Bayesian Virtual Probe: Minimizing Variation Characterization Cost for Nanoscale IC Technologies via Bayesian Inference

Wangyang Zhang and Xin Li
Electrical & Computer Engineering Department
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213
{wangyan1, xinli}@ece.cmu.edu

Rob A. Rutenbar
Computer Science Department
University of Illinois at Urbana-Champaign
201 N. Goodwin Ave, Urbana IL 61801
rutenbar@illinois.edu

ABSTRACT

The expensive cost of testing and characterizing parametric variations is one of the most critical issues for today's nanoscale manufacturing process. In this paper, we propose a new technique, referred to as *Bayesian Virtual Probe* (BVP), to efficiently measure, characterize and monitor spatial variations posed by manufacturing uncertainties. In particular, the proposed BVP method borrows the idea of Bayesian inference and information theory from statistics to determine an optimal set of sampling locations where test structures should be deployed and measured to monitor spatial variations with maximum accuracy. Our industrial examples with silicon measurement data demonstrate that the proposed BVP method offers superior accuracy (1.5× error reduction) over the VP approach that was recently developed in [12].

Categories and Subject Descriptors

B.7.2 [**Integrated Circuits**]: Design Aids – Verification

General Terms

Algorithms

Keywords

Process Variation, Integrated Circuit, Variation Characterization

1. INTRODUCTION

As the feature size of integrated circuits continues to scale down, parametric variation of manufacturing process becomes increasingly difficult to control [1]-[2]. Great uncertainty in circuit behavior has been observed due to process variations, which is a growing issue with technology scaling. To understand and combat process variations, silicon testing and characterization [3] is an important part of the infrastructure that facilitates many statistical IC design techniques, such as statistical timing analysis [4]-[8], post-silicon tuning [9]-[11], etc. However, most traditional silicon characterization approaches are extremely expensive: Hundreds of test structures must be deployed in wafer scribe lines and/or product chips to accurately capture spatial variations.

Recently, a new technique referred to as *Virtual Probe* (VP) has been developed to minimize the cost of silicon characterization [12]. The key idea of VP is to measure very few test structures at a set of sampling locations of a wafer/chip. The parametric variations at other locations are not physically measured by hardware testing. Instead, spatial variation information at these locations is predicted through the use of a numerical algorithm. In other words, unlike the traditional approaches that require a large number of test structures, VP only physically monitors the variability at very few locations, thereby reducing the testing and characterization cost.

DAC'10, June 13-18, 2010, Anaheim, California, USA

While the efficacy of VP has been demonstrated by several industrial examples in [12], it remains an open question how VP could select optimal sampling locations to maximize prediction accuracy. The algorithm proposed in [12] randomly selects a number of sampling locations to collect measurement data. Such a simple sampling scheme may lead to large error, especially if a set of "bad" locations are randomly sampled. More importantly, random sampling does not necessarily minimize error and, hence, does not offer the best-possible accuracy that VP could achieve.

Motivated by these observations, we aim to develop an *optimal* sampling scheme to improve the accuracy of VP. In particular, we ask the following question: Where are the optimal sampling locations that result in maximum accuracy? The answer to this question, however, is *not* trivial. In this paper, we show that the optimal sampling locations strongly depend on the spatial pattern of process variations. Namely, they are different from process to process, from wafer to wafer, and from chip to chip. It is *impossible* to come up with a fixed set of sampling locations that are optimal for all cases. Instead, the best sampling locations must be *adaptively* "learned" in real time.

Towards this goal, we borrow the idea of *Bayesian inference* from statistics [20] to develop a new algorithm referred to as *Bayesian Virtual Probe* (BVP). BVP adaptively determines the optimal sampling locations by explicitly minimizing prediction error. Given a small set of measurement data, it first estimates the error of VP by using Bayes' theorem. Next, the optimal sampling locations are found to collect additional measurement data and improve prediction accuracy. Such an adaptive sampling scheme does not save silicon area, as we must deploy and manufacture all test structures in advance. However, it can substantially reduce testing/characterization time and eliminate many reliability issues for wafer probe test, since we only need to measure a small subset of test structures optimally selected by BVP.

Two important new contributions are made in this paper to uniquely tune the proposed BVP algorithm for our variation characterization application. First, we develop a new scheme to approximate the Bayesian inference by multivariate Gaussian distribution. As such, the posterior distribution (i.e., the error of VP) can be analytically estimated with low computational cost. Without this new approximation scheme, posterior distribution must be estimated by Monte Carlo simulation [20], which is computationally expensive and, hence, not feasible for many practical applications.

In addition, we further derive an analytical formula to estimate the differential entropy (i.e., the prediction uncertainty) based on information theory. Such an entropy formulation enables

us to efficiently select the optimal sampling locations to minimize prediction uncertainty (or equivalently, maximize prediction accuracy). As will be demonstrated by several industrial examples with silicon measurement data in Section 4, our proposed BVP method reduces the number of required sampling points by 1.5× compared to the earlier VP algorithm that relies only on random sampling.

