaller test structure overheads.

Categories and Subject Descriptors

B.7.2 [**B.7.3**]: Integrated CircuitsDesign Aids, Redundant Design

General Terms

Performance, Design

Keywords

Post-Silicon Optimization, Statistical Timing Analysis

1. INTRODUCTION

It is widely accepted today that it is imperative to incorporate the effects of process variations in nanometer-scale VLSI circuits. Broadly speaking, process variations can be classified into inter-die variations, from one chip to another, and intra-die variations, between different parts on the same die. Intra-die variations may be spatially correlated for some process parameters, such as the channel length L and the transistor width W, while other parameters such as the dopant concentration N_A and the oxide thickness T_{ox} show no such correlation structure.

These variations have driven a flurry of research in the area of statistical design to enable a transition from conventional corner-based static timing analysis (STA) to statistical static timing analysis (SSTA) which provides a probability distribution for the delay. Parameterized block-based SSTA methods [1, 2, 4] have emerged as effective frameworks that can incorporate spatial and structural correlations in the circuit, using a canonical form for the delay. The computational efficiency of these methods is made practical through a preprocessing step, proposed in [1], which has shown that Gaussian-distributed correlated variations can be orthogonalized using Principal Component Analysis (PCA).

The diagnostics provided by SSTA at the pre-silicon phase of design can be used to optimize the circuit for robust operation. However, pre-silicon optimizations alone are likely to be inadequate, particularly when the range of variation is large. Therefore, post-silicon testing and optimizations, which improve the probability that a chip meets its specifications after fabrication, form an important and complementary phase of design. Unlike pre-silicon analysis, which determines the range of performance (timing or power) variations over a large population of die, post-silicon analysis and test is typically directed toward determining the performance of an individual fabricated chip. It is inevitable that pre-silicon analysis, more generally applicable to the entire population of manufactured chips, will have a large standard deviation, and post-silicon optimizations typically require more information based on measurements specific to a manufactured die. Moreover, because tester time is prohibitively expensive, it is vital that performance estimations must be made on the basis of a small number of post-silicon measurements.

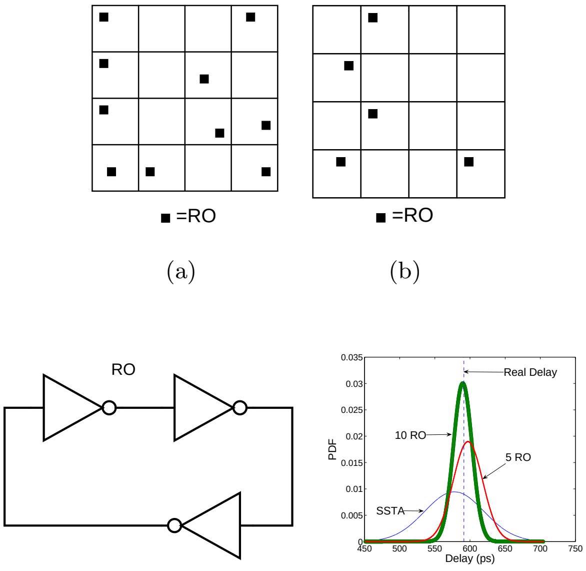
We present a general framework of post-silicon statistical delay prediction: the role of this step is seated between SSTA and full chip testing. Given the *original circuit* whose delay is to be estimated, the primary idea is to determine information from specific on-chip *test structures* to narrow 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, distributed over the area of the chip, to capture the variations of spatial correlated parameters over the die. To illustrate this 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 1(c) shows a sample test structure consisting of a 3-stage ring oscillator (RO). The data gathered from the test structures in Figures 1(a) and 1(b) are used in this paper to determine a new PDF (Probability Density Function) for the delay of the original circuit, conditioned on this data. This has significantly smaller variance than the result of SSTA, as is illustrated

*This research was supported in part by the NSF under award CCF-0205227 and by the SRC under contract 2007-TJ-1572.

[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.

