A Linear-Centric Simulation Framework for Parametric Fluctuations

Emrah Acar, Lawrence T. Pileggi[1] and Sani R. Nassif

IBM Austin Research Labs
11400 Burnet Rd
Austin, TX 78758
{emrah,nassif}@us.ibm.com

[1]*Dept. of ECE., Carnegie Mellon Univ.*
5000 Forbes Ave
Pittsburgh, PA 15213
pileggi@ece.cmu.edu

Abstract

The relative tolerances for interconnect and device parameter variations have not scaled with feature sizes which have brought about significant performance variability. As we scale toward 10nm technologies, this problem will only worsen. New circuit families and design methodologies will emerge to facilitate construction of reliable systems from unreliable nanometer scale components. Such methodologies require new models of performance which accurately capture the manufacturing realities.

Recently, one step toward this goal was made via a new variational reduced order interconnect model that efficiently captures large scale fluctuations in global parameter values. Using variational calculus the linear interconnect systems are represented by analytical models that include the global variational parameters explicitly. In this work we present a framework which extends the previous work to a linear-centric simulation methodology with accurate nonlinear device models and their fluctuations. The framework is applied to generate path delay distributions under nonlinear and linear parameter fluctuations.

1. Introduction

With decreasing MOS transistor geometries for DSM (Deep Submicron) technologies, the influence of fluctuations in process parameters during manufacturing becomes increasingly important since process tolerances are not proportionally scaled with geometries. Typically, the effect of process variations are captured by a set of worst-case device model parameters and the circuit performance is evaluated at these worst-case corners. However, the dominant interconnect in DSM technologies complicates the feasibility of a worst-case corner method by increasing the dimensionality of the problem. Moreover, the worst-case corner methods are known to create overly pessimistic results and in sub-optimal designs. To synthesize reliable fabrics with tomorrow's fabrication technologies requires new models and analyses that account for variations in transistors and interconnect. For future designs, new models are required to capture the parameter fluctuations and manufacturing realities. The nominal and extreme performance evaluations need to be replaced by statistical frameworks that evaluate the stochastic nature of the system performance more accurately.

This work is supported in part by the MARCO/DARPA Gigascale Silicon Research Center (http://www.gigascale.org) and IBM Austin Research Labs. Their support is gratefully acknowledged.

In this work, we develop an efficient framework to assess more realistic performance distributions and extreme case scenarios. We demonstrate our methodology by incorporating variational interconnect models into transistor-level simulation with accurate nonlinear device models and their parameter fluctuations.

In a previous work, a variational reduced order interconnect model was reported to capture global parameter variations in [1]. The efficiency of the linear-centric variational models was demonstrated in statistical analysis of the skew performance of a clock grid from a gigahertz microprocessor[2][3]. The variational interconnect models, developed with variational calculus, relates global interconnect parameters to compact representations of interconnect models. Generally, the projectional methods, PACT[4], PRIMA[5] are used to precharacterize the variational model with a design of experiments. During the library pre-characterization, the projectional reduction algorithms retain the passive nature of the interconnect in the macromodel, however the variational versions of such algorithms are unable to preserve passivity and stability. Without preserving the passive nature of the interconnect, subsequent analyses with nonlinear devices can cause instability, as discussed later in this paper.

The proposed framework is embedded with a waveform evaluation engine TETA[6][7] which was developed for use in timing analysis. TETA provides efficient runtime accuracy trade-offs in handling nonlinear devices and offers several benefits for statistical analyses. Unlike macromodeling approaches, TETA employs interconnect-friendly linear-centric device models without sacrificing accuracy.

Variational analysis of the linear interconnect with nonlinear devices and their associated parameter fluctuations is a formidable task. The complexity of the linear models and the requirements for nonlinear analysis limit the advantages of current methods. In this paper, we propose a linear-centric simulation framework, which incorporates variational interconnect models into the aforementioned transistor-level

waveform evaluation methodology, TETA. Our framework employs efficient models for nonlinear devices and obviates the need for passive interconnect models. The framework also includes proper statistical methods to perform statistical analysis on path delays. We demonstrate the accuracy and efficiency of the framework by various examples.

2. Variational Reduced Order Model

Reduced order modeling constructs a macromodel for a large linear system with an MNA formulation:

$$\text{Original System: } (\boldsymbol{G}+s\boldsymbol{C})\boldsymbol{v}(s) = \boldsymbol{i}(s) \tag{1}$$

where $\boldsymbol{G}$ and $\boldsymbol{C}$ are the admittance and susceptence matrices. The vectors, $\boldsymbol{v}$ and $\boldsymbol{i}$ are the node voltages and currents:

$$\boldsymbol{v} = \begin{bmatrix} \boldsymbol{v}_p^T & \boldsymbol{v}_{int}^T \end{bmatrix}^T \qquad \boldsymbol{i} = \begin{bmatrix} \boldsymbol{i}_p^T & \boldsymbol{i}_{int}^T \end{bmatrix}^T \tag{2}$$

where the N_p dimensional $\boldsymbol{v}_p$ and $\boldsymbol{i}_p$ vectors are associated with the port nodes, and $\boldsymbol{v}_{int}$, $\boldsymbol{i}_{int}$ are for internal nodes. Note that for most of the interconnect problems, the linear system is driven at its ports, hence $\boldsymbol{i}_{int} = \boldsymbol{0}$.