The remainder of this paper is organized as follows. In Section 2 we briefly review the background of VP. In Section 3, we derive the Bayesian inference for our proposed BVP method, and then utilize it to develop a novel optimal sampling method that maximizes prediction accuracy based on information theory. The efficacy of BVP is demonstrated by several industrial examples with silicon measurement data in Section 4. Finally, we conclude in Section 5.

2. BACKGROUND

Let $g(x, y)$ be the two-dimensional function of the performance of interest, where x and y represent the coordinate of a location within the two-dimensional plane. The performance g can be the frequency of a ring oscillator, the threshold voltage of a transistor, etc. We discretize the two-dimensional function $g(x, y)$ and denote the coordinates x and y as integers $x \in \{1,2,...,P\}$ and $y \in \{1,2,...,Q\}$. Mathematically, the relation between the performance value and its frequency-domain component can be represented by a two-dimensional linear transform such as discrete cosine transform (DCT) [19]:

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

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

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

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

On the other hand, $\{g(x, y); x = 1,2,...,P, y = 1,2,...,Q\}$ can be represented as the linear combination of $\{G(u, v); u = 1,2,...,P, v = 1,2,...,Q\}$ by inverse discrete cosine transform (IDCT):

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

Virtual Probe (VP [12]) aims to measure a very small number of (say, M) samples at the locations $\{(x_m, y_m); m = 1,2,...,M\}$ and recover the performance value $g(x, y)$ at other locations $\{(x_m, y_m); m = M+1,M+2,...,PQ\}$ where $M << PQ$. Towards this goal, we formulate the following linear equation based on the measurement data $\{g(x_m, y_m); m = 1,2,...,M\}$:

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

where

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

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

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

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

We need to solve $\{G(u, v); u = 1,2,...,P, v = 1,2,...,Q\}$ from (5)-(9). Once the DCT coefficients are known, the unknown performance values $\{g(x_m, y_m); m = M+1,M+2,...,PQ\}$ can be easily calculated by IDCT in (4):

$$\widetilde{B} = \widetilde{A} \cdot \eta \quad (10)$$

where

$$\widetilde{A} = \begin{bmatrix} A_{M+1,1,1} & A_{M+1,1,2} & \cdots & A_{M+1,P,Q} \\ A_{M+2,1,1} & A_{M+2,1,2} & \cdots & A_{M+2,P,Q} \\ \vdots & \vdots & \vdots & \vdots \\ A_{PQ,1,1} & A_{PQ,1,2} & \cdots & A_{PQ,P,Q} \end{bmatrix} \quad (11)$$

$$\widetilde{B} = [g(x_{M+1},y_{M+1}) \quad g(x_{M+2},y_{M+2}) \quad \cdots \quad g(x_{PQ},y_{PQ})]^T. \quad (12)$$

In (11), $\{A_{m,u,v}; m = M+1,M+2,...,PQ, u = 1,2,...,P, v = 1,2,...,Q\}$ is defined by (7).

Studying (5)-(9), one would notice that M (the number of equations) is vastly less than PQ (the number of unknowns). The linear equation $A \cdot \eta = B$ is profoundly underdetermined. To solve (5), the authors of [12] further assume that the solution η is sparse. Namely, a large number of DCT coefficients are close to zero, but we do not know the exact locations of these zeros. Given this assumption, the solution η can be uniquely determined by solving the following optimization:

$$\begin{aligned} &\underset{\eta}{\text{minimize}} && \|\eta\|_0 \\ &\text{subject to} && A \cdot \eta = B \end{aligned} \quad (13)$$

where $\|\bullet\|_0$ stands for the L_0-norm of a vector, i.e., the number of non-zeros in the vector. The optimization in (13) attempts to minimize the number of non-zeros in η, while satisfying the linear equation $A \cdot \eta = B$. Hence, it results in a unique solution η that is as sparse as possible.

The optimization problem in (13) is NP hard and, hence, is extremely difficult to solve [15]-[16]. A more efficient technique to find sparse solution is based on L_1-norm regularization – a relaxed version of L_0-norm [12], [15]-[16]:

$$\begin{aligned} &\underset{\eta}{\text{minimize}} && \|\eta\|_1 \\ &\text{subject to} && A \cdot \eta = B \end{aligned} \quad (14)$$

where $\|\bullet\|_1$ denotes the L_1-norm of a vector, i.e., the summation of the absolute value of all elements in the vector. The L_1-norm regularization in (14) can be re-formulated as a linear programming problem and solved efficiently [12].

