Post-Silicon Timing Characterization by Compressed Sensing

ABSTRACT

We address post-silicon timing characterization of the unique gate delays and their distributions on each manufactured IC. Our proposed approach is based upon the new theory of compressed sensing that enables efficient sampling and reconstruction of sparse signals using significantly fewer measurements than previously thought possible. The first step in performing timing measurements is to find the sensitizable paths via traditional testing methods. Next, we show that variations are sparse in wavelet domain. Then, using the compressed sensing theory, our method estimates the delay distributions by a small number of timing measurements. We discuss how the post-silicon characterization method can enable a range of new and emerging applications including improved simulation models, post-silicon optimization, and IC fingerprinting. Experimental results on benchmark circuits show that using compressed sensing theory can characterize the post-silicon variations with a mean accurately of 95% in the pertinent sparse basis.

1. INTRODUCTION

Modern integrated circuits are variable and complex. Continuous CMOS scaling has made possible integration of billions of gates into a single multi-layer chip. Scaling to the physical device limitations and mask imprecisions have created nondeterminism in the chip's characteristics. In the new regime, traditional models, design, and test methods have a limited effectiveness.

Process variations may be die-to-die (inter-die), or within one die (intra-die). A number of process parameters such as transistor width/length maybe spatially correlated, whereas other parameters, e.g., oxide thickness, may not show such structural relationships. Furthermore, with miniaturization of devices beyond 65nm, the impact of intra-die variations and the spatial correlations are becoming more prominent [1]. Several key areas have been impacted. For example the number of critical paths is increasing with variations, rendering the traditional test methodologies based on a few critical paths inexpressive [2]. As another example, in *statistical static timing analysis (SSTA)*, instead of the single valued delays utilized in traditional models, delay probability distributions and their correlations were used [3, 4].

The models and analysis produced by SSTA are pre-silicon. They are utilized for determining the impact of component variations on the circuit response to optimize the design to be robust to variations [5]. Recently, a post-silicon timing analysis of chips was proposed [6]. The method works by collecting data from a few on-chip test points (e.g., via ring oscillators), and integrating this data with the SSTA models to form the chip-specific distribution of the delays.

Our objective is to perform post-silicon timing characterization of each specific chip. Our method uses the revolutionary theory of *compressed sensing* [7, 8] and the set of the sensitizable paths known from the testing phase to perform post-silicon delay modeling with very few measurements. Traditional sampling methods were based on Nyquist density which states that a signal must be sampled at a rate at least twice its highest frequency. Compressed sensing theory has shown that it is possible to reconstruct signals that are sampled by a rate far smaller than the Nyquist theorem, such that the pertinent signals can be most often reconstructed accurately, and even sometimes exactly. We demonstrate how this method can be used for testing the chips, such that the narrow and specific distribution of the chip's timing can be estimated by a few number of post-silicon measurements. Compressed sensing exploits the sparsity of the distribution matrix [4], to translate the compressed analog sample measurements to a reconstructed information. Our contributions are as follows:

- We introduce the first post-silicon timing characterization method that is based on the compressed sensing. Our method does not add on-chip test structures and relies on external tests to keep the number of measurements low.
- We create a systematic method for exploiting the sparsity of the timing variations for post-silicon characterization.
- The approach does not impose any restrictions on the shape of the process parameter distribution, except for sparsity which has been extensively used in the modeling and validation of timing variations [3, 4].
- We present modifications to the original compressed sensing framework that is based upon regular grid-based sampling, so it can include the irregularities of the spatial placement and thus spatial correlations of the gate delays.
- The method exploits the correlations to approximate the timing variations of the gates that are unobservable and uncontrollable because of their placement on insensitizable paths. The key for post-silicon gate characterization is the delay variations' sparsity.
- We demonstrate how the extracted features are insensitive to the selected inputs. This can be explained by the nature of the compressed sensing theory, that can find accurate solutions for random independent measurement inputs.
- Knowledge of the within-die correlation models will be readily available via our method, producing a feedback from manufacturing to SSTA. In turn, it may increase the performance yield by lessening the timing models pessimism.
- We discuss a number of emerging applications that are enabled by the proposed method.

The remainder of the paper is as follows. Sections 2 and 3 discuss related works and preliminaries. We introduce variation estimation by delay measurements in Section 4. In Section 5, we use sparsity of the variations in the wavelet domain to recover variations with a small number of delay measurements. Applications and Evaluation results are presented in Sections 6 and 7. We conclude in Section 8.

2. RELATED WORK

SSTA is a method that performs pre-silicon analysis of timing variations of the full chip under the assumption of the process parameter uncertainty. The proposed models range from simple to sophisticated ones, including nonparametric and higher order models [3,4,9–11]. A recent study compared compared 5 different SSTAs and their associated correlation models on real chip measurement data [1]. A group of Intel researchers studied the benefits of changing from STA to SSTA for optimizations that target gate sizing [5]. They concluded that under the current variation models, the power reduced only by 2%. They emphasize that to achieve 4-6% power reduction, process variation need to increase by a factor of 2x and 4x respectively.

The recent post-silicon statistical approach predicts the delays by collecting data from a small number of on-chip test sensors and then combining this information to find the narrow and die-based timing distributions [6]. They report that their method extracts the variability-based distribution with 83.5% smaller on average than the SSTA results. While our approach also concentrates on post-silicon optimizations, it is not based on insertion of additional sensors/circuitry.