When global interconnect parameters fluctuate, the circuit matrices can be written in variational forms:

$$\boldsymbol{G}(\boldsymbol{w}) = \boldsymbol{G_0} + \boldsymbol{dG_1}w_1 + \boldsymbol{dG_2}w_2 \tag{3}$$

$$\boldsymbol{C}(\boldsymbol{w}) = \boldsymbol{C_0} + \boldsymbol{dC_1}w_1 + \boldsymbol{dC_2}w_2 \tag{4}$$

Variational reduced order modeling finds a compact representation of the interconnect macromodel by including the variation parameters. In doing so, it creates a macromodel which can be efficiently evaluated in terms of the global interconnect parameters. As described in [1], projection based reduced order modeling methods (PACT, PRIMA) in a variational manner, form the variational reduced order models as:

$$\boldsymbol{G_r}(\boldsymbol{w}) = \begin{bmatrix} \boldsymbol{A}(\boldsymbol{w}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}(\boldsymbol{w}) \end{bmatrix} \quad \boldsymbol{C_r}(\boldsymbol{w}) = \begin{bmatrix} \boldsymbol{B}(\boldsymbol{w}) & \boldsymbol{R}(\boldsymbol{w}) \\ \boldsymbol{R}^T(\boldsymbol{w}) & \boldsymbol{E}(\boldsymbol{w}) \end{bmatrix} \tag{5}$$

where $\boldsymbol{G_r}(\boldsymbol{w})$ and $\boldsymbol{C_r}(\boldsymbol{w})$ are the reduced order matrices for a particular parameter sample $\boldsymbol{w}$. The related matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{D}, \boldsymbol{E}, \boldsymbol{R}$ are computed by using the pre-characterized model library. The resulting MNA formulation of the newly constructed reduced order model:

$$\text{Reduced System: } (\boldsymbol{G_r}(\boldsymbol{w}) + s\boldsymbol{C_r}(\boldsymbol{w}))\boldsymbol{v_r}(s) = \boldsymbol{i_r}(s) \tag{6}$$

relates a new set of voltage and current vectors:

$$\boldsymbol{v_r} = \begin{bmatrix} \boldsymbol{v}_p^T & \boldsymbol{v}_{nodes}^T \end{bmatrix}^T \qquad \boldsymbol{i_r} = \begin{bmatrix} \boldsymbol{i}_p^T & \boldsymbol{i}_{nodes}^T \end{bmatrix}^T. \tag{7}$$

$\boldsymbol{v}_{nodes}$ and $\boldsymbol{i}_{nodes}$ are related to the new set of internal nodes that mimic the original linear system. Similar to the original formulation, $\boldsymbol{i}_{nodes} = \boldsymbol{0}$. The number of internal nodes in the (6) is much less than the original dimension of the system making the macromodel more efficient.

A practical implementation of variational reduced order modeling can be made via variational Krylov vectors in a projection based method:

$$\boldsymbol{X}(\boldsymbol{w}) = \boldsymbol{X_0} + \boldsymbol{dX_1}w_1 + \boldsymbol{dX_2}w_2 \tag{8}$$

Following the PRIMA algorithm, the reduced order macromodel matrices can be computed via congruence transformations. Using the variational forms for Krylov vectors, the projection based methods compute the first order variational admittance matrix for a single parameter as:

$$\boldsymbol{G_r}(w_1) = \boldsymbol{X}^T(w_1)\boldsymbol{G}(w_1)\boldsymbol{X}^T(w_1) \tag{9}$$

$$\boldsymbol{G_r}(w_1) = (\boldsymbol{X_0} + \boldsymbol{dX_1}w_1)^T(\boldsymbol{G_0} + \boldsymbol{dG_1}w_1)(\boldsymbol{X_0} + \boldsymbol{dX_1}w_1) \tag{10}$$

$$\begin{aligned} \boldsymbol{G_r}(w_1) = {} & \boldsymbol{X_0^T G_0 X_0^T} + \\ & w_1 \times (\boldsymbol{dX_1^T G_0 X_0} + \boldsymbol{X_0^T dG_1 X_0} + \boldsymbol{X_0^T G_0 dX_1}) + \boldsymbol{h \cdot o \cdot t} \end{aligned} \tag{11}$$

As seen in (11), the first-order variational admittance macromodel is not a congruence transformation which is essential for macromodel passivity. If the impractical higher order terms were included for proper congruence transformation, the passivity would become provable. However, it is impractical to store and apply the higher order variational matrices. For efficiency and accuracy reasons, these higher order terms are often ignored. Therefore, unlike the nominal case, the practical variational reduced order models do not generally preserve passivity and stability. Hence their interfaces with general transistor-level analysis tools have potential divergent behavior. As opposed to stability checks, a practical test for passivity is not available and remains as an open problem. However, one could easily check for macromodel stability by monitoring the poles in the frequency domain pole/residue description[8]. Furthermore, unstable macromodels can be forced to be stable by removing their unstable modes and compensating the stable modes to preserve the dc conditions. Such useful heuristics are not available for non-passive macromodels due to the theoretical difficulties in detection and correction strategies.

As we need new models and methods to capture manufacturing realities in greater detail, we need to combine nonlinear device models and their parameter fluctuations with variational interconnect models. Two major considerations need to be addressed. First, as we described above, the variational interconnect models do not maintain passivity making linear models incompatible for simulation with nonlinear devices. Therefore, we have to obviate the need for passive macromodels in our framework. Second, the inclusion of nonlinear device models and their parameter fluctuations is a formidable task due to its extreme complexity. A practical approach can be taken with using linear-centric models that can sustain limited parametric variations within a fast waveform evaluation engine[9].

In the next section, we briefly describe the proposed simulation framework that analyzes digital integrated circuits with variational interconnect models.