The Virtual Probe (VP) method proposed in [12] randomly selects a number of sampling locations to collect measurement data and build the linear equation $A \cdot \eta = B$ in (5). Such a simple sampling scheme may lead to large error, especially if a set of "bad" locations are randomly sampled. In addition, random sampling does not lead to minimum error and, hence, does not offer the best-possible accuracy. Motivated by this observation, this paper aims to develop a new Bayesian Virtual Probe (BVP) algorithm that adaptively determines the optimal sampling locations to achieve minimum prediction error.

3. BAYESIAN VIRTUAL PROBE

The key novelty of our proposed Bayesian Virtual Probe (BVP) lies in an iterative optimal sampling technique that maximizes prediction accuracy. Starting from a small set of initial

samples, BVP determines a set of optimal sampling locations to collect additional measurement data so that the prediction error is minimized. Such an optimal sampling scheme is facilitated by two critical techniques: (1) efficient error prediction by Bayesian inference, and (2) optimal sample selection based on information theory. These two core techniques enable us to efficiently estimate and, consequently, minimize the prediction error. In this section, we describe the mathematical formulation of BVP and highlight its novelties.

3.1 Bayesian Inference

Bayesian inference is an efficient statistical method for error estimation [20]. Unlike the L_1-norm regularization in (14) that only solves a deterministic solution, Bayesian inference aims to find a probability density function (PDF) that quantitatively measures the prediction error. For instance, a large variance of the PDF implies large uncertainty and low accuracy.

Towards this goal, we first need to define a so-called *prior distribution* for η. Intuitively, such a prior distribution represents our prior knowledge about η. In this paper, we construct the following prior distribution for each element of the vector $\eta = [\eta_1, \eta_2, ..., \eta_{PQ}]^T$:

$$pdf(\eta_i) = \frac{1}{\sigma_i\sqrt{2\pi}} \cdot \exp\left(-\frac{\eta_i^2}{2\sigma_i^2}\right) \quad (i = 1,2,\cdots,PQ). \tag{15}$$

The prior PDF in (15) is modeled as a zero-mean Normal distribution with a unique variance σ_i^2 assigned to each η_i. In other words, we model the DCT coefficients $\{\eta_i;\ i = 1,2,...,PQ\}$ by different probability distributions. Figure 1 shows two zero-mean Normal distributions with $\sigma = 0.1$ and $\sigma = 1$, respectively. Studying Figure 1, one would notice that if the prior distribution $pdf(\eta_i)$ has a small (or large) variance, the DCT coefficient η_i is likely to be zero (or non-zero).

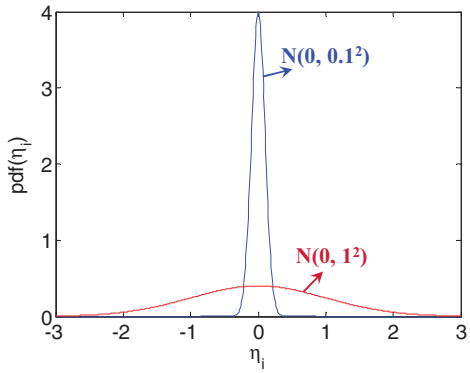

Figure 1. Two zero-mean Normal probability density functions ($\sigma = 0.1$ and $\sigma = 1$ respectively) as prior distributions.

Determining the appropriate values for $\{\sigma_i;\ i = 1,2,...,PQ\}$, however, is not trivial. We only know that the vector η is sparse; however, we do not know the exact locations of these zeros. To address this difficulty, we re-use the idea of L_1-norm regularization described in Section 2. Namely, given a small set of initial samples, we solve the optimization in (14) to "estimate" the values of $\{\eta_i;\ i = 1,2,...,PQ\}$. If the solution η_i from (14) is far away from zero, the corresponding variance σ_i^2 should be large, implying that the coefficient η_i is likely to be non-zero. Otherwise, if the solution η_i from (14) is close to zero, the corresponding variance σ_i^2 should be small, implying that the coefficient η_i is likely to be zero. From this point of view, the proposed BVP technique re-uses the result of VP, i.e., the vector η solved from (14), to further create a Bayesian inference for error estimation and optimal sampling.

Based on the above discussion, we formulate the following equation to determine $\{\sigma_i;\ i = 1,2,...,PQ\}$:

$$\sigma_i = \lambda \cdot \left|\eta_i^{L1}\right| \quad (i = 1,2,\cdots,PQ) \tag{16}$$

where η_i^{L1} represents the i-th element of the solution η solved from the L_1-norm regularization in (14), and $\lambda > 0$ is a scaling factor that controls the variance of the distribution. In practice, however, it is not necessary to know the exact value of λ. As will be shown in Section 3.2, the final decision of optimal sampling locations is independent of λ. It should be noted that while the model in (16) looks simple, it has been successfully applied to a number of practical problems related to sparse regression and demonstrated with promising results, e.g., [17]-[18].

