Adjustment-Based Modeling for Statistical Static Timing Analysis with High Dimension of Variability

Lin Xie, Azadeh Davoodi, Jun Zhang†, Tai-Hsuan Wu
Department of Electrical and Computer Engineering, †Department of Statistics
University of Wisconsin, Madison, WI 53706

Abstract—This paper presents an adjustment-based modeling framework for Statistical Static Timing Analysis (SSTA) when the dimension of parameter variability is high. Instead of building a complex model between the circuit timing and parameter variability, we build a model which adjusts an approximate variation-aware timing into an accurate one. The intuition is that it is simpler to build a model which adjusts an approximate estimate into an accurate one. It is also more efficient to obtain an approximate circuit timing model. The combination of these two observations makes the use of an adjustment-based model a good choice for SSTA with high dimension of parameter variability. To build the adjustment model, we use a simulation-based approach, which is based on Gaussian Process. Combined with intelligent sampling, we show that an adjustment-based model can more effectively capture the nonlinearity of the circuit timing with respect to parameter variability compared to polynomial modeling. We also show that with only 200 samples of the circuit timing and 42 independent parameter variations, adjustment-based modeling obtains higher accuracy than direct SSTA using quadratic modeling.

I. INTRODUCTION

In today’s sub-45nm era, we expect an increase in the dimension of parameter variability as well as in their ranges [7], [5]. Consequently, an efficient and accurate Statistical Static Timing Analysis (SSTA), which can handle high dimension of variability, becomes a crucial need. It might be necessary during the sign-off stage especially when accurate interconnect and device parameter variability should be incorporated. It can also be used to more effectively make optimization decisions or to perform sensitivity analysis. The majority of the existing SSTA techniques are based on polynomial model of the circuit timing such as [3]-[4], [6], [10]-[12]. Many of them experimented with only few number of parameter variabilities (for example one global $V_t$, $L_{on}$, $t_{oa}$ for all devices) and reported high modeling accuracy. With the increase in the dimension of parameter variability, we require complex polynomials to achieve the same accuracy.

In this paper, we introduce a new SSTA technique which is useful for high dimension of variability. We first study Gaussian Process (GP) as an alternative to model the nonlinear dependencies of the circuit timing as a function of parameter variability. GP models are built using a simulation-driven approach in which the models are constructed solely using the input-output behavior collected for a limited number of intelligently chosen variation samples. We show this type of modeling has a simpler form than high-order polynomials but the modeling complexity depends on the number of samples.

We then introduce a novel adjustment-based modeling framework for SSTA. Instead of building a complex model between the circuit timing and the parameter variability, we build a simple model which adjusts an approximate circuit timing into an accurate one. The adjustment model is built using GP, which is simulation-driven. However, we show that the number of required samples can be greatly reduced for an adjustment-based model. This is because approximate models have already captured some nonlinear dependencies in terms of the input parameter variability.

We introduce two ways to obtain approximate circuit timing. The first one obtains the circuit timing from an extracted subcircuit which is expected to be timing-critical. The second one is to use a cheap linear SSTA. We then adjust these approximate timings into accurate ones. Here, by accurate timing, we mean a timing that is obtained using Monte Carlo (MC) simulation over the entire circuit. Our simulation results show that an adjustment-based timing model can achieve higher accuracy compared to a quadratic model of the circuit timing, which is directly a function of parameter variables. We also show that for 42 independent parameter variables, we only need 200 samples to build an adjustment-based model while fitting-based quadratic modeling requires at least 10,000.

II. PREVIOUS WORKS AND SHORTCOMINGS

Previous SSTA works [3]-[4], [6], [10]-[12] always express the gate delay and arrival times using polynomial model in terms of variations in transistor or interconnect parameters. Particularly, [4], [6], [12] used quadratic forms to express the circuit timing, in which high modeling accuracy are reported. However, the dimension of the parameter variability used for accuracy verification in all these works is low.

With the scaling into sub-45nm era, the dimension of variability will become high even after implementing efficient parameter order reduction techniques. [7] and [5] reported 25 and 30 sources of uncorrelated variations after reduction, respectively. In addition, as illustrated in [2], low-order polynomial models become inaccurate with the increase in the dimension and the range of the variability. We also would like to note that quadratic model in this case also have very high computation complexity. As an example, a quadratic expression with 30 uncorrelated variables has over 400 terms if cross-terms are not ignored. Therefore, existing polynomial-based SSTA approaches need to be revisited. To address the aforementioned problems, we propose to express the circuit timing using Gaussian Process (GP) model, which is simpler than polynomial model and can flexibly capture a large class of complex nonlinear relationships [2], [9]. To further reduce the complexity of GP model, we propose adjustment-based model in Section IV.