DAC 2007, June 4-8, 2007, San Diego, CA, USA

in Figure 1(d); detailed experimental results are available in Section 6. When no test structures are used and no tests are performed, the PDF of the original circuit is the same as that computed by SSTA. 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, as shown in Figure 1(d). In other words, by using more or fewer test structures, the approach is *scalable* in terms of statistical confidence.



(a) (b)

(c) (d)

Figure 1: (a), (b): Two different placements of test structures under the grid spatial correlation model (c) An example test structure (ring oscillator) (d) Reduced-variance PDFs, obtained from statistical delay prediction, using data gathered from the test structures in (a) an (b)

A use case scenario for this method corresponds to a post-silicon optimization method such as Adaptive Body Bias (ABB) [5, 6, 7]. 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 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 [8].

2. PROBLEM FORMULATION

Intra-die variations for some parameters are spatially correlated[2]: this means that devices placed close together are more likely to have similar characteristics than those placed far away. Under spatial correlations, while one test structure may not reveal the characteristic of the whole chip, it can reveal some characteristics for the devices nearby. Therefore, our proposed statistical delay prediction approach uses a number of test structures, placed at different locations on chip, to provide diverse test data.

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* $\mathbf{d}_t = [d_{t,1} \quad d_{t,2} \quad \cdots \quad d_{t,n}]^T$ 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* $\mathbf{d}_r = [d_{r,1} \quad d_{r,2} \quad \cdots \quad d_{r,n}]^T$. 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 $f(d|\mathbf{d}_t = \mathbf{d}_r)$.

Ideally, given enough test structures, we can compute the delay of the original circuit with a great deal of confidence 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 test structures cannot be arbitrarily large. Another factor that limits the accuracy of these measurements is the fact that the variations in some parameters, such as T_{ox} and N_A, are widely believed to show no spatial correlation 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 a test circuit will not track variations in the original circuit. However, 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.

3. STATISTICAL DELAY PREDICTION

3.1 Statistical Static Timing Analysis Framework

We assume the process parameters, which affect both the original circuit and test structures, are Gaussian distributed. For the chip being considered, containing the original circuit and the test structures, it is assumed that there are m normalized underlying independent sources of variation for

[2] Inter-die variations can be considered to be a special case of intra-die variations, where the correlation region is the entire die.

spatially correlated variations (equivalent to Principal Components (PCs) in [1]), and these can be obtained by applying PCA to the covariance matrix of each spatially correlated process parameter variation. In addition, there may also be other independent uncorrelated sources of variation. Performing a parameterized SSTA technique such as [1], we can use a canonical form to represent the delay of the original circuit 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 2, and μ is the mean of d obtained from SSTA, and is also an approximation of its nominal value. The random variable p_i corresponds to the ith PC, and is normally distributed as $N(0,1)$; note that p_i and p_j for $i \neq j$ are uncorrelated by definition, due to the 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. For simplicity, we refer to $\mathbf{p} = [p_1 \quad p_2 \quad \cdots \quad p_m]^T \in \mathbf{R}^m$ as the *PC vector* and $\mathbf{a} = [a_1 \quad a_2 \quad \cdots \quad a_m]^T \in \mathbf{R}^m$ as the *coefficient vector* for the original circuit.

Equation (1) is general enough to incorporate both inter-die and intra-die variations. As is pointed out in [4], 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.

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

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

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

We define $\boldsymbol{\mu}_t = [\mu_{t,1} \quad \mu_{t,2} \quad \cdots \quad \mu_{t,n}]^T \in \mathbf{R}^n$ as the *mean vector*, $\mathbf{R}_t = [R_{t,1} \quad R_{t,2} \quad \cdots \quad R_{t,n}]^T \in \mathbf{R}^n$ 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 $\mathbf{A}_t = [\mathbf{a}_{t,1} \quad \mathbf{a}_{t,2} \quad \cdots \quad \mathbf{a}_{t,n}]$. 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 2.

To illustrate the procedure, we will first assume, in the remainder of this section and in Section 4, 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 5, and illustrate 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:

$$\sigma^2 = \mathbf{a}^T\mathbf{a} \text{ and } \boldsymbol{\Sigma}_t = \mathbf{A}_t^T\mathbf{A}_t. \tag{4}$$

3.2 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 delays, $\mathbf{d}_r$, measured from the test circuits. To achieve this, we first introduce a theorem below; a sketch of the proof can be found in [9].