To complete our definition of the prior distribution, we further assume that all DCT coefficients in the vector η are mutually independent. Hence, the joint probability density function $pdf(\eta)$ is:

$$pdf(\eta) = \frac{1}{(2\pi)^{PQ/2} \cdot \sqrt{\det(\Sigma_\eta)}} \cdot \exp\left(-\frac{1}{2} \cdot \eta^T \cdot \Sigma_\eta^{-1} \cdot \eta\right) \tag{17}$$

where $\det(\bullet)$ denotes the determinant of a matrix, and the covariance matrix Σ_η is equal to:

$$\Sigma_\eta = \lambda^2 \cdot diag\left(\left|\eta_1^{L1}\right|^2, \left|\eta_2^{L1}\right|^2, \cdots, \left|\eta_{PQ}^{L1}\right|^2\right). \tag{18}$$

The independence assumption simply implies that we do not know the correlation of $\{\eta_i;\ i = 1,2,...,PQ\}$ in advance. The correlation information will be taken into account, once the prior distribution $pdf(\eta)$ is combined with our measurement data $A \cdot \eta = B$ to determine the posterior distribution.

To derive the proposed Bayesian inference, we re-write the prior distribution in terms of the measured performance $B = A \cdot \eta$ and the unmeasured performance $\tilde{B} = \tilde{A} \cdot \eta$ that are defined in (5) and (10), respectively. Given the zero-mean Normal distribution $pdf(\eta)$ in (17)-(18) and the linear mapping in (5) and (10), it is easy to verify that the prior distribution of $(B,\tilde{B})$ is also zero-mean and jointly Normal [20]:

$$pdf\left(B,\hat{B}\right) = \frac{1}{(2\pi)^{PQ/2} \cdot \sqrt{\det\left(\Sigma_{B,\hat{B}}\right)}} \cdot \exp\left(-\frac{1}{2} \cdot \begin{bmatrix} B \\ \hat{B} \end{bmatrix}^T \cdot \Sigma_{B,\hat{B}}^{-1} \cdot \begin{bmatrix} B \\ \hat{B} \end{bmatrix}\right) \tag{19}$$

where the covariance matrix $\Sigma_{B,\tilde{B}}$ is equal to:

$$\Sigma_{B,\tilde{B}} = \lambda^2 \cdot \begin{bmatrix} A \\ \tilde{A} \end{bmatrix} \cdot diag\left(\left|\eta_1^{L1}\right|^2, \left|\eta_2^{L1}\right|^2, \cdots, \left|\eta_{PQ}^{L1}\right|^2\right) \cdot \begin{bmatrix} A \\ \tilde{A} \end{bmatrix}^T. \tag{20}$$

Note that $\Sigma_{B,\tilde{B}}$ is *not* diagonal, implying that the random variables in $(B,\tilde{B})$ are statistically correlated.

Next, we take the measurement data $A \cdot \eta = B$ into account. Given these measurement data, the vector B is fixed. Namely, we know the exact value of B (except for measurement noise), and the vector B now becomes deterministic, instead of being modeled as a multivariate random variable. The proposed Bayesian inference aims to utilize the measured performance value B to extract extra information about the unmeasured performance value $\tilde{B}$. This is possible, because B and $\tilde{B}$ are statistically correlated, as shown in (20). Mathematically, the extra knowledge about $\tilde{B}$ that we can learn from B is represented by the *posterior distribution*, i.e., the conditional probability $pdf(\tilde{B}|B)$. Based on Bayes' theorem and the prior distribution in (19)-(20), the posterior distribution $pdf(\tilde{B}|B)$ is also jointly Normal and its mean and covariance are respectively equal to [20]:

$$\mu_{\tilde{B}|B} = -\Delta_{\tilde{B}\tilde{B}}^{-1} \cdot \Delta_{\tilde{B}B} \cdot B \tag{21}$$

$$\Sigma_{\tilde{B}|B} = \lambda^2 \cdot \Delta_{\tilde{B}\tilde{B}}^{-1} \tag{22}$$

where $\Delta_{\tilde{B}B}$ and $\Delta_{\tilde{B}\tilde{B}}$ are the block partitions of the matrix $\lambda^2 \cdot (\Sigma_{B,\tilde{B}})^{-1}$:

$$\begin{bmatrix} \Delta_{BB} & \Delta_{B\tilde{B}} \\ \Delta_{\tilde{B}B} & \Delta_{\tilde{B}\tilde{B}} \end{bmatrix} = \left\{ \begin{bmatrix} A \\ \tilde{A} \end{bmatrix} \cdot diag\left(\left|\eta_1^{L1}\right|^2, \left|\eta_2^{L1}\right|^2, \cdots, \left|\eta_{PQ}^{L1}\right|^2 \right) \cdot \begin{bmatrix} A \\ \tilde{A} \end{bmatrix}^T \right\}^{-1}. \tag{23}$$

It is worth mentioning that other Bayesian inferences can also be derived, e.g., using a Laplace distribution as the prior [12]. However, the posterior distribution cannot be easily solved in such cases. Hence, these Bayesian inferences do not fit the need of our application in this paper.