III. OVERVIEW OF GAUSS PROCESS MODEL

Given the input uncertainty vector $\mathbf{x} = \{x_1, x_2, ..., x_d\} \in \mathbb{R}^d$, we model the deterministic circuit timing as [9]:

$$y(\mathbf{x}) = \beta \mathbf{f}^T(\mathbf{x}) + \epsilon(\mathbf{x}) = \sum_{i=0}^{n} \beta_i f_i(\mathbf{x}) + \epsilon(\mathbf{x}), \quad (1)$$

where $f(x) \triangleq \{f_1(x_1), ..., f_n(x_n)\}$ represents pre-specifed functions, which are typically a constant or a linear expression.

The $\epsilon(x)$ in Eq. (1) is a stationary Gaussian Process (GP) with mean 0 and variance $\sigma^2$. Its correlation structure defined between any two input sample vectors $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ is typically modeled as

$$R(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = \prod_{i=1}^{d} \exp(-\theta_i |x_i^{(1)} - x_i^{(2)}|^2), \quad (2)$$

where $x_i^{(j)}$ denotes the $i^{th}$ element of $\mathbf{x}^{(j)}$.
Fig. 1. The flow of the adjustment-based modeling framework.

In order to estimate the unknown parameters \( \beta \) in Eq. (1) as well as \( \theta = \{ \theta_1, \ldots, \theta_d \} \) in Eq. (2) and \( \sigma^2 \) which defines \( \epsilon(x) \), we first collect \( N \) input/output samples, and each sample refers to the input parameter variations and their corresponding output circuit timing. We denote the input/output of the \( i \)th sample by \( x^{(i)} \) and \( y^{(i)} \), respectively. We also denote the unknowns after estimation as \( \hat{\beta} \) and \( \hat{\theta} \), respectively.

As shown in [9], we can obtain the estimation \( \hat{\theta} \) via:

\[
\max_{\theta} -\frac{1}{2} \left[ N \log \sigma^2(\theta) + \log \left( \det(\mathbf{R}(\theta)) \right) + N \right],
\]

where

\[
\hat{\sigma}^2(\theta) \triangleq \frac{1}{N} \left[ (y - \mathbf{F}\hat{\beta})^T \mathbf{R}^{-1}(\theta) (y - \mathbf{F}\hat{\beta}) \right],
\]

and \( \mathbf{R}(\theta) \) has its \( ij \)th entry as \( R(x^{(i)}, x^{(j)}) \) as defined in Eq. (2).

In Eq. (4), \( y \) is a column vector with its \( i \)th entry as the \( i \)th observation (i.e., \( y^{(i)} \)), \( \mathbf{F} \) is a matrix with its \( ij \)th entry as \( f_j(x^{(i)}) \), and \( \hat{\beta} \) is defined as

\[
\hat{\beta} \triangleq \left( \mathbf{F}^T \mathbf{R}^{-1} \mathbf{F} \right)^{-1} \mathbf{F}^T \mathbf{R}^{-1} y.
\]

The optimization problem in Eq. (3) can be efficiently solved using mleg in "R" [1]. Using these estimated coefficients, we can compute the circuit timing for any given input variation vector \( x^* \) using

\[
y(x^*) = \hat{\beta}^T f(x^*) + \mathbf{r}^T \mathbf{R}^{-1} (y - \mathbf{F}\hat{\beta}),
\]

where \( \mathbf{r} \) is a vector with its \( i \)th entry as \( R(x^*, x^{(i)}) \).

**Complexity of GP Model:** As illustrated in Eq. (3), the number of unknowns to implement GP modeling is equal to \( d + n + 1 \), where \( d \) is the dimension of the variability and \( n \) is the number of \( f_i(x) \) (\( n \) is always very small). Note that these unknowns will maintain even when the Probability Density Function (PDF) of \( x \) changes.

As illustrated in Eq. (6), to predict the circuit timing with different \( x^* \) only requires the computation of \( f(x^*) \) and \( \mathbf{r} \). All other components such as \( \mathbf{R}, \mathbf{F}, \) and \( \hat{\beta} \) will not change, and thus can be precomputed after GP modeling is complete. Note that the dimension of \( \mathbf{r} \) is equal to the number of samples (i.e., \( N \)) used in GP modeling. Therefore, decreasing \( N \) becomes necessary, such that efficient Latin Hypercube Design (LHD) is always used [9].