3. Linear-Centric Simulation Framework

3.1. Need For Passivity

The general time-domain simulators are composed of two major techniques: numerical integration and nonlinear algebraic equation solution. They often employ Newton-based nonlinear solvers which linearize the nonlinear circuit elements and iteratively solve the linearized system. For digital circuits that are made of nonlinear drivers and a linear load, the linearization of the nonlinear circuit elements transforms the nonlinear driver into a Norton equivalent of which the parameters vary for every

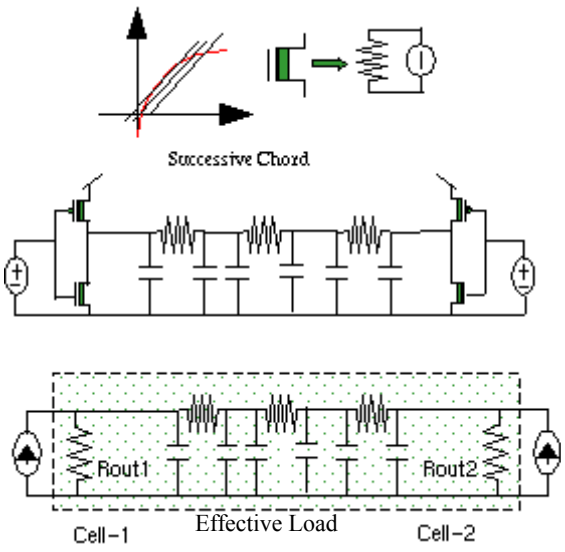

Figure 1. TETA's linear-centric device models and the interconnection of chord models with the linear load to form the effective load. Effective load requires to be stable.

operating point. When interfaced with a nonpassive linear load model, the effective load seen by the Norton current source can be unstable, resulting a possible divergence. This simple explanation clearly describes the need for passive linear models for conventional circuit simulators. The simulation of linear models with nonlinear devices using a general Newton based linearization approach strongly needs a passive linear model [7].

3.2. TETA: Linear-Centric Simulation Engine

TETA is a fast transistor-level timing simulation engine with almost SPICE accuracy[6][7][9]. Unlike other timing simulators, TETA uses Successive Chords (SC) technique to solve the associated nonlinear system of equations creating fixed affine linearizations for nonlinear circuit elements. The SC method and splitting the linear and nonlinear portions of the circuit, create constant impedances seen as each nonlinear element during solver iterations. The linearized impedances can be lumped into a final Norton equivalent model. In TETA, this is done by a linear centric device model which is called as chord model [6][7][9]. The chord models are chosen prior to the analysis and are used in every nonlinear iteration.

TETA was developed to evaluate the delay and output waveforms of multi-port coupled logic stages whose load models include large number of linear circuit elements. Its effectiveness is extremely significant when the complexity of waveform analysis is due to these linear circuit elements.

When applied to simulate strongly connected components coupled via a large multiport interconnect, TETA does not require a strict passivity condition for the linear models. As shown in Figure 1, it solves the *passivity bottleneck of nonlinear simulation* by incorporating the aggregate Norton conductance of the nonlinear devices (resulting from the chord models) with the linear load prior to simulation. This enables the use of more efficient stable macromodeling algorithms which do not maintain passivity. For simulating the variational interconnect models, it is seemingly impossible to maintain passivity. Therefore, the use of the SC method and linear-centric device models in TETA are crucial for our framework to accommodate nonpassive variational interconnect models in nonlinear simulation.

3.3. Stable Variational Reduced Order Models

In this work, we target variational delay analysis of the paths made of logic stages that may include coupling. For that purpose, we derive our models for multiple nonlinear drivers that are coupled via large multi-port interconnect. Simulation of such structures is the major problem in DSM's timing verification.

To solve the stability/passivity problem of the variational interconnect models, we apply a TETA-like approach that includes the impact of a nonlinear driver as a lumped Norton equivalent model with a constant output conductance. Note that the SC-based linearization of nonlinear drivers (chords) can be independent of the interconnect and device model parameters, which are subject to fluctuations. Therefore, in a statistical analysis, the output conductances of the SC-linearized drivers, $\boldsymbol{G_{out}}$ remain constant. Therefore, we may use the same chord models under nonlinear device parameter variations. Even if the device model parameters fluctuate, the output conductance in the SC-based linearizations do not need to be changed. Therefore the chord models driven for nominal model parameters can be used for different types of device and interconnect models efficiently.

In our approach, the effective linear load seen by Norton current sources of nonlinear drivers (refer to Figure 1) is transformed into a variational reduced order form. This process explicitly includes the output conductances seen for the nonlinear drivers, $\boldsymbol{G_{out}}$, into the variational macromodel. Hence, the MNA formulation for the effective load can be written as:

$$\text{New System: } (\boldsymbol{G}(\boldsymbol{w}) + \boldsymbol{G_{sc}} + s\boldsymbol{C}(\boldsymbol{w}))\boldsymbol{v}(s) = \boldsymbol{i}(s) \qquad (12)$$

where $\boldsymbol{G_{sc}}$ is a diagonal matrix, and its first N_p diagonal entries are equivalent to the output conductances seen for the SC-based linearizations of nonlinear drivers. TETA provides these values during LU factorization of the admittance matrix. We have to note that the inclusion of $\boldsymbol{G_{sc}}$ updates all of the variational matrices and not just the diagonal of $\boldsymbol{A}(\boldsymbol{w})$.