The posterior covariance $\Sigma_{\tilde{B}|B}$ in (22) offers a quantitative measure of the uncertainty of $\tilde{B}$. Namely, if the prediction of $\tilde{B}$ is accurate, the covariance matrix $\Sigma_{\tilde{B}|B}$ should be close to zero. Otherwise, if the prediction has large error, the posterior distribution should carry a large variance. It, in turn, provides a useful guideline for us to optimally collect extra measurement data to minimize prediction error. Intuitively, if the k-th element of $\tilde{B}$ (i.e., $\tilde{B}_k$) has a large posterior variance, we should directly measure $\tilde{B}_k$ to minimize uncertainty. In what follows, we will derive a mathematical framework based on information theory to determine optimal sampling locations.

3.2 Optimal Sampling

According to information theory, the uncertainty of a statistical system $pdf(\tilde{B}|B)$ can be quantitatively measured by the following differential entropy [20]:

$$H\left(\tilde{B} \mid B\right) = -\int pdf\left(\tilde{B} \mid B\right) \cdot \log\left[pdf\left(\tilde{B} \mid B\right)\right] \cdot d\tilde{B}. \tag{24}$$

Intuitively, the differential entropy in (24) is large, if the system is profoundly undetermined. In other words, the differential entropy should be minimized in order to reduce the uncertainty and, hence, achieve an accurate estimation. Given the jointly Normal posterior distribution $pdf(\tilde{B}|B)$ for which the covariance is specified in (22), the differential entropy in (24) can be re-written as:

$$H\left(\tilde{B} \mid B\right) = \frac{PQ-M}{2} \cdot \left(1 + \log 2\pi\right) + \frac{1}{2} \cdot \det\left(\lambda^2 \cdot \Delta_{\tilde{B}\tilde{B}}^{-1}\right) \tag{25}$$

where PQ–M is the dimension of $\tilde{B}$, i.e., $\tilde{B} \in R^{PQ-M}$. Note that for a jointly Normal distribution with a given dimension, the differential entropy is uniquely determined by the covariance matrix.

The objective of BVP is to select several (say, K) additional samples that we should further measure to minimize the prediction error or, equivalently, minimize the differential entropy. To simplify our notation, we re-order the vector $\tilde{B}$ such that:

$$W \cdot \tilde{B} = \begin{bmatrix} \tilde{B}_K \\ \tilde{B}_{\tilde{K}} \end{bmatrix} \tag{26}$$

where W is a permutation matrix and $\tilde{B}_K$ is a vector that contains the K additional samples that we should measure. Given the covariance matrix in (22) and the permutation in (26), the covariance matrix of $W\tilde{B}$ is:

$$\Sigma_{\tilde{B}_K, \tilde{B}_{\tilde{K}}|B} = \lambda^2 \cdot W \Delta_{\tilde{B}\tilde{B}}^{-1} W^T = \lambda^2 \cdot \begin{bmatrix} \Sigma_{KK} & \Sigma_{K\tilde{K}} \\ \Sigma_{\tilde{K}K} & \Sigma_{\tilde{K}\tilde{K}} \end{bmatrix} \tag{27}$$

where $\Sigma_{\tilde{B}_K, \tilde{B}_{\tilde{K}}|B}$ is partitioned into four blocks: Σ_{KK}, $\Sigma_{K\tilde{K}}$, $\Sigma_{\tilde{K}K}$ and $\Sigma_{\tilde{K}\tilde{K}}$. Once $\tilde{B}_K$ is measured, the conditional probability $pdf(\tilde{B}_{\tilde{K}}|\tilde{B}_K,B)$ is still jointly Normal and its covariance matrix is [20]:

$$\Sigma_{\tilde{B}_{\tilde{K}}|\tilde{B}_K,B} = \lambda^2 \cdot \left(\Sigma_{\tilde{K}\tilde{K}} - \Sigma_{\tilde{K}K} \Sigma_{KK}^{-1} \Sigma_{K\tilde{K}}\right). \tag{28}$$

Hence, the differential entropy of $pdf(\tilde{B}_{\tilde{K}}|\tilde{B}_K,B)$ can be written as:

$$H\left(\tilde{B}_{\tilde{K}} \mid \tilde{B}_K, B\right) = \frac{PQ-M-K}{2} \cdot \left(1 + \log 2\pi\right) + \frac{1}{2} \cdot \det\left[\lambda^2 \cdot \left(\Sigma_{\tilde{K}\tilde{K}} - \Sigma_{\tilde{K}K} \Sigma_{KK}^{-1} \Sigma_{K\tilde{K}}\right)\right] \tag{29}$$

where PQ–M–K is the dimension of $\tilde{B}_{\tilde{K}}$, i.e., $\tilde{B}_{\tilde{K}} \in R^{PQ-M-K}$. Based on the block partitions in (27) and the theory of Schur complement [21], $H(\tilde{B}_{\tilde{K}}|\tilde{B}_K,B)$ can be re-formulated as:

$$H\left(\tilde{B}_{\tilde{K}} \mid \tilde{B}_K, B\right) = \frac{PQ-M-K}{2} \cdot \left(1 + \log 2\pi\right) + \frac{1}{2} \cdot \frac{\det\left(\lambda^2 \cdot W \Delta_{\tilde{B}\tilde{B}}^{-1} W^T\right)}{\det\left(\lambda^2 \cdot \Sigma_{KK}\right)}. \tag{30}$$