**IV. PROPOSED ADJUSTMENT-BASED MODEL**

As illustrated in Section III, we can utilize LHD to reduce \( N \), the required number of samples to obtain a desired modeling accuracy. However, \( N \) might be still large, especially when the dimension of variability is rather high. Moreover, larger \( N \) indicates higher evaluation complexity. Therefore, in this paper, we propose a novel adjustment-based model to characterize the circuit timing under high-dimension of variability. This adjustment-based model has recently been proposed in the statistics community and can significantly reduce the number of collected samples [8]. To the best of our knowledge, we are the first to introduce its usage and discuss its benefits in the CAD area for SST A.

**A. Overview**

Fig. 1 shows the flow of the adjustment-based modeling framework. Instead of directly modeling the circuit timing with respect to the process variations using GP model, we build a model which can adjust an approximate circuit timing (subject to error) into an accurate one. The intuition behind this modeling framework is the following principle:

**Principle of Accuracy:** The approximate and accurate simulations should generate different output values for the same input sample. However, the approximate simulation can still reflect the relationship between the inputs and their corresponding accurate outputs. This reflection is subject to error but is sufficient to capture the complexities between the input and outputs to some extent.

After satisfying the above principle, building a model, which adjusts an approximate circuit timing into an accurate one, becomes a simpler task [8]. In other words, the task of building an adjustment-based model requires much fewer samples than direct GP modeling, which indicates the reduction of the evaluation complexity. Therefore, the adjustment-based model can be more suitable for modeling in presence of high dimension of variability and can perform much better than GP model.

As shown in Fig. 1, in order to build an adjustment-based model, apart from generating accurate reference samples for the circuit timing, we also need to generate samples for approximate circuit timing, which should be able to satisfy the principle of accuracy. These approximate samples can be obtained, for example, based on relaxing assumptions/models and can be generated much faster than the accurate ones, as we further illustrate. Moreover, even though the approximate samples are generated based on relaxed assumptions, they can still capture some nonlinear relationship between the parameter variations and the actual circuit timing.

**B. Adjustment-Based Modeling Process**

In order to build the adjustment-based model, we first generate a small number (i.e., \( N_a \)) of input samples, and obtain their corresponding approximate and accurate observations. For the convenience of notations, we denote the approximate and accurate observations by \( y_A(x) \) and \( y_A(x) \) respectively. Then, we utilize the GP modeling procedures introduced in Section III to build the adjustment function between \( y_A(x) \) and \( y_A(x) \):

\[
y_A(x) = \rho(x) y_A(x) + \delta(x),
\]

where

\[
\rho(x) = \rho_0 + \sum_{i=1}^{d} \rho_i x_i,
\]

\( x_i \) is the \( i \)th element of \( x \in \mathbb{R}^d \) for \( i = 1, 2, \ldots, d \), and \( \delta(\cdot) \) is a Gaussian Process with variance \( \sigma_\delta^2 \) and \( \theta_\delta \), as defined in Section III. Note that the parameters \( \sigma^2, \theta_\delta \) and \( \delta_0 \) in Eqs. (7) and (8) all need to be estimated.

Eq. (7) resembles the GP model in Eq. (1) in the following way

\[
\begin{align*}
\{ & f_i(x) = y_A(x), \beta_i = \rho_0, \quad \text{when } i = 0, \\
& f_i(x) = y_i(x) \cdot x_i, \beta_i = \rho_i, \quad \text{when } i = 1, 2, \ldots, d, \\
& f_i(x) = 1, \beta_i = \delta_0, \quad \text{when } i = d + 1.
\end{align*}
\]

Consequently, we can similarly obtain the unknowns (i.e., \( \sigma^2, \theta_\delta, \rho_0, \delta_0 \) by following the estimation procedures for GP in Section III. We can then build the final adjustment-based model, and accurately compute the output performance for any new input vector using Eq. (6).

**Remark:** When Eq. (7) has all \( \rho_0 \)’s equal to 0, it is equivalent to the original GP form in Eq. (1), where \( f_i \) is a constant. Therefore, we can expect not to loose accuracy by applying our adjustment-based model compared to building GP model based on accurate simulation.
V. Adjustment-Based Model for SSTAs

We propose two methods to generate approximate variation-aware static timing analysis efficiently. We obtain accurate samples using Monte Carlo based SSTAs on the entire circuit (MC-SSTA-EC).

A. Approximate Timing Using MC-SSTA-CS

The first way to obtain approximate circuit timing is to conduct MC simulation over a Critical Subcircuit, which we refer to as MC-SSTA-CS. In presence of process variations, the critical path will not be a constant path. However, we still extract the top critical paths of a circuit in the same way as under the nominal case. We first traverse the circuit topologically from the POs towards the PIs. At arrival times in the circuit, two atomic operations, namely summation and interconnect parameter variations. Particularly, in computing the combination of Gaussian random variables representing different device times of other fanins are close to this maximum arrival time by a given threshold of $\tau\%$, we also include these fanins in this critical subtree. We apply Monte-Carlo based SSTA over the extracted critical subcircuit and record these as the approximate observations.