A suit of new post-silicon tests have been developed by the testing community with the objective of integrating the impact of variations in traditional timing and functional tests. A number of methods have utilized the long-known relation that under certain mild assumptions, the delay of each path in the circuit can be expressed as a linear combination of others [2,12–14]. In this work, we use the linear dependence of the paths and the functionally sensitizable paths that are extracted by testing methods.

In the past four year, the compressed sensing theory has emerged [7,8,15]. According to the class Shannon/Nyquist theory, the number of required samples for a signal to be reconstructed without error - the length of the shortest interval containing the support of the pertinent signal. Compressive sensing has shown that compressible images and signals can be reconstructed from far fewer measurement samples. The new theory suggests that the analog data (say a scene) is readily captured into its compressed form. The class of compressible signals are referred to by sparse.

3. PRELIMINARIES

3.1 Variation Model and Delay Model

Variations in an IC are categorized as systematic variations and random variations [16]. Systematic variations refer to the variations caused by imperfectness of fabrication tools. Since the properties of fabrication tools are known, systematic variations are deterministic and they are known beforehand. Random variations include inter-die and intra-die variations. Inter-die variations represent the variation among various dies in a wafer and intra-die variations represent variation among different devices in a die. Note that systematic variation can also divided into inter-die and intra-die variations. Thus, total variation, ψ_u^{total}, in a gate g_u will be [16]

$$\psi_u^{\text{total}} = \psi_u^{\text{inter}} + \psi_u^{\text{intra}} + F_u\beta \tag{1}$$

Where ψ_u^{inter} and ψ_u^{intra} represent inter-die and intra-die variation, respectively. ψ_u^{intra} is a multivariate Gaussian random vector. $F_u\beta$ models systematic variations ; if (x_u, y_u) is the location of the gate g_u on the IC, then $F = [1, x_u, y_u]^T$ and β is a 3×1 constant vector.

Transition delay is usually modeled as a linear function of transistor feature size variations [4,17,18]. For example, consider a NAND2 gate that one of its inputs is 1 and the other input, at time $t = 0$, transits from 0 to 1. Because of propagation delay of the NAND2 gate, output transit from 1 to 0 at time $t = d_r$. When there are variation in the feature size of the transistors, rising propagation delay, d_r, varies among different NAND2 gates in the IC. i.e. [17]

$$d_r(\psi_u^{\text{total}}) = d_r^0 + a\psi_u^{\text{total}} \tag{2}$$

where a is a constant.

Note that, even if we model the propagation delay quadratic (or higher order) [19], we can use the same approach by assuming new variables for higher order parameters.

3.2 Sensitizable Paths

A path in an IC is defined as a sequence of logic gates from an input of the IC to one of its output pins. To find propagation delay in a path, one should find an appropriate input vector to the IC. If such an input vector exists, the path is called *sensitizable*; otherwise it is called *unsensitizable.*

To find sensitizable paths, we use the path selection method that is introduced by Murakami et al. [20]. This method is based on finding inconsistent transitions in paths. The method provides a list of potentially sensitizable paths. While a potentially sensitizable path might not be sensitizable, they show about 99% of potentially sensitizable paths are sensitizable. Without loss of generality, we use potentially sensitizable path in this paper; one can exhaustively run ATPG algorithms [21] to find list of all sensitizable paths.

3.3 Compressed Sensing

Compressed Sensing is a recently emerging signal acquisition method that exploits sparse signal models to reduce the signal acquisition burden [7,15]. Specifically, we assume that the signal of interest is a K-sparse vector $\mathbf{x}$ in an N-dimensional space, i.e., that it only has K non-zero components. Using compressed sensing we can sample and reconstruct this vector by acquiring only $M = O(K \log(N/K))$ linear measurements.

$$\mathbf{p} = A\mathbf{x} + e, \tag{3}$$

where A denotes the measurement matrix, of dimension $M \times N$, $\mathbf{p}$ denotes the M-dimensional measurement vector, and e denotes the measurement noise.

Despite the dimensionality reduction and the rank deficiency of A, we can reconstruct the sparse vector of interest, $\mathbf{x}$ from the measurement vector $\mathbf{p}$ using the following convex optimization:

$$\min ||\mathbf{x}||_1 + \lambda||\mathbf{p} - A\mathbf{x}||_2^2, \tag{4}$$

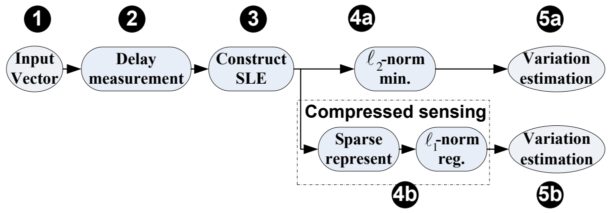

Figure 1: Global flow.

in which λ is a parameter chosen according to the noise variance and $\|\mathbf{x}\|_p = (\sum_{i=1}^{N}|x_i|)^{\frac{1}{p}}$. If the measurement matrix A satisfies certain conditions, we can show that the reconstruction using Equation 4 is exact [7].

The compressed sensing model is robust even when the acquired vector $\mathbf{x}$ is approximately sparse, often referred to as compressible. A vector is compressible if it has very few (say K) coefficients with large magnitude and the remaining coefficients are approximately 0. Compressible vectors can be approximated very well using the best K-term approximation, i.e. using the K most significant coefficients and setting the remaining to 0. An often used class of compressible vectors are the vectors lying in a weak ℓ_p ball, where $p < 1$. These vectors have the property that their sorted coefficients follow a power law decay:

$$|\mathbf{x}_{(i)}| \leq ri^{-\frac{1}{p}}, 1 \leq i \leq N, \tag{5}$$

in which $x_{(i)}$ is i-th largest coefficient of the vector $\mathbf{x}$ [7].