Since $\det(W) \cdot \det(W^T) = 1$ for the permutation matrix W and the parameter λ appears at both the numerator and the denominator, $H(\tilde{B}_{\tilde{K}}|\tilde{B}_K,B)$ can be further simplified as:

$$H\left(\tilde{B}_{\tilde{K}} \mid \tilde{B}_K, B\right) = \frac{PQ-M-K}{2} \cdot \left(1 + \log 2\pi\right) + \frac{1}{2} \cdot \frac{\det\left(\Delta_{\tilde{B}\tilde{B}}^{-1}\right)}{\det\left(\Sigma_{KK}\right)}. \tag{31}$$

Two important observations can be made from (31). First, to minimize the differential entropy $H(\tilde{B}_{\tilde{K}}|\tilde{B}_K,B)$, we need to optimally select K elements (i.e., K sampling locations) out of $\tilde{B}$ so that the determinant of the corresponding covariance matrix $\det(\Sigma_{KK})$ is maximized. In particular, if only a single sampling location is selected, maximizing $\det(\Sigma_{KK})$ is equivalent to finding the k-th element of $\tilde{B}$ that has the maximum variance. Second, the value of $H(\tilde{B}_{\tilde{K}}|\tilde{B}_K,B)$ is independent of λ. Hence, we can simply set $\lambda = 1$ in our numerical computation. It does not change the optimal sampling locations that we select.

3.3 Summary

Algorithm 1: Bayesian Virtual Probe (BVP)

1. Start from a given integer N representing the total number of sampling locations to be measured.
2. Set the scaling factor $\lambda = 1$ for our numerical computation.
3. Randomly select a small number of (say, $M << N$) sampling locations and collect the measurement data at these locations: $\Omega = \{g(x_m, y_m);\ m = 1,2,...,M\}$.
4. Formulate the linear equation in (5) and solve the L_1-norm regularization problem in (14) to estimate η.
5. If $M = N$, stop iteration and go to Step 9.
6. Define the prior distribution $pdf(B,\tilde{B})$ in (19)-(20) and calculate the posterior distribution $pdf(\tilde{B}|B)$ using (21)-(23).
7. Find the k-th element of $\tilde{B}$ that has the maximum posterior variance. The corresponding sampling location is denoted as (x_k, y_k).
8. Collect the measurement data $g(x_k, y_k)$. Set $\Omega = \Omega \cup \{g(x_k, y_k)\}$ and $M = M+1$. Go to Step 4.
9. Apply IDCT in (4) to recover the performance function $\{g(x, y);\ x = 1,2,...,P,\ y = 1,2,...,Q\}$ spatially across the wafer/chip.

Algorithm 1 summarizes the major steps of the proposed BVP method. It repeatedly applies L_1-norm regularization to estimate η, and then build the proposed Bayesian inference to optimally select the next sampling location. Note that even though Algorithm 1 only finds a single optimal sampling location at one time, the proposed BVP technique can be easily extended to select multiple sampling locations simultaneously.

Our current implementation of Algorithm 1 requires to know N (i.e., the total number of sampling points) in advance. In practice, the required number of sampling points can be determined by monitoring the difference of the prediction results between two successive iteration steps. Furthermore,

measurement cost (i.e., the maximum number of sampling points that are affordable) must also be considered when selecting *N*. Due to space limit, the details of these implementation options are not discussed here.

Finally, it is important to mention that BVP requires a small number of initial samples to start the iteration. This initial set of samples is randomly generated in Algorithm 1, although other deterministic schemes [15]-[16] can also be applied for initial sampling. After this initialization step, the main body of Algorithm 1 does not involve any other random sampling, thereby eliminating the possibility of getting "bad" samples. As will be demonstrated by our industrial examples with silicon measurement data in Section 4, the proposed BVP algorithm achieves superior accuracy (more than 1.5× error reduction) over the original VP method. Furthermore, the proposed Bayesian inference can be analytically solved by simple matrix inverse and, hence, is extremely efficient with low computational cost.