THEOREM 3.1. *For 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,*

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

then the distribution of $\mathbf{X}_1$ *conditional on* $\mathbf{X}_2 = \mathbf{x}$ *is multivariate normal, and is given by*

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

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

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

To map our problem to the theorem, we call $\mathbf{X}_1$ the *original subspace*, and $\mathbf{X}_2$ the *test subspace*. By stacking d and $\mathbf{d}_t$ together, a new vector $\mathbf{d}_{all} = \begin{bmatrix} d\\ \mathbf{d}_t\end{bmatrix}$ 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:

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

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, as:

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

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

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

3.3 Interpretation of the Conditional PDF

We now 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 $\mathbf{d}_r$, 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 before testing, 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 $\boldsymbol{\Sigma}_t$. Therefore, the value of the conditional variance can be obtained by adjusting the value of $\mathbf{A}_t$, which is achieved by varying the number of test structures and their locations. Intuitively, this implies that the value of the conditional variance depends on how well the test structures are distributed, in the sense of capturing spatial correlations between variables.

Due to the nature of our problem, $\mathbf{A}_t^T \in \mathbf{R}^{n\times m}$, where n is usually less than m. Theorem 3.1 assumes that $\boldsymbol{\Sigma}_t$ is of full rank and has an inverse, which means $\mathbf{A}_t^T$ must have full row rank. Detailed discussion about the ranks of $\mathbf{A}_t^T$ and $\boldsymbol{\Sigma}_t$ can be found in Section 4. 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 $\boldsymbol{\Sigma}_t^{-1} = (\mathbf{A}_t^T\mathbf{A}_t)^{-1} = \mathbf{A}_t^{-1}(\mathbf{A}_t^T)^{-1}$ into Equation (8b), we get $\bar{\mu} = \mu + \mathbf{a}^T(\mathbf{A}_t^T)^{-1}(\mathbf{d}_r - \boldsymbol{\mu}_t)$. The term $(\mathbf{A}_t^T)^{-1}(\mathbf{d}_r - \boldsymbol{\mu}_t)$ is the solution of the linear equations

$$\mathbf{d}_t = \boldsymbol{\mu}_t + \mathbf{A}_t^T\mathbf{p} = \mathbf{d}_r \tag{9}$$

for $\mathbf{p}$. Therefore, in this case $\bar{\mu}$ is equival to d. And it is easy to derive that $\bar{\sigma}^2 = 0$. 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, referring to $\mathbf{p}$ as the original subspace, and $\mathbf{d}_t$ as the test subspace. From this, we obtain the conditional distribution of $\mathbf{p}$, using Theorem 3.1, as:

$$\text{PDF}(\mathbf{p}_{cond}) = \text{PDF}\left(\mathbf{p}|(\mathbf{d}_t = \mathbf{d}_r)\right) \sim N(\bar{\boldsymbol{\mu}}_{\mathbf{p}}, \bar{\boldsymbol{\Sigma}}_{\mathbf{p}}) \qquad (10a)$$

$$\bar{\boldsymbol{\mu}}_{\mathbf{p}} = \mathbf{A}_t\boldsymbol{\Sigma}_t^{-1}(\mathbf{d}_r - \boldsymbol{\mu}_t) \text{ and } \bar{\boldsymbol{\Sigma}}_{\mathbf{p}} = \mathbf{I} - \mathbf{A}_t\boldsymbol{\Sigma}_t^{-1}\mathbf{A}_t^T \qquad (10b)$$

where $\mathbf{I}$ represents the identity matrix, which is the unconditional covariance matrix of $\mathbf{p}$. The result (10) 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.

With the PDF of $\mathbf{p}_{cond}$, using basic statistical properties, we can get the same result as in (8). However, dividing the derivation into two steps, as we have done here, provides additional insight into the problem.