In most practical applications, such as in this paper, a vector is not compressible in the canonical domain. Usually, in practice, a sparsity inducing basis W is necessary to expose the sparsity. The theory accommodates this case using the basis expansion

$$\mathbf{s} = W\mathbf{x}, \tag{6}$$

in which case W is the sparsity inducing transform, and the basis expansion vector $\mathbf{s}$ is sparse instead of the vector of interest $\mathbf{x}$. In this case Equation 3 becomes

$$\mathbf{p} = AW^{-1}\mathbf{s} + e. \tag{7}$$

This is the same formulation as Equations 3 and 4, with only a change of variables. We now aim to recover a sparse representation $\mathbf{s}$ from the measurements $\mathbf{y}$, which are acquired with a measurement matrix AW^{-1}. The signal is subsequently recovered from the representation using Equation 6.

3.4 Global Flow

Figure 1 shows the global flow of the work. At the first step, we feed the circuit with a number of input vectors that provide sensitizable paths. In step 2, propagation delay is measured for every sensitizable path. Based on the measured propagation delays, we construction a System of Linear Equations (SLE) with gate variations as its unknown parameters. Then, we estimate variations by two methods (4a and 4b). The first method is based on the traditional ℓ_2-norm minimization (4a.) In the second method, we show sparsity of the variations in wavelet domain and we use compressed sensing (ℓ_1-norm regularization) to estimation variation more efficiently.

4. DELAY ESTIMATION BY ℓ_2-NORM MINIMIZATION

In this section, we propose a method for post-silicon gate delay estimation by measuring the input/output path delays. First, we measure the signal propagation delays of a number of sensitizable paths. Then, based on the measured delays, we construct linear equations with the scaling factors of gate delays (defined in Section 3.1) as the unknown parameters. Finally, using the linear equations, we estimate the gate variations by solving for the scaling factors. In Section 5, we utilize the variations in spatial correlations to improve the scaling factor estimations.

An example of path delay analysis is shown in Figure 2. Lines labeled by a, b, c, and d are the circuit's primary inputs and the line n is the circuit's primary output. We want to sensitize the delay of the highlighted path, P_1: (a-g_1-z-e-g_3-f-g_4-s-g_6-k-g_7-n). This is because as we discussed in Section 3, we can only find the delays on the sensitizable paths. Thus, we need to find an input vector that guarantees a transition in input a that would propagate through the path. Let us assume a rising transition in a (input a transits from 0 to 1). To allow propagation through the gate g_1, we need to set b to be equal to 0. Then, there would be a falling ($1 \rightarrow 0$) and a rising ($0 \rightarrow 1$) transition in lines e and f, respectively. If g is equal to 1 and m is equal to 0, then the rising transition propagates in the lines s, k and n. To guarantee that g is equal to 1 and m is equal to 0, we just need to set the input $c = 0$.

The input assignments above allow the transition in input a to propagate through the path P_1 :a-g_1-z-e-g_3-f-g_4-s-g_6-k-g_7-n. Thus using the delay bounding method introduced in [13], one can measure the total delay of the underlying path. i.e., we can measure the time difference between the transitions in line a and in line n. Let us denote the total delay of the path P_1 for the rising transition by $d_r(P_1)$.

The total path delay is an additive composition of the delays of its elements. For example, delay of the path P_1 can be written as the summation of the delays in line a, gate g_1, line k, line e, gate g_3, and so on. i.e.,

$$\begin{aligned} d_r(P_1) &= d(a) + d_r(g_1) + d(k) + d(e\ ) + d(e\ ) \\ &+ d_f(g_3) + d(f\ ) + d_r(g_4) + d(s\ ) + d_f(g_6) \\ &+ d(k\ ) + d_r(g_7) + d(n\ ), \end{aligned} \tag{8}$$

where $d(x)$ is the delay of the line x; $d_r(g_i)$ and $d_f(g_i)$ are rising and falling delay of the gate g_i, respectively.

In this paper we assume interconnect delays (line delays) are zero. This assumption is just to keep the clarity of the presentation and the approach. The proposed method can be easily extended to cases with non-zero interconnect delays. Note that, it maybe the case that variations in the interconnects have a separate statistical representation. In such scenarios, one may consider compressed sensing methods that address the summation of two distinct distributions in one framework [15]. Assuming zero interconnect delays, equation 8 reduces to:

$$d_r(P_1) = d_r(g_1) + d_f(g_3) + d_r(g_4) + d_f(g_6) + d_r(g_7). \tag{9}$$

In Section 3, we illustrated that because of the process variation, delays of the gates deviate from their nominal values, i.e. [17],

$$d_r(g_i) = d_r^{\text{nominal}}(g_i) + \xi_{r,g_i} l_{g_i}, \tag{10}$$

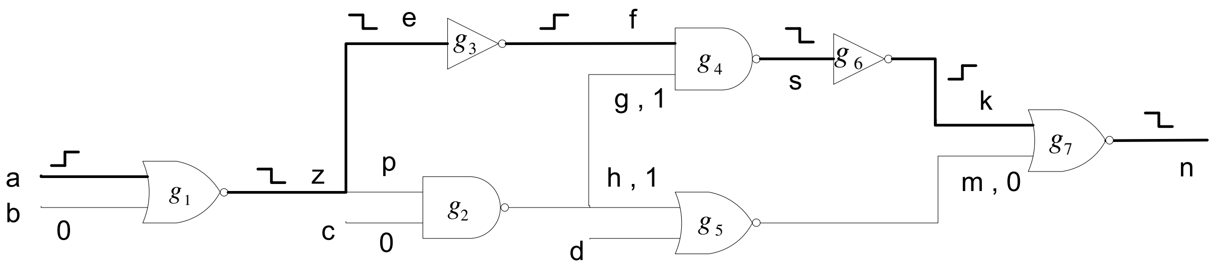

Figure 2: A sensitizable path from input to the output. Inputs to the circuit are set such that a rising (falling) transition in input a can propagate to the output b.