4. NUMERICAL EXAMPLES

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

4.1 Flush Delay Measurement Data

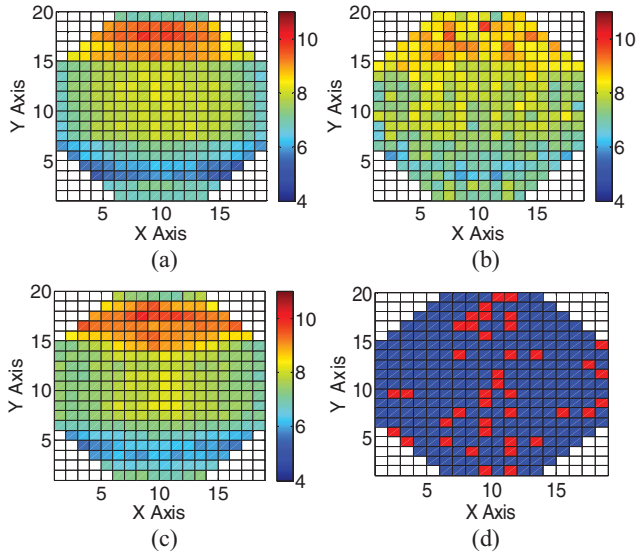

Figure 2. (a) Measured flush-delay values (normalized by a randomly selected constant) of 282 industrial chips from the same wafer. (b) Recovered flush-delay values from 40 tested chips by VP. (c) Recovered flush-delay values from 40 tested chips by BVP. (d) Optimal 40 sampling locations (marked by red color) selected by BVP.

We consider the flush-delay values measured from 282 industrial chips on the same wafer, as shown in Figure 2(a). In this example, the measured delay significantly varies from chip to chip due to process variations; however, there is a strong spatial correlation in delay variability. Our objective is to predict such a spatial variation by measuring few silicon chips.

For testing and comparison, two different techniques are implemented to predict spatial variation: (1) the traditional VP with random Latin Hypercube sampling [12], and (2) the proposed BVP with optimal sampling location selection. In this example, BVP first measures 10 randomly selected chips and it iteratively collects additional measurement data based on Algorithm 1. Figure 2(b) and Figure 2(c) show the recovered flush-delay values from 40 tested chips by VP and BVP respectively. Figure 2(d) further shows the locations of these 40 chips optimally selected by BVP.

Two important observations can be made from the results in Figure 2. First, by testing 40 chips only, BVP accurately capture the spatial variation of delay, while VP fails to achieve the same accuracy. In other words, BVP provides superior accuracy over VP in this example. Second, as shown in Figure 2(d), BVP automatically selects a large portion of tested chips in the middle of the wafer along the Y axis. This direction exactly follows the gradient of delay variation. It, therefore, enables us to accurately predict the spatial variation information in this example.

Figure 3 shows the average prediction error as a function of the number of samples (i.e., tested chips) for both VP and BVP. Note that BVP achieves more than 1.5× error reduction over VP. The error of BVP is around 5% when 40 chips (out of 282 chips in total) are tested. To achieve the same accuracy, VP has to measure 60~70 chips (1.5× more).

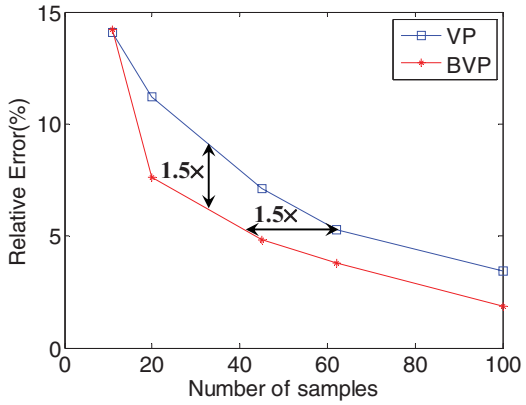

Figure 3. Comparison of prediction error for VP and BVP with different number of samples.

4.2 Leakage Current Measurement Data

We consider the leakage-current measurement collected by IDDQ test for the same silicon wafer. Figure 4(a) shows the normalized leakage-current values $\log_{10}(I_{Leak})$ (after logarithmic transform) as a function of the spatial locations. Compared to the delay variation shown in Figure 2(a), the spatial pattern of leakage-current measurement is less regular. It, in turn, implies that leakage prediction is inherently more difficult than delay prediction for this wafer being tested.

Similar to the previous example, we apply both VP and BVP to estimate the spatial variation of leakage current. In our experiment, BVP first measures 10 randomly selected chips and it iteratively collects additional measurement data based on Algorithm 1. Figure 4(b) and Figure 4(c) show the predicted leakage-current values from 50 tested chips by VP and BVP respectively. Figure 4(d) shows the locations of these 50 chips optimally selected by BVP.

Comparing Figure 4(b) and Figure 4(c), one would notice that the proposed BVP method offers substantially better accuracy than the traditional VP approach. The accuracy is improved by BVP, because it can optimally select the sampling locations to accommodate the unique spatial pattern of leakage variation.

Figure 5 further shows the average prediction error as a function of the number of samples (i.e., tested chips) for both VP and BVP. When 50 chips (out of 282 chips in total) are tested, the error of BVP is around 6.5%. To achieve the same accuracy, VP has to measure 70~80 chips (1.5× more).