4. LOCALLY REDUNDANT GLOBALLY INSUFFICIENT TEST STRUCTURES

Because the PC vector is obtained by PCA, we do not know where these independent variation sources lie on the chip. As a result, it is possible that we place too many test structures which collectively capture only a small portion of PCs, with the coefficients of other PCs being zeros. That is to say, in some portion of the chip, the number of test structures exceeds 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 $\boldsymbol{\Sigma}_t$ would be rank deficient. The most trivial case is when two rows in $\mathbf{A}_t^T$ are identical. Under a grid based spatial correlation model, this corresponds to placing two test structures in the same grid, which is an obvious redundancy and can easily be avoided even before PCA. Therefore, we assume that such redundancies are removed, and that no two rows of $\mathbf{A}_t^T$ are identical.

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

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

where $\mathbf{B}_{11} \in \mathbf{R}^{s \times q}$, with $s > q$. Since we have prohibited two test structures 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 $\boldsymbol{\Sigma}_t$ also has a rank less than n and is singular. In this case, Equation (9) can be divided into two sets of equations:

$$\mathbf{B}_{11}\mathbf{p}_u = \mathbf{d}_{r,u} \qquad (12)$$

$$\mathbf{B}_{21}\mathbf{p}_u + \mathbf{B}_{22}\mathbf{p}_v = \mathbf{d}_{r,v} \qquad (13)$$

where $\mathbf{p}_u$, $\mathbf{p}_v$, $\mathbf{d}_{r,u}$, $\mathbf{d}_{r,v}$ are sub-vectors of the PC vector $\mathbf{p}$ and the result vector $\mathbf{d}_r$, correspondingly. Note that $\mathbf{B}_{11}$ is not square, and Equation (12) is an over-determined system. This can be solved in several ways, and we take the least-squares solution as its equivalence.

$$\bar{\mathbf{p}}_u = (\mathbf{B}_{11}^T\mathbf{B}_{11})^{-1}\mathbf{B}_{11}^T\mathbf{d}_{r,u} \qquad (14)$$

Under conditions (14) as well as (13), the conditional PDF of d can be computed using the same technique introduced in Section 3. The detailed derivation is omitted due to limited space.

5. SPATIALLY UNCORRELATED PARAMETERS

In Section 3, 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 3, 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 [4], with mean 0 and variance σ_R^2. Considering this effect, the variance of the original circuit is adjusted to be

$$\sigma'^2 = \mathbf{a}^T\mathbf{a} + \sigma_R^2. \qquad (15)$$

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

$$\boldsymbol{\Sigma}_t' = \mathbf{A}_t^T\mathbf{A}_t + diag[\sigma_{R_{t,1}}^2, \sigma_{R_{t,2}}^2, \cdots, \sigma_{R_{t,n}}^2]. \qquad (16)$$

The same kind of technique from Section 3 can still be applied. However, in this case, due to the diagonal matrix added to $\boldsymbol{\Sigma}_t$, $\bar{\sigma}$ 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.

6. EXPERIMENTAL RESULTS

The proposed post-silicon statistical delay prediction approach can be summarized as follows:

1. Perform SSTA on both the original circuit and the test structures, get μ, $\mathbf{a}$, $\boldsymbol{\mu}_t$, $\mathbf{A}_t$, and σ_R, $\sigma_{R_{t,1}} \sim \sigma_{R_{t,n}}$ if spatially uncorrelated parameters are considered.

2. After fabrication, test the delay of the test structures on chip to obtain $\mathbf{d}_r$.

3. Compute the conditional mean $\bar{\mu}$ and variance $\bar{\sigma}^2$ for the original circuit using the expressions in Equation (8).

We use the software package *MinnSSTA* [1] to perform SSTA. Because of the difficulty in accessing process data, we 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 3-stage ring oscillator (RO), as shown in Figure 1(c).

The grid model in [3] 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 $P \cdot G$. 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, two metal tiers (each corresponding to one horizontal and one vertical layer) are considered. Parameters of different metal tiers are different and are not correlated, but the two metal layers within a tier are taken to have similar characteristics. Table 1 lists the levels of parameter variations assumed in this work. The 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; the only exception is N_A, where 30% of the variations are assumed to be due to intra-die effects. We assume this variation model is accurate in our simulation. In practice the model should be tailered according to manufacturing data.