Gate	Rising (pS/μm)	Falling (pS/μm)
Inverter	86.9	40.77
NAND2	176.9	507.7
NOR2	95.4	1106.2

Table 1: Transition propagation rate for different gates. The rising and the falling transitions do not enforce the same delay rates.

where $d_r^{\text{nominal}}(g_i)$ is the nominal delay for rising transition and l_{g_i} is the scaling factor of the variation for the gate g_i; and ξ_{r,g_i} is a constant coefficient. Table 4 shows the constant coefficient for NAND2 gate. Similarly for the falling transition,

$$d_f(g_i) = d_f^{\text{nominal}}(g_i) + \xi_{f,g_i} l_{g_i}. \qquad (11)$$

Thus, Equation 9 becomes

$$\begin{aligned} d_r(P_1) &= d_r^{\text{nominal}}(g_1) + \xi_{r,g_1} l_{g_1} \\ &+ d_f^{\text{nominal}}(g_3) + \xi_{f,g_3} l_{g_3} \\ &+ d_r^{\text{nominal}}(g_4) + \xi_{r,g_4} l_{g_4} \\ &+ d_f^{\text{nominal}}(g_6) + \xi_{f,g_6} l_{g_6} \\ &+ d_f^{\text{nominal}}(g_7) + \xi_{r,g_7} l_{g_7}, \end{aligned} \qquad (12)$$

or

$$\xi_{r,g_1} l_{g_1} + \xi_{f,g_3} l_{g_3} + \xi_{r,g_4} l_{g_4} + \xi_{f,g_6} l_{g_6} + \xi_{r,g_7} l_{g_7} = b_{P_1}$$

$$\begin{aligned} b_{P_1} &= d_r(P_1) - d_r^{\text{nominal}}(g_1) - d_f^{\text{nominal}}(g_3) \\ &- d_r^{\text{nominal}}(g_4) - d_f^{\text{nominal}}(g_6) - d_f^{\text{nominal}}(g_7) \end{aligned}$$

b_{P_1} is a constant. Thus, each sensitizable path in the circuit leads to a linear relation among the variation elements, l_{g_i}. The falling and rising coefficient (ξ_{f,g_i} and ξ_{r,g_i}) are known and our goal is to estimate the variations, l_{g_i}.

Assume that $P_1, P_2 \ldots P_M$ are M sensitizable paths in a general combinational circuit C with N gates. For each path P_j, if it is stimulated by a rising transition,

$$\sum_{i=1}^{N} \alpha_{P_j}(i) \xi_{\lambda^r(P_j,g_i),g_i} l_{g_i} = b_j^r \qquad (13)$$

where

$$\alpha_{P_j}(i) = \begin{cases} 1 & \text{if } g_i \text{ belongs to the path } P_j; \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\lambda^r(P_j, i) = \begin{cases} f & \text{if } g_i \text{ has a falling transition when path } P_j \\ & \text{is stimulated by a rising transition;} \\ r & \text{otherwise.} \end{cases}$$

Similarly for a falling transition,

$$\sum_{i=1}^{N} \alpha_{P_j}(i) \xi_{\lambda^f(P_j,g_i),g_i} l_{g_i} = b_j^f \qquad (14)$$

where

$$\lambda^f(P_j, i) = \begin{cases} f & \text{if } g_i \text{ has a falling transition when path } P_j \\ & \text{is stimulated by a falling transition;} \\ r & \text{otherwise.} \end{cases}$$

To write Equations 13 and 14 in a compact form, we define matrix A and measurement vector $\mathbf{b}$ and variation vector $\mathbf{l}$ as follows.

$$A = \begin{pmatrix} \alpha_{P_1}(1)\xi_{\lambda^r(P_1,g_1),g_1} & \cdots & \alpha_{P_1}(N)\xi_{\lambda^r(P_1,g_N),g_N} \\ \alpha_{P_2}(1)\xi_{\lambda^r(P_2,g_1),g_1} & \cdots & \alpha_{P_2}(N)\xi_{\lambda^r(P_2,g_N),g_N} \\ \vdots & & \vdots \\ \alpha_{P_M}(1)\xi_{\lambda^r(P_M,g_1),g_1} & \cdots & \alpha_{P_M}(N)\xi_{\lambda^r(P_M,g_N),g_N} \\ \alpha_{P_1}(1)\xi_{\lambda^f(P_1,g_1),g_1} & \cdots & \alpha_{P_1}(N)\xi_{\lambda^f(P_1,g_N),g_N} \\ \alpha_{P_2}(1)\xi_{\lambda^f(P_2,g_1),g_1} & \cdots & \alpha_{P_2}(N)\xi_{\lambda^f(P_2,g_N),g_N} \\ \vdots & & \vdots \\ \alpha_{P_M}(1)\xi_{\lambda^f(P_M,g_1),g_1} & \cdots & \alpha_{P_M}(N)\xi_{\lambda^f(P_M,g_N),g_N} \end{pmatrix},$$

$$\mathbf{b} = (b_1^r, b_2^r, \ldots b_M^r, b_1^f, b_2^f, \ldots b_M^f)^T,$$

and

$$\mathbf{l} = (l_1, l_2 \ldots l_N)^T.$$

Finally, we use following optimization to estimation variation, $\mathbf{l}$.

$$\min ||A\mathbf{l} - \mathbf{b}||_2^2. \qquad (15)$$

We call this method ℓ_2 minimization method.