In the first step, the output conductances, $\boldsymbol{G_{out}}$ are computed. Since their values depend on the topology of the nonlinear driver and timestep resolution of the analysis, $\boldsymbol{G_{out}}$ values can be computed for each driver in the library. Then the variational reduced order modeling algorithm creates the macromodel of the effective linear load $\boldsymbol{Y_r^{lin}}(s) = \boldsymbol{G_r^{lin}}(\boldsymbol{w}) + s\boldsymbol{C_r^{lin}}(\boldsymbol{w})$. These matrices can be used to create a subcircuit description or they can be directly stamped into a nonlinear system. However, as we pointed out in the previous section, they do not guarantee passivity and therefore may cause divergence in solution.

To extract the frequency domain behavior of the linear macromodels, it is more efficient to apply a transformation to obtain a pole/residue description. The impedance matrix for the effective linear load, $\boldsymbol{Z_r^{lin}}(s)$ is

$$\boldsymbol{Z^{lin}}(s) = [v_i(s)/I_j(s)] = [Z_{ij}^{lin}(s)] \qquad (13)$$

The numerical procedures to obtain the effective impedance macromodel are summarized below:

$$Z_{ij}^{lin}(s) = \boldsymbol{e_i^T}(\boldsymbol{G_r^{lin}} + s\boldsymbol{C_r^{lin}})^{-1}\boldsymbol{e_j} \qquad i,j = 1\ldots N_p \qquad (14)$$

$$(\boldsymbol{Y_r^{lin}})^{-1} = (\boldsymbol{I} - s\boldsymbol{T^{lin}})^{-1}(\boldsymbol{G_r^{lin}})^{-1} \qquad (15)$$

$$\boldsymbol{T^{lin}} = -(\boldsymbol{G_r^{lin}})^{-1}\boldsymbol{C_r^{lin}} = \boldsymbol{SDS^{-1}} \qquad (16)$$

$$Z_{ij}^{lin}(s) = \boldsymbol{e_i^T}(\boldsymbol{I} - s\boldsymbol{SDS^{-1}})(\boldsymbol{G_r^{lin}})^{-1}\boldsymbol{e_j} \qquad (17)$$

$$Z_{ij}^{lin}(s) = \boldsymbol{e_i^T S}(\boldsymbol{I} - s\boldsymbol{D})\boldsymbol{S^{-1}}(\boldsymbol{G_r^{lin}})^{-1}\boldsymbol{e_j} \qquad (18)$$

$$\mu_{ij} = \boldsymbol{e_i^T S} \qquad \upsilon_{ij} = \boldsymbol{S^{-1}}(\boldsymbol{G_r^{lin}})^{-1}\boldsymbol{e_j} \qquad d_{kk} = \boldsymbol{e_k^T D e_k} \qquad (19)$$

$$Z_{ij}^{lin}(s) = \sum_{k=1}^{N}(\mu_{ijk}\upsilon_{ijk})/(1 - sd_{kk}) \qquad i,j = 1\ldots N_p \qquad (20)$$

Since the matrices for the reduced order model are relatively smaller than those of the original system, the steps above are not very expensive. Furthermore, eigen-decomposition of $\boldsymbol{T^{lin}}$ is done only once and reused for each entry in $\boldsymbol{Z^{lin}}$.

The real advantage of a pole/residue transformation is the ability to conduct practical strategies to avoid macromodel instability. Macromodel instability manifests itself with positive poles in the pole/residue representation. The poles with positive real parts are mainly due to the high frequency components, near-singularities, approximation error and ill-conditioned numerical computations. With very small residues, they generally do not possess significant information on the system behavior. The following practical two-step strategy is very effective for filtering such unstable modes. The first step eliminates the real positive poles and the second step adjusts the stable residues with a common multiplier factor, β, in order to match the dc (first moment) behavior of the original system.

$$Z_{ij}^{lin}(s) = \sum_{Re(p_k)\le 0} r_k/(s - p_k) + \sum_{Re(p_k)>0} r_k/(s - p_k) \qquad (21)$$

$$Z_{ij}^{lin,\,stable}(s) = \sum_{Re(p_k)\le 0} \hat{r}_k/(s - p_k) \qquad \hat{r}_k = r_k\beta \qquad (22)$$

$$\beta = (\sum_k r_k/p_k)/(\sum_{Re(p_k)\le 0} r_k/p_k) \qquad (23)$$

This simple procedure captures the dominant modes and removes instability that may be caused by the variational modeling. Using the corrected stable reduced order models in pole/residue form (22), we may conduct the time-domain simulation with TETA.

The flow of creating variational reduced order models is depicted in Table 1. For more technical details and an alternative approach, the reader can refer to [9]. Next, we discuss about the application of analysis of path delay variability.

Construction:

1. Calculate/retrieve output conductances of SC-based linearizations of nonlinear drivers, $\boldsymbol{G_{out} = diag(G_{SC})}$
2. Include driver conducatnces with multiport load to create effective load, $\boldsymbol{G^{lin} = G + G_{SC}}$
3. Create variational reduced order model library based on the effective load model: $\boldsymbol{(G^{lin}, C) \rightarrow (A, B, D, E, R(w))}$

Evaluation of a particular global interconnect parameter:

1. Apply the variational algebra[1] to obtain the reduced order model: $\boldsymbol{G_r(w), C_r(w)}$
2. Convert the reduced order model into the frequency domain: $\boldsymbol{Z^{lin}}(s)$
3. Filter the unstable pole/residues and make necessary corrections: $\boldsymbol{Z^{lin}}(s) \rightarrow \boldsymbol{Z^{lin,\,stable}}(s)$
4. Simulate the load with the nonlinear drivers using TETA.