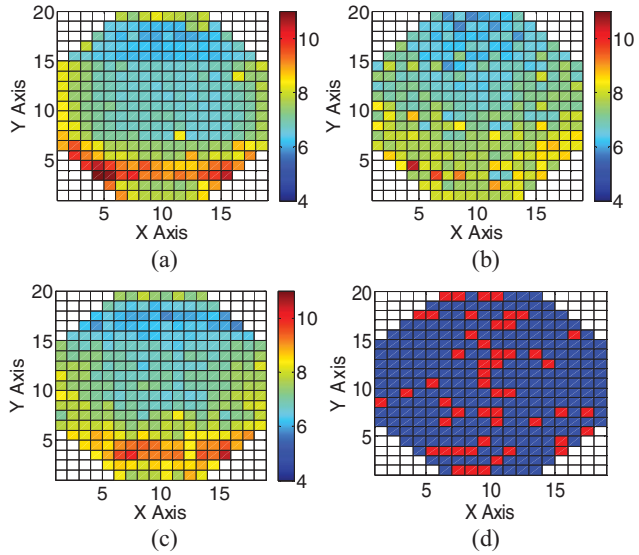

Figure 4. (a) Measured leakage-current values $\log_{10}(I_{Leak})$ (normalized by a randomly selected constant) of 282 industrial chips from the same wafer. (b) Recovered leakage-current values from 50 tested chips by VP. (c) Recovered leakage-current values from 50 tested chips by BVP. (d) Optimal 50 sampling locations (marked by red color) selected by BVP.

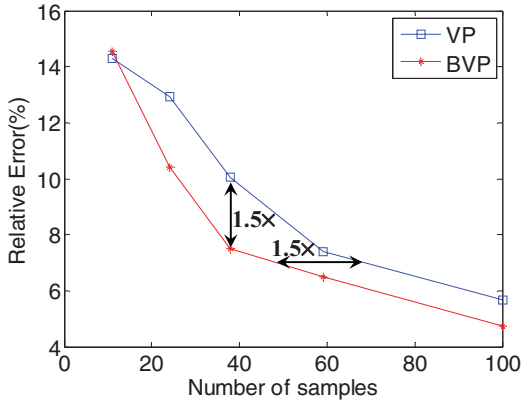

Figure 5. Comparison of prediction error for VP and BVP with different number of samples.

5. CONCLUSIONS

In this paper, we propose a novel Bayesian Virtual Probe (BVP) method to generate an optimal set of sampling locations for minimum-cost silicon testing and characterization. The proposed BVP technique applies an efficient Bayesian inference to estimate the prediction error based on posterior distribution. In addition, an optimal sample selection algorithm is derived from information theory to minimize the prediction error. Several industrial examples with silicon measurement data demonstrate that the proposed BVP technique offers superior accuracy (1.5× error reduction) over the traditional VP approach [12]. Our future research will further improve the BVP algorithm by considering various non-ideal effects (e.g., measurement noise) and apply it to practical integrated circuit design and testing problems such as wafer probe testing and post silicon tuning.

6. ACKNOWLEDGEMENTS

The authors acknowledge the support of the C2S2 Focus Center, one of six research centers funded under the Focus Center Research Program (FCRP), a Semiconductor Research Corporation entity. This work is also supported in part by the National Science Foundation under contract CCF–0915912.

7. REFERENCES

[1] S. Nassif, "Modeling and analysis of manufacturing variations," *IEEE CICC*, pp. 223-228, 2001.

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

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

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

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

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

[7] S. Burns, M. Ketkar, N. Menezes, K. Bowman, J. Tschanz and V. De, "Comparative analysis of conventional and statistical design techniques," *IEEE DAC*, pp. 238-243, 2007.

[8] D. Blaauw, K. Chopra, A. Srivastava and L. Scheffer, "Statistical timing analysis: from basic principles to state of the art," *IEEE Trans. CAD*, vol. 27, no. 4, pp. 589-607, Apr. 2008.

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

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

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

[12] X. Li, R. Rutenbar and R. Blanton, "Virtual probe: a statistically optimal framework for minimum-cost silicon characterization of nanoscale integrated circuits," *IEEE ICCAD*, pp. 433-440, 2009.

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

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

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

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

[17] I. Gorodnitsky and B. Rao, "Sparse signal reconstruction from limited data using FOCUSS: a re-weighted minimum norm algorithm," *IEEE Trans. Signal Processing*, vol. 45, no. 3, pp. 600-616, Mar. 1997.

[18] E. Candes, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted l1 minimization," *Journal of Fourier Analysis and Application*, vol. 14, no. 5, pp. 877-905, Dec 2008.

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

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

[21] R. Horn and C. Johnson, *Matrix Analysis*, Cambridge Press, 2007.