Note that it may not be possible to find the variation of all gates by this method. For example in Figure 2, if we want to find another sensitizable path that includes g_4, we should fix $f = 1$ (none-controlling value) causing $e = 0$ and $g = 1$. Thus, the transition cannot propagate on the line g and path P_0 is the only path that includes the gates g_3, g_4 and g_6. As a result, there is at most two equations (falling and rising) that includes variation of the gates g_3, g_4 and g_6; it is impossible to find the variation of the three gates separately. We refer to such cases as ambiguous gates.

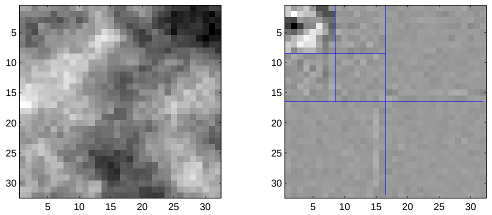

Figure 3: Left: Spatial variation in a typical IC. Right: wavelet transform of the variation. Because of the spatial correlation the variation is sparse in the wavelet domain.

5. DELAY ESTIMATION USING COMPRESSED SENSING

Section 4 presents a system of linear equations to estimate variations of the gates. However, the optimization problem in Equation 15 does not consider the spatial correlation of the delay variations. Incorporating the spatial correlation in the model significantly improve the results and allows resolving the ambiguities described in the previous section. This section incorporates sparsity in the wavelet domain as a model for the spatial correlation of the timing variation. Thus we can use compressed sensing theory to measure and estimate the variation.

5.1 Sparse Representation of Variations

To describe the form of spatial correlation of the variations we use a wavelet basis expansion. Wavelet basis expansions have a number of significant advantages that make them suitable for the problem at hand [22]. Specifically, wavelet expansions are very efficient to compute using well-studied fast algorithms. Furthermore they are very good in sparsely describing smooth functions, such as spatial correlations. Figure 3 demonstrates the effectiveness of the wavelet transform in representing spatial variation. The left side of the figure is the image plot of the variation in a typical IC, generated using the Gaussian model in [16]. The spatial correlation is evident in the figure. The right side of the figure represents the wavelet transform of the left hand side. Most of the transform coefficients are zero. Only the top-left part of the figure has a dense amount of significant non-zero elements.

Figure 4 presents the decay rate of the wavelet coefficients for a number of different wavelet transforms. A transform appropriate for compressed sensing should have a fast decay rate. The faster the decay, the sparser the signal under this transform, and the fewer the measurements necessary to acquire the variation vector. The figure demonstrates that the (3,5) Biorthogonal wavelet basis best describes the spatial variations. We use this wavelet basis for the remainder of this paper.

5.2 Gates on the Regular Grids

The derivations in this section assume that all the gates are located on a regular grid. This assumption is relaxed in Section 5.3 where the general case is considered.

When gates are located on a regular grid, the two-dimensional wavelet transform of the variation, $\mathbf{s}$, can be expressed as the product of the variation vector, $\mathbf{l}$, with the wavelet transform matrix W.

$$\mathbf{s} = W\mathbf{l}. \tag{16}$$

As discussed in Section 5.1, $\mathbf{s}$ is assumed sparse because of the spatial correlation in the variation. We enforce the sparsity prior by regularizing Equation 15 using the ℓ_1 norm of $\mathbf{s}$, as described in Section 3.3:

$$\min ||A\mathbf{l} - \mathbf{b}||_2^2 + \lambda||\mathbf{s}||_1 \tag{17}$$

or, equivalently,

$$\min ||AW^{-1}\mathbf{s} - \mathbf{b}||_2^2 + \lambda||\mathbf{s}||_1, \tag{18}$$

where λ is the regularization coefficient. Sparsity of the variations wavelet transformation,$\mathbf{s}$, provide a new piece of information; Equations 17 and 18 essentially add the knowledge of sparsity to the optimization. We call this method ℓ_1 regularization method.

5.3 Gates on the Irregular Grids

In practice, gates are not placed on a regular layout grid. Thus, in this section, we extend the proposed delay characterization method to irregular grids.

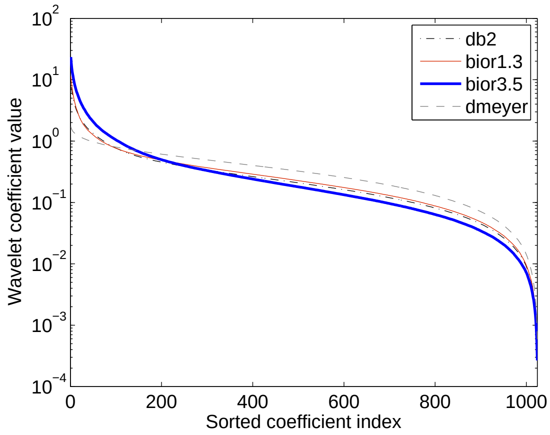

Figure 4: Sorted wavelet coefficients for different bases. bio3.5 bases results in the most sparse representation.