Table 1. Flow for variational reduced order modeling

4. Calculating the Path Delay Variations

Among the important performance metrics for digital integrated circuits is the critical path delay. A critical path is defined as the performance limiting path that has the longest (or shortest) sensitizable delay between its primary inputs and primary outputs. Like others, it is a string of a number of stages which can include the interconnected wires and effective neighboring lines. In DSM technologies, the inclusion of the electrical activity in the local vicinity of the signal path into timing analysis (signal integrity) can be imperative.

The accuracy in the statistical delay analysis is crucial due to the impact of the noise from modeling errors. To alleviate this impact, in current design practice, the critical paths delay statistics are often evaluated by circuit simulation[10] with proper input vectors. To predict the timing yield of the critical path delay, a large number of simulations are required. General purpose simulators suffer from several reasons. First, these methods do not model the interconnect efficiently and becomes very slow for large number of linear circuit elements, even for a nominal simulation. Second, the dynamic nature of the path elements, effects for simultaneous switching and parasitic coupling become extremely hard to be included in the entire path simulation. Third, some general purpose simulators like SPICE, could not exploit the natural spatial and temporal latency of the critical path and cost more runtime.

In this section, we describe two methods to evaluate the critical path delay variations using our linear-centric circuit simulation framework. Unlike gate-level macromodeling approaches which surrender accuracy for efficiency, we conduct an accurate simulation strategy to capture the true nature of the performance variations with as little numerical error as possible.

4.1. Statistical Methods

We first review the major statistical methods, which are implemented in our framework.

4.1.1 Principal Component Analysis

Integrated circuit device and wire models are often complex and include a significant number of model parameters. These model parameters exhibit spatial and temporal correlation since they depend on a few common factors in the processing and operating environment. Therefore, it is wiser to conduct a Principal Component Analysis (PCA) prior to a statistical sampling. PCA discovers the independent factors which explain the majority of the parameter variations. As an example, [11] reports that the variations in BSIM3 device model parameters, of 60 device model parameters, can be explained by 10 uncorrelated, factors. In PCA, these factors are simply the linear combinations of the varying parameters and one could retrieve

the original model parameters using a by-product reverse transformation. Since PCA reduces the dimensionality of the variational problems and the corresponding sample size, it is implemented in many statistical analysis frameworks. Additional variance-reduction methods are also available to enhance its performance.

4.1.2 Monte-Carlo Methods

A rigorous way to predict the statistical distribution of the performance is Monte-Carlo analysis. Considering the variability in device and wire models, a Monte-Carlo analysis creates a sample of parameters that are subject to variation and exhaustively evaluates the performance for each sample. Performing a Monte-Carlo simulation requires to know the nature of variation sources in advance. The procedure can be improved by advanced sampling techniques and selecting an uncorrelated set of main variation sources using PCA. PCA provides a more manageable set of uncorrelated factors for sampling and reduces the analysis errors. While the global wire parameters are subject to fluctuations, similar considerations need to be considered. However one could argue that the global wire parameters are less likely correlated and they could be directly handled in Monte-Carlo analysis.

4.1.3 Gradient Analysis

Gradient Analysis[12] (GA) is a more simple and computationally efficient technique that consists of a gradient analysis of the performance. This approach evaluates the standard deviation of the circuit performance with a linear model in terms of variation sources. If the performance, the path delay (D_{path}) is related to uncorrelated variation sources, w_l, $1 \le l \le N_W$, then the standard deviation of the performance can be written as:

$$\sigma_{Dpath} = \sqrt{\sum_{w_l} \sigma_{w_l}^2 (\partial D_{path} / \partial w_l)^2} \qquad (24)$$

If there is correlation between the variation sources, they can be added into the formulation. However, if the factors of a PCA analysis is used in GA, they will be uncorrelated. The GA method requires the sensitivity or gradient computation of the performance with respect to the variation sources. This generally requires less work than a Monte-Carlo analysis. The required gradients can be evaluated using finite difference methods.

4.2. Formulation

To formulate of the path delays, we need to abstract the waveform and the signal transfer models as [13]. In timing analysis, a typical switching signal waveform, $v(t)$ can be expressed as a waveform function:

$$v(t) = w(t, P_w) \qquad (25)$$

where P_w denotes the waveform function parameters. The most popular waveform function is the saturated ramp model, which has the slew and the start time for its parameters.

Given a set of loading and operating conditions, the input/ output behavior of a particular stage can be expressed in terms of the waveform function parameters. This input/output relation can be expressed in terms of waveform function parameters:

$$P_w^{out} = \Gamma_{stage}(P_w^{in}, L, E) \qquad (26)$$

where L represents the loading and E holds for device model parameters and other operational factors. With this abstraction, the intermediate output waveforms of a path can be defined with a recurrence relation and a special termination condition:

$$P_{w,k}^{out} = P_{w,k+1}^{in} = \Gamma_k(P_{w,k}^{in}, L_k, E_k) \qquad k = 1 \ldots N_{path} \qquad (27)$$

$$v_{Npath}(t) = w(t, P_{w,Npath}^{out}) \qquad (28)$$

N_{path} is the number of stages in the path, and $v_{Npath}(t)$ denotes the final output waveform. We assume that the input of the initial stage is given as $P_{w,0}^{in}$.

4.3. Calculating Path Delay Statistics

The variability in the path delay can be captured by a statistical analysis that performs multiple simulations of the logic stages to evaluate the path delay over a parameter space. One way to evaluate the path delay is simply a sequential evaluation of the stages in their topological order. This method evaluates the stage delay for pre-defined input vectors at a particular model parameter sample. An alternative method is the application of statistical analysis for the individual stages and combining the results to predict the path delay statistics. This section explains these two different approaches in more detail.