Table 1: Parameters used in the experiments

	L (nm)	W (nm)	W_{int} (nm)	T_{int} (nm)	H_{ILD} (nm)	N_A (10^{17}cm^{-3}) N/PMOS
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 3 to calculate the delay of the original circuit.

The result is shown as a scatter plot in Figure 2. 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.

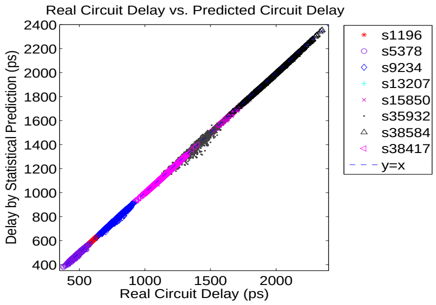


Figure 2: The scatter plot: real circuit delay vs. predicted circuit delay

The precise test error for each benchmark is listed in Table 2. 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 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.

Table 2: Test errors considering L

Benchmark	Average Error	Maximum Deviation (% of mean)	Error at Maximum Deviation
s1196	0.21%	19.0%	0.09%
s5378	0.73%	25.7%	0.02%
s9234	0.48%	22.7%	0.76%
s13207	0.20%	28.0%	0.16%
s15850	0.27%	24.9%	0.13%
s35932	1.52%	26.1%	1.47%
s38584	0.17%	21.4%	0.37%
s38417	0.18%	22.2%	0.16%

Note that in theory, according to the discussion in Section 3, 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 such as 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.

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 whole 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 17.3%, while our approach has a maximum error of 8.46%, an improvement of more than 50%.

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 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 get 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 put 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 3 for each benchmark circuit. It

[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 3: Prediction results with insufficient number of test structures (considering L): Case 1 and Case 2 are distinguished by the number of ring oscillators (RO) available for each circuit

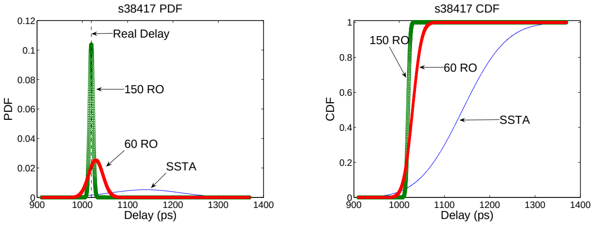
Benchmark			SSTA Results		Case 1			Case 2		
Name	#Cells	#Grids	μ(ps)	σ(ps)	#RO	Avg. $\bar{\sigma}(ps)$	Avg. $\sigma_{reduction}$	#RO	Avg. $\bar{\sigma}$(ps)	Avg. $\sigma_{reduction}$
s1196	547	16	577.06	35.32	10	6.48	81.64%	5	11.97	66.1%
s5378	2958	16	475.97	29.84	10	5.96	80.02%	5	10.77	63.9%
s9234	5825	16	775.36	51.51	10	9.50	81.55%	5	18.85	63.4%
s13207	8260	256	1399.8	92.81	150	9.63	89.62%	60	18.56	80.0%
s15850	10369	256	1573.7	100.48	150	8.25	91.79%	60	16.88	83.2%
s35932	17793	256	1359.5	82.17	150	11.08	86.52%	60	27.69	66.3%
s38584	20705	256	1994.0	120.83	150	16.54	86.31%	60	29.96	75.2%
s38417	23815	256	1139.8	76.38	150	9.40	87.69%	60	17.87	76.6%

Table 4: Prediction results considering all parameter variations: Case I, Case II and Case III are distinguished by different number of ring oscillators(RO) available for each circuit