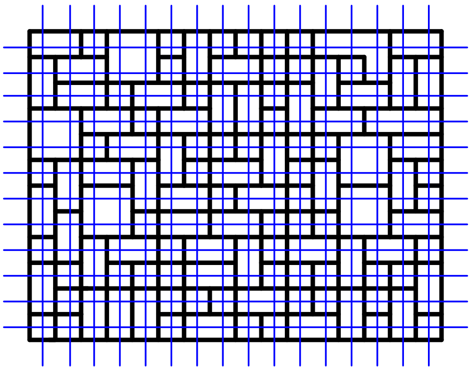

Figure 5: When gates are placed on irregular grids, we cover the circuit with a fine grid and map the gates to the point on the fine grid.

Figure 5 shows an example of an IC in which gates are placed on an irregular grid. To address the irregular placement, we cover the IC with fine regular grids. Then, using procedure 1, each gate is assigned to a point on the regular grid. At the first step of Procedure 1, we label all the regular grid points *unmarked.* It means that none of the regular points is assigned to any gate. In the second step, for every gate, we find its closest regular point that is *unmarked.* Finally, in 2.c, to prevent multiple selection, we mark the selected regular grid.

In procedure 1, each gate is uniquely assigned to its closest regular grid that is not assigned to any other gate.

PROCEDURE 1
Mapping from irregular gates to fine regular grids

```
(1) Set all the regular grid points unmarked
(2) for all gates, $g_i$
  a. $p$ = the closest grid point to the gates that is unmarked
  b. assign gate $g_i$ to $p$
  c. it Mark regular grid point $p$
```

Then, we assign auxiliary variables to the points in the fine grid that are not assigned to any gates. We also modify the measurement matrix A to be consistent with the fine regular grids. i.e., for each auxiliary variable, we add an appropriate zero column to the matrix A. Since the coefficients of auxiliary variables in the measurement matrix are zero, they do not affect the optimization.

6. APPLICATIONS

The proposed timing characterization method is effective, inexpensive, and fast. A range of technical applications can profit from the extracted post-silicon delay characteristics. The emerging applications includes:

(1) Post-silicon optimization. Fast and noninvasive IC characterization, enables application of chip-specific optimization, e.g., post-silicon adaptive body bias [6] [23] [24].

(2) Improving simulations. The post-silicon models can be integrated within the simulation platforms to enable more efficient and accurate simulations.

(3) Improving SSTA methods. The aggregate statistics gathered from post-silicon characterization can also be used to enhance the quality of the pre-silicon models, such as SSTA.

(4) Manufacturing process characterization. The processes and technologies of the state-of-the-art silicon manufacturing facilities are considered classified information that are not typically available to the users. The new noninvasive characterization method could make accurate post-silicon estimation of a number of important process parameters.

(5) IC identification. Since the variations are unique and unclonable for each manufactured IC, they can be used as the chip's ID/fingerprint.

7. EVALUATION RESULTS

In this section, we evaluate the performance of the proposed variation estimation methods on the MCNC benchmarks.

7.1 Measurement Matrix and Estimation in Subspaces

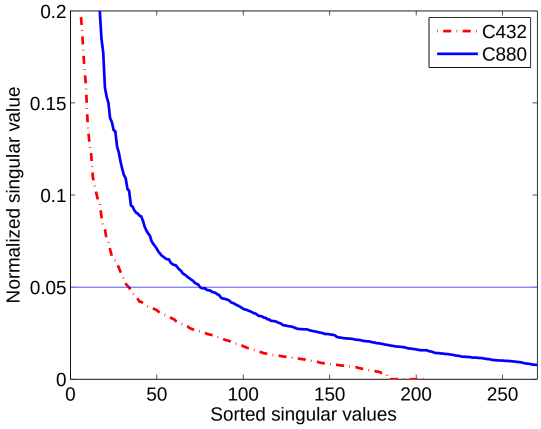

Figure 6: Singular values of the measurement matrices decay very fast.

As mentioned in Section 4, due to the existence of ambiguities (path dependencies), it may not be possible to find variations of all the gates in the circuit. Thus, the measurement matrix, A, is not necessary a full-rank matrix. Most often the measurement matrix is ill-conditioned and its singular value decay very fast. Figure 6 shows singular values of the measurement matrix for C880 and C432 circuit. The singular values are normalized to have the maximum value equal to 1. The singular values decay to 5% of the first one after 34-th (C432) and 75-th (C880) singular values. Note that C432 and C880 have 206 and 353 gates, respectively.

Hence, it is not possible to find the variations of all gates. We measure estimation error in the space of singular values. The estimation error is minimum at the direction of the singular vector corresponding to the largest singular value and so on. We say estimation subspace is n_e, when we project estimation error to the space of the first n_e singular vectors (singular vectors are sorted based on their corresponding singular values).

7.2 Variation Estimation Evaluation

To evaluate the performance of the proposed methods, we simulate the variation model (Section 3.1) on a number of MCNC benchmark circuits. A total of 12% random variations is assumed. Correlated intra-die variation is 60% of the total variation [25] [26]; 20% of the total variation is uncorrelated intra-die variation and the remaining variation is allotted to the inter-die variation.

We used SIS software to map the benchmark circuits to NAND2, NAND3, NAND4, NOR2, NOR3, NOR4, and inverter gates. Then, using Dragon, a placement software package [27], gates are placed on the IC. Since various gates cover different areas on the IC, gates are located on irregular grids.