4.3.1 Monte-Carlo Approach

To evaluate the path delay statistics with a Monte-Carlo analysis, we employ the variational interconnect models and the linear-centric simulation engine for all stages on the critical path. The variation sources are classified and transformed into a more compact set via PCA analysis. Then each stage along the path is simulated in the topological order. The stage output waveform is propagated as the input of the subsequent stage.

One advantage of the proposed simulation framework is that the nonlinear chord models do not vary with the wire and device parameter variation and are kept constant throughout the Monte-Carlo process. Therefore, only a single characterization of the variational interconnect models is required. This is a significant advantage for stages with a large number of linear circuit elements, since the cost of the reduced order modeling may surpass the cost of nonlinear simulation. Additionally, we propagate a fine resolution waveform model which captures almost the exact waveform. The waveforms are represented by a piece-wise linear model that adaptively selects the breakpoints.

We have observed that the stage-by-stage path simulation is a lot more effective than the entire path simulation via traditional circuit simulators. A few orders of magnitude speedup can be obtained by exploiting the functional and temporal behavior of the path, and most importantly the single stage-simulation can easily incorporate the proper settings for environmental factors and coupling.

4.3.2 Gradient Analysis Approach

To compute the path delay statistics via the GA approach, we need to evaluate the sensitivity of the stage delays with respect to the variation sources. These sensitivities can be derived from the recurrence relation (28). To simplify the formulation, a saturated ramp waveform function can be used with parameters of 50%

arrival point M_k and slope S_k, for k^{th} stage input waveform, i.e.

$$P_{w,k} = (M_k, S_k). \quad (29)$$

Then, the input/output waveform relation can be expressed with two scalar functions as:

$$\begin{bmatrix} M_k^{out} \\ S_k^{out} \end{bmatrix} = \Gamma_k(M_k^{in}, S_k^{in}, L_k, E_k) = \begin{bmatrix} \Pi_k(P_{w,k}, L_k, E_k) \\ \Psi_k(P_{w,k}, L_k, E_k) \end{bmatrix} \quad (30)$$

For any variation source w_l, derivatives of the waveform model can be expressed recursively:

$$\frac{\partial M_k^{out}}{\partial w_l} = \frac{\partial \Pi_k}{\partial w_l} + \frac{\partial \Pi_k}{\partial M_k^{in}} \frac{\partial M_k^{in}}{\partial w_l} + \frac{\partial \Pi_k}{\partial S_k^{in}} \frac{\partial S_k^{in}}{\partial w_l} \quad (31)$$

A similar expression can be written for derivatives of S_k^{out} as well. Hence, the GA approach requires the derivatives of the stage input/output relations with respect to (M_k^{in}, S_k^{in}) and w_l, i.e. input waveform parameters and all variation sources. Besides, the derivatives of the waveform function parameters have to be propagated to next stages for further evaluation. Similar steps need to be taken for a more complex waveform function model with additional parameters.

In our framework, each derivative term in (31) is computed with a few simulations. For instance, $\delta\Pi_k / \delta S_k^{in}$ can be evaluated by perturbing S_k^{in} around its nominal value keeping all the other parameters constant. The derivatives with respect to the variation sources can be handled similarly. The described sensitivity computation is equivalent to the first-order expansion of the performance function around the nominal performance and model parameters. The total number of simulations required in the described procedure is generally less than the number of Monte-Carlo simulations. Once all the stages are completed, the path delay sensitivity is simply found by:

$$\partial D_{path} / \partial w_l = \partial M_{Npath}^{out} / \partial w_l \quad (32)$$

The standard deviation of the path delay can be found via (24).

After finishing a particular stage, we propagate the nominal waveform model along with the required derivative terms. The nominal waveform model is used to evaluate derivatives for variation sources. The waveform model, (M_k^{in}, S_k^{in}) is used to obtain the gradients for input waveform, i.e. $\partial \Pi_k / \partial M_k^{in}$.

The GA approach can be considered as a differential version of timing analysis of the stage. It processes the nominal arrival times and the derivative terms using the stage input/output relation. The GA approach is more accurate when the performance function has a linear relation with the parameter of interest. It relies on the first-order expansion of the performance function. For stage input/output relations of digital circuits, this condition is easier to achieve compared to more general analog circuits.

5. Results

5.1. Example 1

Our first example demonstrates the unstable nature of the variational reduced order models. The circuit is a coupled RC line shown in Figure 2. The symmetric two-port RC line is modeled in three segments of which the electrical model is given in Table 2. A normalized spatial parameter affects the electrical model as depicted. The element values are assumed to have linear relation with p. For simplicity, the two-port RC is transformed into a one-port RC load by shunting the second port with 100 ohms.

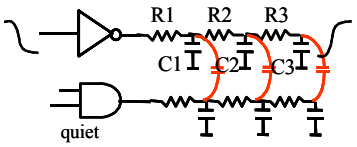

Figure 2. Circuit in Example 1

p	R1	R2	R3	C1	C2	C3	CC1	CC2	CC3
0	10	2	30	2pf	2pf	2pf	2pf	2pf	2pf
0.1	15	2	40	3pf	2pf	3pf	3pf	2pf	3pf

Table 2. Electrical model of the circuit in Example 1.

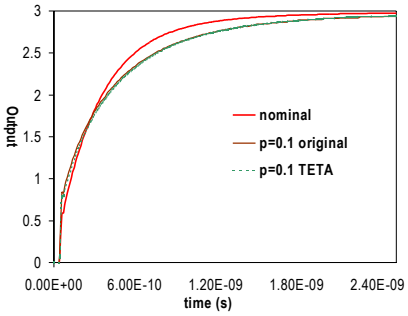

Figure 3. Result for nominal (p=0), extreme (p=0.1) and the reconstructed macromodel.

p	0.05	0.06	0.08	0.09	0.1
unstable pole	2.93e15	3.54e13	8.43e12	5.41e12	3.75e12

Table 3. The unstable poles during construction of variational reduced order model for circuit in Example 1.