Benchmark		SSTA Results		Case I			Case II			Case III		
Name	#Grids	μ	σ	#RO	$\bar{\sigma}$	$\sigma_{reduction}$	#RO	Avg. $\bar{\sigma}$	Avg.	#RO	Avg. $\bar{\sigma}$	Avg.
		(ps)	(ps)		(ps)			(ps)	$\sigma_{\text{reduction}}$		(ps)	$\sigma_{\text{reduction}}$
s1196	16	577.68	42.15	16	9.06	78.5%	10	11.63	72.4%	5	15.85	62.4%
s5378	16	477.04	34.32	16	4.36	87.3%	10	7.07	79.4%	5	11.12	67.6%
s9234	16	777.63	58.28	16	6.64	88.6%	10	12.18	79.1%	5	19.58	66.4%
s13207	256	1403.03	106.68	256	16.3	84.7%	150	19.4	81.8%	60	25.4	76.2%
s15850	256	1578.86	115.67	256	15.15	86.9%	150	17.58	84.8%	60	22.56	80.5%
s35932	256	1372.45	96.35	256	18.50	80.8%	150	22.26	76.9%	60	27.46	71.5%
s38584	256	2006.76	143.64	256	29.16	79.7%	150	34.76	75.8%	60	43.09	70.0%
s38417	256	1144.47	88.80	256	16.16	81.8%	150	19.36	78.2%	60	25.40	71.4%

is observed that there is a trade-off between test structure overhead and $\sigma_{reduction}$.

Figure 3 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 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 be very confident in saying that it can meet a 1050 ps delay requirement, and using 60 test structures we can be confident in saying that it can meet a 1100 ps delay requirement.



(a) (b)

Figure 3: PDF and CDF with insufficient number of test structures for circuit s38417 (considering L)

Finally, 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 results of applying it to real circuits under the full set of parameter variations listed in Table 1. 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 $\bar{\sigma}$ and average $\sigma_{reduction}$ for Case II and Case III obtained from the post-silicon statistical delay prediction are listed in Table 4. The distribution plot for this set of experiment is similar to that in Figure 3, and the conditional PDFs of one particular sample of the circuit s1196 for Case II and Case III are shown in Section 1 as Figure 1(d), with the SSTA PDF as a comparison. Note that the conditional PDF obtained by our approach would be even sharper for Case I.

7. REFERENCES

[1] H. Chang and S. S. Sapatnekar. Statistical Timing Analysis Considering Spatial Correlations using a Single Pert-Like Traversal . In *Proc. IEEE/ACM ICCAD*, pp. 621–625, 2003.

[2] C. Visweswariah *et al.* First-Order Incremental Block-Based Statistical Timing Analysis . In *Proc. ACM/IEEE DAC*, pp. 331–336, June 2004.

[3] A. Agarwal *et al.* Path-Based Statistical Timing AnalysisConsidering Inter- and Intra-die Correlations . In *Proc. TAU*, pp. 16–21, 2002.

[4] H. Chang and S. S. Sapatnekar. Statistical Timing Analysis under Spatial Correlations . In *IEEE Trans. on CAD*, vol. 24, pp. 1467–1482, Sep. 2005.

[5] J. Tschanz *et al.* Adaptive Body Bias for Reducing Impacts of Die-to-Die and Within-Die Parameter Variations onMicroprocessor Frequency and Leakage . In *IEEE Journal of Solid-State Circuits*, vol. 37, pp. 1396–1402, Nov. 2002.

[6] J. Tschanz *et al.* Effectiveness of Adaptive Supply Voltage and Body Bias for Reducing the Impact of Parameter Variations in Low Power and High Performance Microprocessors . In *IEEE Journal of Solid-State Circuits*, vol. 38, pp. 826–829, May 2003.

[7] J. Tschanz *et al.* Adaptive Circuit Techniques to Minimize Variation Impacts on Microprocessor Performance and Power . In *Proc. IEEE ISCAS*, pp. 23–26, May 2005.

[8] S. V. Kumar, C. H. Kim, and S. S. Sapatnekar. Mathematically Assisted Adaptive Body Bias (ABB) for Temperature Compensation in Gigascale LSI Systems . In *Proc. ASP DAC*, pp. 559–564, Jan. 2006.

[9] R. A. Johnson, D. W. Wichern. *Applied Multi-variate Analysis (3rd ed.)*, Prentice Hall, New Jersey, 1992.