According to the Principle of Accuracy, the circuit delays obtained from MC-SSTA-CS should reflect the relationship between the parameter variations and the accurate circuit timing. To illustrate this principle, we selected circuit S39532 as an example. Fig. 2(a) shows the similarity between the circuit timing obtained using MC-SSTA-CS (i.e., $Y$-axis) and that using MC-SSTA-EC (i.e., $X$-axis). These samples were generated for 42 parameter variation cases as described in Section VI. We also plot the line $y = x$ to indicate the similarity between these two sets of circuit timing. As seen in this figure, in some variation cases, the approximate circuit timing is equal to accurate one (i.e. circuit timing is indeed determined by the critical subcircuit extracted under the nominal case). However, in other variation cases, the critical subcircuit does not determine the circuit timing, but they still exhibit a correct pattern of increase or decrease in the circuit timing. This is likely due to the common (global) source of variations which result in identical increase or decrease in the parameters of all the gates and interconnects for each variation sample. Furthermore, different gate types have similar range of sensitivities to different parameters.

B. Approximate Timing Using LG-SSTA

The second way to obtain approximate circuit timing, which can follow the Principle of Accuracy, is to use Linear Gaussian SSTAs (i.e., LG-SSTAs) [11]. In LG-SSTA, the arrival time at each gate is approximated to have Gaussian distribution. These arrival times, along with the individual gate delays are expressed as a linear combination of Gaussian random variables representing different device and interconnect parameter variations. Particularly, in computing the arrival times in the circuit, two atomic operations, namely summation and maximization should be computed. Summation operation is computed exactly. However, maximum of Gaussian arrival times is approximated as a Gaussian arrival time which can be represented in the same linear form. This approximation is based on Clark’s approach. Given two Gaussian arrival times $X$ and $Y$, we define $Z$ as maximum of $X$ and $Y$, and approximate it as:

$$Z \triangleq \max(X, Y) \approx tX + (1 - t)Y + \Delta,$$  \hspace{1cm} (9)

where $t$ is the tightness probability defined as

$$t \triangleq \int_{-\infty}^{\lambda} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)dx,$$  \hspace{1cm} (10)

and $\Delta$ is an additional random variable to ensure that this approximation has the same mean and covariance as $Z$.

In Eq. (10), $\lambda$ is defined as $\frac{\mu_{X-Y}}{\sigma_{X-Y}}$, where $\mu_{X-Y}$ and $\sigma_{X-Y}$ denote the mean and standard deviation of $X - Y$, respectively. The Gaussian approximation in Eq. (9) makes maximum operation easy to implement, but suffers from high error. The authors in [12] pointed that the Clark’s approximation in LG-SSTA introduces high error when $X$ and $Y$ have very similar mean but very different variances, and when $X$ and $Y$ have very similar mean but highly negative correlation. Also, LG-SSTA is based on moment matching technique and can only guarantee the correctness of the statistical information, such as mean and covariance at each maximization but the actual circuit timing is under variations and subject to error.

Despite the intrinsic problems in LG-SSTA illustrated above, we can still use LG-SSTA to obtain approximate circuit timing. First, LG-SSTA is the fastest known SSTAs with a complexity linear in the number of gates in the circuit; for each node with multiple fanins, it only needs to approximate the mean and variance of the maximum of their arrival times. Therefore, using LG-SSTA for approximate timing will not add much overhead for our modeling. Second, the LG-SSTA is subject to error; for example, the mean of the circuit timing obtained from LG-SSTA is proven to always be an underestimation of the exact mean. However, LG-SSTA can still capture the output/input relationship Fig. 2(b) plots the circuit timing for S39532 using LG-SSTA in $Y$-axis and that using MC-SSTA-EC in $X$-axis. For LG-SSTA, we obtain the samples by doing Monte Carlo simulation on the linear Gaussian expression of timing obtained using LG-SSTA. Even though the circuit timing using LG-SSTA suffers from high error, it can still reflect the output/input relationship compared to the accurate simulation.

VI. Simulation Results

We synthesized ISCAS89 and ITC99 benchmark suites with 90nm TSMC library. Using the spatial correlation model in [5], we assumed 42 independent random variables (RVs) with Gaussian distribution to represent parameter variations in channel length and zero-body threshold voltage. All these RVs have their standard deviation to be 7% of their mean. We also assumed a linear gate delay model in terms of these RVs.