When the one-port RC load is reduced with variational reduced order modeling algorithm (PACT) with fourth order, the reduced order model produces unstable poles for the driving point admittance model for a range of parameter values. These unstable poles are tabulated in Table 3. We used SPICE3f5[10] (SPICE) to simulate the SPICE-subcircuit created from reduced order macromodel with a large inverter designed in 0.6 micron CMOS technology. Since the reduced order model was nonpassive and unstable, SPICE couldn't converge and reported error when p>0.05. However, with our methods in the framework, we could be able to obtain very accurate results for all p values, p<0.1. The results for nominal, extreme and the variational macromodel for p=0.1 case agree well as shown in Figure 3.

5.2. Example 2

This next example demonstrates the efficiency of our approach to analyze the stages with large number of interconnect wires. A 4-port stage is given in Figure 4. The input signals for

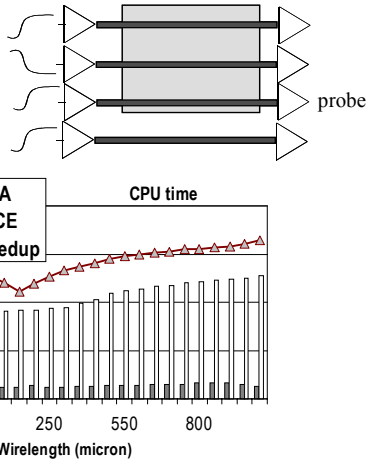

Figure 4. 4-port stage (Example 2)

Figure 5. CPU time comparison with different wirelengths. (Example 2).

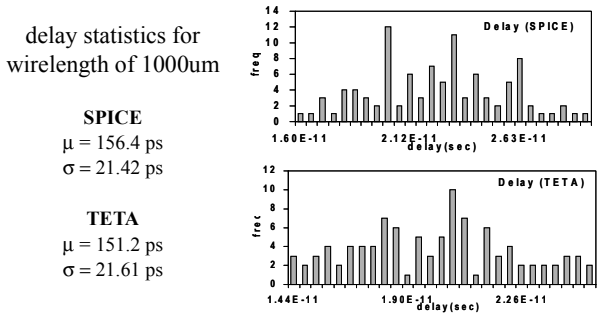

Figure 6. Delay histograms (Example 2).

the stage shown and the delay at the probe line is measured. For this simulation study, we modeled the parallel lines identical from minimum width wire geometries. The geometrical parameters for the wires, width (W), thickness (T), spacing (S), resistivity (ρ) and the inter-layer-dielectric thickness (H) values are taken from [14]. Sakurai's formulas are used to generate the electrical circuit elements [15]; but any similar model could be used. The experiment is done for variations in W, T, S, H and ρ assuming uniform distributions with tolerances specified in [14]. The wires are divided into coupled RC segments at each micron length. The delays are evaluated for 100 samples which are generated by Latin Hypercube Sampling.

The cpu-time comparison in Figure 5 for various wirelengths shows a significant speedup with respect to SPICE, especially when the number of linear circuit elements is quite large. To reflect the accuracy of the variational reduced order models, the histograms of the delays measured at probe node are shown in Figure 6. The mean and standard deviation statistics for both methods are in the order of numerical precision error of the circuit simulation tool.

5.3. Example 3

The third example demonstrates the application of the statistical critical path delay analysis for device and wire

Circuit	No of Stages	No of linear elements between stages	Speedup wrt. SPICE
s27	5	10	8.12
		500	74.2
s208	9	10	18.59
		500	78.76
s444	12	10	12.47
		500	84.62
s1423	54	10	25.25
		500	120.42
s9234	58	10	20.3
		500	100.6

Table 4. Speedup obtained with the framework.

parameters for the ISCAS-89 benchmark circuits. The gate-level descriptions of the benchmarks are transformed into transistor level circuit netlists. In the benchmark set, ten different logic cells are used. The latch-to latch paths are extracted and ordered by a unit-delay based timing analyzer. The delay variation in the longest path is analyzed in the presence of device and wire parameters.

Throughout the simulations 0.18 micron MOSFET models are used. The nominal and 3σ tolerances for 0.18 micron device and wire technology parameters are taken from [14]. As in Example 2, the electrical parameters are computed via [15]. For simplicity, each variation terms are assumed as independent distributed normal random variables. Similarly, one could use the PCA method to obtain an uncorrelated set out of the measure data.

A Monte-Carlo analysis with 100 samples is performed via SPICE and TETA using the analytical level-1 model from [10]. We observed that the number of samples are large enough to estimate the standard deviation of the distribution within 1%. The samples for channel length, W and H parameters are generated from normal distribution. The runtime speedups are tabulated in Table 4 for different numbers of linear circuit elements. The accuracy of the simulations are reasonable and the simulation speedup increases linearly with the number of linear elements. Since a simple device model is used in both simulations, the obtained speedup is mainly attributed to utilizing the variational interconnect reduced order models.

Using the same set, we apply the GA approach to evaluate the standard deviation of the critical path delay under nonlinear device model variations in threshold voltage and channel length reduction. The waveform function model parameter gradients are evaluated numerically using five simulations per each variation source.

More results for longest path delays are given in Table 5. Figure 7 shows the histograms for the longest path delays of s27 and s208 obtained by Monte Carlo and the GA approaches.