To calculate the falling and rising coefficients (ξ_{f,g_u} and ξ_{r,g_u} in Equation 13), we implemented all the gates with 65nm CMOS transistor technology. Then, we used the HSPICE software to fit the linear model for all the gates.

Figure 7 shows variation estimation error for ℓ_2 minimization and ℓ_1 regularization methods. The horizontal axis is delay measurement noise and the vertical axis is variations estimation error. ℓ_1 regularization causes a 100%

Circuit properties						3% noise		6% noise		9% noise	
name	#gates	#inputs	#meas	$\frac{\sigma_{N/4}}{\sigma_1}$	subspace	ℓ_1 error	ℓ_2 error	ℓ_2 error	ℓ_2 error	ℓ_1 error	ℓ_2 error
C432	206	36	309	0.035	26	3.76	6.82	4.34	12.86	5.23	17.25
					52	6.57	12.58	7.75	21.22	9.5	30.846
C499	532	41	798	0.045	67	4.05	4.78	4.74	6.91	5.70	9.35
					135	11.52	12.28	12.48	15.11	13.80	18.60
C880	353	60	529	0.043	44	2.65	5.45	4.27	10.61	5.99	22.49
					89	5.34	11.56	7.93	21.71	10.9	36.5
C1355	517	41	775	0.038	65	2.55	4.11	4.17	7.87	5.90	11.69
					131	5.22	7.10	8.21	13.19	11.41	19.47
C1908	615	33	992	0.052	78	2.56	2.77	4.05	71.61	5.68	100
					156	4.78	5.25	7.57	70.94	10.58	97.21
C2670	900	233	1350	0.019	114	2.26	3.03	3.48	5.54	4.84	8.17
					229	5.22	7.27	7.66	13.29	10.51	19.60
alu2	360	10	540	0.0519	45	2.54	10.69	3.74	21.30	5.17	38.78
					91	4.88	25.70	7.89	51.28	11.28	78.55
alu4	733	14	1099	0.036	93	3.63	12.79	6.01	100	9.76	100
					186	6.42	20.41	10.22	102.93	15.76	102.93
comp	163	32	244	0.061	20	1.16	1.78	1.71	3.11	2.34	4.51
					41	2.63	4.43	3.81	8.05	5.19	11.87
cordic	102	23	153	0.099	13	3.37	5.11	5.04	9.41	6.93	13.90
					26	8.38	15.93	13.10	29.89	16.91	44.17
b9	113	41	169	0.15	14	1.62	11.19	2.13	22.34	2.75	33.50
					28	3.17	13.13	4.11	25.48	5.24	38.01
c8	165	28	247	0.22	20	2.32	9.43	4.12	18.72	5.85	28.03
					41	5.10	14.09	9.33	27.95	13.10	41.84

Table 2: Performance of ℓ_2-norm minimization and ℓ_1-norm regularization for a number of MCNC benchmark circuits.

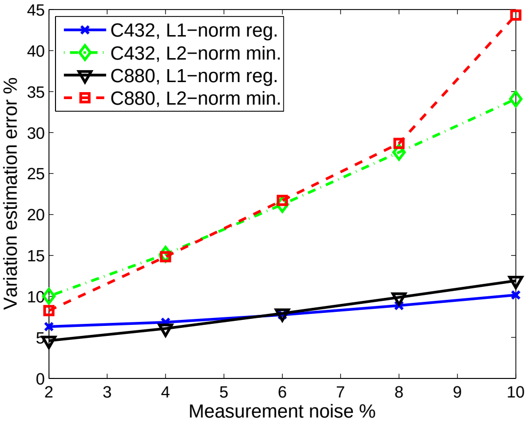

Figure 7: Variation (delay) estimation error vs. measurement error.

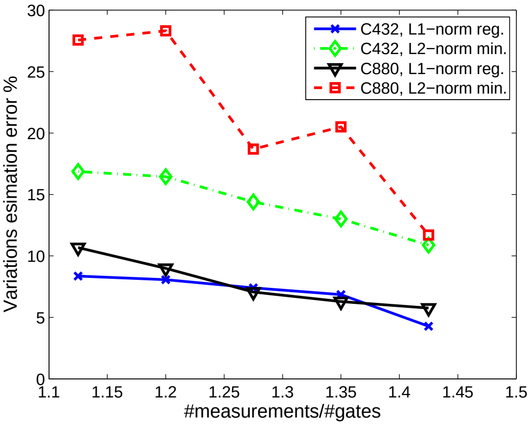

Figure 8: Variation (delay) estimation error vs. the number of measurements.

improvement over ℓ_2 minimization. The estimation subspace is 52 and 89 for C432 and C880 circuits, respectively. When measurement noise is small, delay measurements provides enough information to estimate variations accurately. As measurement noise increase, sparsity provides more irredundant information. Thus, performance of ℓ_1 regularization over ℓ_2 minimization increases as measurement noise increases.

The effect of the number of measurements is illustrated in Figure 8. The horizontal axis is the number of delay measurements divided by the number of the gates. Again, ℓ_1 regularization performs about 100% better than ℓ_2 minimization. On the figure, the estimation subspace is 52 and 89 for C432 and C880 circuits, respectively.