Evaluation of Adjustment-Based SSTAs Using Critical Subcircuit Extraction: To evaluate the performance of our proposed adjustment-based SSTAs using critical subcircuit, we implemented 3 types of SSTAs, which are based on Quadratic model (QR), GP model, and Adjustment-based model (ADI), respectively. Note that by adjusting $t$, we extracted a small critical subcircuit in ADI. In all these SSTAs approaches, we first generated a large number of samples for the 42 RVs and their corresponding circuit timing. We then fitted the samples to the model used in each approach to get a circuit timing expression in terms of these RVs. Particularly, we considered 4 variations of these 3 types of SSTAs, which are listed as follows:
**TABLE 1**

<table>
<thead>
<tr>
<th>BENCH</th>
<th>5%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>95%</th>
<th>Avg Err = 10.67%</th>
</tr>
</thead>
<tbody>
<tr>
<td>S35932</td>
<td>-1.554</td>
<td>-0.541</td>
<td>-1.159</td>
<td>-2.182</td>
<td>-3.995</td>
<td>23.50X</td>
</tr>
<tr>
<td>S3847</td>
<td>0.335</td>
<td>-0.567</td>
<td>-0.609</td>
<td>-0.551</td>
<td>-0.994</td>
<td>143.39X</td>
</tr>
<tr>
<td>S38584</td>
<td>-0.101</td>
<td>-0.207</td>
<td>-0.234</td>
<td>-0.283</td>
<td>-0.311</td>
<td>301.31X</td>
</tr>
<tr>
<td>b1b</td>
<td>1.150</td>
<td>1.300</td>
<td>1.151</td>
<td>1.154</td>
<td>1.156</td>
<td>11.72X</td>
</tr>
<tr>
<td>b20</td>
<td>2.000</td>
<td>2.221</td>
<td>2.573</td>
<td>2.611</td>
<td>2.773</td>
<td>38.31X</td>
</tr>
<tr>
<td>b31</td>
<td>1.149</td>
<td>1.149</td>
<td>1.186</td>
<td>1.202</td>
<td>1.193</td>
<td>62.43X</td>
</tr>
<tr>
<td>b22</td>
<td>-0.346</td>
<td>0.232</td>
<td>0.258</td>
<td>0.621</td>
<td>-0.311</td>
<td>1.186</td>
</tr>
<tr>
<td>Ave [Err]</td>
<td>0.984</td>
<td>0.888</td>
<td>1.026</td>
<td>1.229</td>
<td>1.078</td>
<td>87.47X</td>
</tr>
</tbody>
</table>

**FIG. 3.** Histograms of Relative error for S35932 (a): QR-UNI; (b): QR-LHD; (c): GP-200; (d): ADJ-CS-200

**Evaluation of Adjustment-based SSTA Using LG-SSTA:** We implemented the adjustment-based SSTA based on LG-SSTA as follows. We first generated 200 LHD samples for the process variations and obtained their corresponding approximate and accurate circuit delays using LG-SSTA and MC-SSTA-EC, respectively. We then built the adjustment model using “R” and denote this SSTA approach by ADJ-LG-200. In order to evaluate the accuracy of ADJ-LG-200, we make comparison with MC-SSTA-EC with 10,000 samples. For S35932 benchmark, ADJ-LG-200 had errors of 3.44%, 2.07%, 1.14%, 0.51% at CDF percentile points 25%, 50%, 75%, 95%, respectively.

**Discussion on Efficiency:** We draw the following conclusions: 1. The simulation runtime of ADJ-CS-200 was nearly proportional to the size of the extracted subcircuit which was much lower than the simulation runtime for the total circuit size. The runtime to collect the approximate circuit timing using MC-SSTA-CS was on average about 5% of that of MC-SSTA-EC. Also, we required collection of only 200 approximate and accurate samples. Similarly, the runtime of ADJ-LG-200 was very fast; this is because Clark’s method is known to have linear runtime complexity in the number of nodes in the timing graph and is the fastest SSTA technique [11].

2. To build the adjustment-based model, Eq. (3) can be efficiently solved using “R”. Note that this model can be directly used for robust analysis under other different PDFs of process variations.

**VII. Conclusions**

We proposed a novel adjustment-based model for SSTA which translates an approximate variation-aware circuit timing into an accurate one. We proposed two ways to generate approximate circuit timing to satisfy the principle of accuracy. For large circuits and with high dimension of variability, our improvement is mainly due to significant reduction in the number of samples using adjustment-based modeling over conventional quadratic and GP modeling.

**REFERENCES**