Based on our observations, GA approach becomes more efficient for short paths with few variation sources. Relying on a linearity assumption, it starts to fail for large-scale problems that have more potential nonlinearities. Monte-Carlo approach is more reliable and robust on large-scale examples, especially with large number of variation sources. Its efficiency can be enhanced

Circuit	std (DL)	std (VT)	Method	mean (ps)	std (ps)
s27 (5 stages)	0.33	0	GA	308.61	7.72
			MC	308.98	11.13
s208 (9 stages)			GA	577.64	13.56
			MC	580.38	21.02
s832 (9 stages)			GA	468.00	14.19
			MC	468.12	16.70
s444 (12 stages)			GA	1051.55	28.09
			MC	1054.08	38.12
s1423 (21 stages)			GA	1004.42	29.61
			MC	1000.97	39.54
s27 (5 stages)	0.33	0.33	GA	308.61	10.09
			MC	308.62	9.49
s208 (9 stages)			GA	577.64	16.69
			MC	579.73	24.40
s832 (9 stages)			GA	468.00	16.64
			MC	467.64	19.08
s444 (12 stages)			GA	1051.55	36.07
			MC	1052.90	45.29
s1423 (21 stages)			GA	1004.42	34.99
			MC	1000.98	44.83

Table 5. Statistics of longest path delays (Example 3)

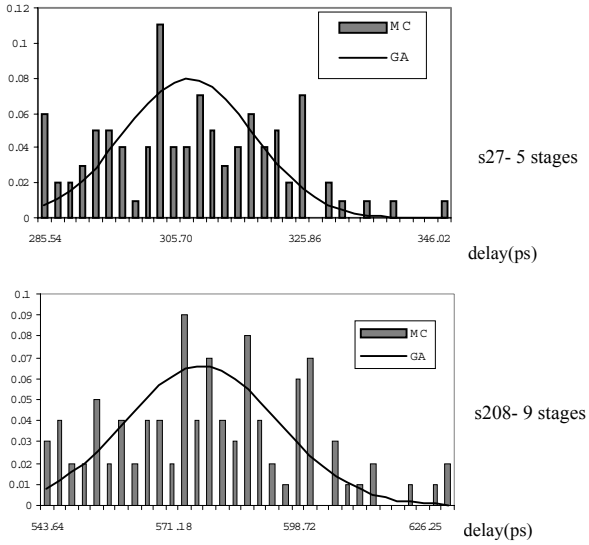

Figure 7. Histograms for the longest path delays obtained by the MC and GA analysis (under DL and VT variations)

by advanced simulation and sampling techniques.

Another distinction between these two approaches is that the required number of samples grows linearly in the GA approach with the number of variation sources, whereas in MC approach, it increases by a smaller rate.

6. Conclusion

In this paper, we described an accurate and robust simulation framework for statistical timing evaluation. Our framework efficiently handles logic stages made of variational interconnect models and nonlinear devices. The efficiencies of variational reduced order modeling and the linear-centric simulation approach are demonstrated. The proposed framework offers solutions for potential passivity/stability problem of the interconnect models, and provide great accuracy for nonlinear device models. Additionally, the utilized chord models for nonlinear devices minimizes the number of macromodeling steps.

As demonstrated by the examples given, the proposed linear centric simulation framework can be used to assess the manufacturability of next-generation integrated circuits.

7. References

[1] Liu, Y., et al., "Model-order reduction of RC(L) interconnect including variational analysis", *Proc. IEEE/ACM DAC*, 1999.

[2] Liu, Y. et al. "Impact of interconnect variations on the clock skew of a gigahertz microprocessor", *Proc. IEEE/ACM DAC*, 2000.

[3] Acar, E. et al. "Assessment of true worst case circuit performance under interconnect parameter variations", *Proc. IEEE/ISQED-2001*, 2001.

[4] Kerns, K.J. and A.T.Yang, "Stable and efficient reduction of large, multiport RC networks by pole analysis via congruence transformations", *IEEE Trans on CAD*, Apr 1997.

[5] Odabasioglu, A., etal., "PRIMA: passive reduced-order interconnect modeling algorithm", *IEEE Trans. on CAD*, Aug 1998.

[6] Dartu, F. and L. Pileggi, "TETA: Transistor-level engine for timing analysis", *Proc. IEEE/ACM DAC*, Jun 1998.

[7] Acar. E., F. Dartu, and L. T. Pileggi, "TETA: Transistor-level waveform evaluation engine for timing analysis", to appear in *IEEE Trans. on CAD.*

[8] Anastasakis, D., etal., "On the stability of approximations in asymptotic waveform evaluation", *Proc. IEEE/ACM DAC*, 1992

[9] Acar, E. *Linear-Centric Simulation Approach for Timing Analysis,* Ph.D. Thesis, Carnegie Mellon Univ. 2001.

[10] SPICE3f5 Manual, http://www.eecs.berkeley.edu/~spice.

[11] PDFAB White Paper, http://www.pdfab.com.

[12] Bolt. M, etal., "Realistic statistical worst-case simulations of VLSI circuits", *IEEE Trans. Semiconductor Manufacturing*, vol. 4. no. 3 Aug 1991.

[13] Gattiker, A., etal. "Timing yield estimation from static timing analysis", *Proc. of IEEE/ISQED-2001.* Pg 437-442.

[14] Nassif, S. R., "Modeling and analysis of manufacturing variations", *Proc IEEE Conf.-Custom Integ. Circuits*, 2001

[15] Sakurai, T. "Closed-form expressions for interconnection delay, coupling, and crosstalk in VLSIs", *IEEE Trans. ED*, vol 40. Jan 1993