Finally, Table 7.1 shows results of variation estimation on 12 benchmark circuits. After the benchmarks' name, the first, the second and the third columns are the number of gates, the number of inputs in the circuit, and the number of delay measurements, respectively. The fourth column is the ratio of the $N/4$-th singular value to the first singular value in the measurement matrix (N is number of gates.) This column shows how fast singular values decay; or how the measurement matrix is well conditioned. The fifth column is the estimation subspace. The rest of the columns represent the estimation error (in percent) for ℓ_2 minimization and ℓ_1 regularization with 3%, 6%, and 9% percent measurement noise.

8. CONCLUSION

We have introduced a novel approach for post-silicon circuit timing characterization. The approach leverages the new theory of compressed sensing for accurate estimation of the sparse delay characteristics and distribution by using only a few noninvasive measurement data. To implement the approach, our compressed sensing-based framework employed the set of sensitizable paths (identified during the testing phase), sparse representation of the delay variations, structural logic relations, and methods to account for irregularity of the gate layouts. Experimental results demonstrate that by using the method, the post-silicon timing of the benchmark circuits could be characterized with an average accuracy of 95% in the pertinent subspace.

9. REFERENCES

[1] B. Cline, K. Chopra, D. Blaauw, and Y. Cao, "Analysis and modeling of CD variation for statistical static timing," in *ICCAD*, 2006, pp. 60–66.

[2] S. Lu, P. Hsieh, and J. Liou, "Exploring linear structures of critical path delay faults to reduce test efforts," in *ICCAD*, 2006, pp. 100–106.

[3] A. Srivastava, D. Sylvester, and D. Blaauw, *Statistical Analysis and Optimization for* VLSI*: Timing and Power*, ser. Series on Integrated Circuits and Systems. Springer, 2005.

[4] A. Ramalingam, G. Nam, A. Singh, M. Orshansky, S. Nassif, and D. Pan, "An accurate sparse matrix based framework for statistical static timing analysis," in *ICCAD*, 2006, pp. 231–236.

[5] S. M. Burns, M. Ketkar, N. Menezes, K. A. Bowman, J. W. Tschanz, and V. De, "Comparative analysis of conventional and statistical design techniques," in *DAC*, 2007, pp. 238–243.

[6] Q. Liu and S. Sapatnekar, "Confidence scalable post-silicon statistical delay prediction under process variations," in *DAC*, 2007, pp. 497–502.

[7] E. Candes, "Compressive sampling," in *Int. Congress of Mathematics*, 2006, pp. 1433–1452.

[8] R. Baraniuk, "A lecture on compressive sensing," *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 118–121, 2007.

[9] A. Agarwal, D. Blaauw, and V. Zolotov, "Statistical timing analysis for intra-die process variations with spatial correlations," in *ICCAD*, 2003, p. 900.

[10] H. Chang and S. Sapatnekar, "Statistical timing analysis under spatial correlations," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, no. 9, pp. 1467–1482, 2005.

[11] Y. Zhan, A. J. Strojwas, X. Li, L. T. Pileggi, D. Newmark, and M. Sharma, "Correlation-aware statistical timing analysis with non-gaussian delay distributions," in *Proceedings of the 42nd annual conference on Design automation*, 2005, pp. 77–82.

[12] J. Lesser and J. Shedletsky, "An experimental delay test generator for LSI logic," *IEEE Trans. on Computers*, vol. C-29, no. 3, pp. 235–248, 1980.

[13] M. Sharma and J. Patel, "Bounding circuit delay by testing a very small subset of paths," in *IEEE VLSI Test Symposium*, 2000, pp. 333–341.

[14] ——, "Finding a small set of longest testable paths that cover every gate," in *IEEE International Test Conference*, 2002, pp. 974–982.

[15] D. Donoho, "Compressed sensing," *IEEE Trans. on Info. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[16] F. Liu, "A general framework for spatial correlation modeling in vlsi design," in *DAC*, 2007, pp. 817–822.

[17] X. Lu, Z. Li, W. Qiu, D. M. H. Walker, and W. Shi, "Longest path selection for delay test under process variation," in *ASP-DAC: electronic design and solution fair*, 2004, pp. 98–103.

[18] V. Iyengar, J. Xiong, S. Venkatesan, V. Zolotov, D. Lackey, P. Habitz, and C. Visweswariah, "Variation-aware performance verification using at-speed structural test and statistical timing," in *ICCAD*, 2007, pp. 405–412.

[19] Z. Feng, P. Li, and Y. Zhan, "Fast second-order statistical static timing analysis using parameter dimension reduction," in *DAC*, 2007, pp. 244–249.

[20] A. Murakami, S. Kajihara, T. Sasao, I. Pomeranz, and S. M. Reddy, "A test structure for characterizing local device mismatches," in *IEEE International Test Conference*, 2000, p. 376.

[21] M. Bushnell and V. Agrawal, *ssentials of Electronic Testing for Digital, Memory, and Mixed-Signal VLSI Circuits*, 2000.

[22] S. Mallat, *A Wavelet Tour of Signal Processing.* Sandiego, USA: Academic Press, 1999.

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

[24] J. Gregg and T. W. Chen, "Post silicon power/performance optimization in the presence of process variations using individual well-adaptive body biasing," *IEEE Trans. VLSI*, vol. 15, no. 3, pp. 366–376, 2007.

[25] Y. Cao and L. T. Clark, "Mapping statistical process variations toward circuit performance variability: an analytical modeling approach," in *DAC*, 2005, pp. 658–663.

[26] J. Xiong, V. Zolotov, and L. He, "Robust extraction of spatial correlation," in *international symposium on Physical design*, 2006, pp. 2–9.

[27] "http://er.cs.ucla.edu/dragon/."