INTEGRATION, the VLSI journal 43 (2010) 167–175

Contents lists available at ScienceDirect

INTEGRATION, the VLSI journal

journal homepage: www.elsevier.com/locate/vlsi

Statistical analysis of large on-chip power grid networks by variational reduction scheme ☆

Duo Li, Sheldon X.-D. Tan *

Department of Electrical Engineering, University of California, Riverside, CA 92521, USA

ARTICLE INFO

Article history:
Received 24 February 2009
Received in revised form
18 December 2009
Accepted 25 January 2010

Keywords:
Statistical
Power grid analysis
Model order reduction
Truncated balanced realization

ABSTRACT

One of the most critical challenges in today's CMOS VLSI design is the lack of predictability in chip performance at design stage. One of the process variabilities comes from the voltage drop variations in on-chip power distribution networks. In this paper, we present a novel analysis approach for computing voltage drops of large power grid networks under process variations. The new algorithm is very efficient and scalable for huge networks with a large number of variational variables. This approach, called variational extended truncated balanced realization (*varETBR*), is based on model order reduction techniques to reduce the circuit matrices before the variational simulation. It performs the parameterized reduction on the original system using variation-bearing subspaces. After the reduction, Monte Carlo based statistical simulation is performed on the reduced system and the statistical responses of the original system are obtained thereafter. *varETBR* calculates variational response Grammians by Monte Carlo based numerical integration considering both system and input source variations in generating the projection subspace. *varETBR* is very scalable for the number of variables and flexible for different variational distributions and ranges as demonstrated in experimental results. Experimental results, on a number of IBM benchmark circuits up to 1.6 million nodes, show that the *varETBR* can be 1900*X* faster than the Monte Carlo method and is much more scalable than one of the recently proposed approaches.

1. Introduction

Reliable on-chip power delivery is one of the major concerns for 90 nm and below VLSI technology. This situation becomes worse as technology continues to scale to 32 nm and below owing to the increasing process-induced variability [2,3]. The process induced variations manifest themselves at different levels (wafer level, die-level and within a die) and they are caused by different sources (lithograph, materials, aging etc.) [4,5]. Some of the variations are systematic, like those caused by chemical mechanical polishing (CMP), while some are purely random, like the doping density of impurities and edge roughness. As the technology moves to 65 nm and comes near to 45 nm, variation will become more and more pronounced for both systemic and random components.

One of the process variabilities comes from the voltage drop variations in on-chip power distribution networks. Voltage drops has significant impacts on the circuit timing [6]. Variability on voltage drops will also affect the statistical timing analysis. A number of research works have been proposed recently to address the variational voltage drop issues in the on-chip power delivery networks under process variations. The voltage drop of power grid networks subject to leakage current variations was first studied in [7,8]. This method assumes that the log-normal distribution of the node voltage drop is caused by log-normal leakage current inputs, and is based on a localized Monte Carlo (sampling) method to compute the variance of the node voltage drop. However, this localized sampling method is limited to the static DC solution of power grids modeled as resistor-only networks. Therefore, it can only compute the responses to the standby leakage currents. However, dynamic leakage currents are becoming more significant, due to the intensive use of sleep transistors for reducing leakage powers. In [9,10], impulse responses are used to compute the mean and variances of node voltage responses caused by general current variations. But this method requires the impulse response from all the current sources to all the nodes, which is expensive to compute for a large network. Methods proposed in [11,12] use orthogonal polynomial chaos expansion of random processes to represent and solve for the stochastic responses of

☆ Some preliminary results of this paper appeared in Proceedings of the Asia South Pacific Design Automation Conference (ASPDAC'09) [1]. This work is funded in part by NSF Grant under no. CCF-0448534, in part by NSF Grant under no. OISE-0929699 and in part by National Natural Science Foundation of China (NSFC) Grant under no. 60828008.

* Corresponding author.
E-mail address: stan@ee.ucr.edu (S.X. Tan).

0167-9260/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.vlsi.2010.01.004

linear systems. But existing approaches can only consider Gaussian distributions, and analysis times increase with the number of variables. The methods have been improved by the StoEKS method [13,14], where reduction is performed on the variational circuit matrices before the simulation.

In this paper, we present a novel scalable statistical simulation approach for large power grid network analysis considering process variations. The new algorithm is very scalable for large networks with a large number of random variables. Our work is inspired by the recent work on variational model order reduction using fast balanced truncation method (called variational Poor man's TBR method, or varPMTBR [15]). The new method, called *varETBR*, is based on the recently proposed extended truncated balanced realization (ETBR) method [16,17]. To consider the variational parameters, we extend the concept of response Grammian, which was used in ETBR to compute the reduction projection subspace, to the variational response Grammian. Then Monte Carlo based numerical integration is employed to multiple-dimensional integrals. Different from traditional reduction approaches, varETBR calculates the variational response Grammians, considering both system and input source variations, to generate the projection subspace. In this way, much more efficient reduction can be performed for interconnects with massive terminals like power grid networks [18]. Furthermore, the new method is based on the globally more accurate balanced truncation reduction method instead of the less accurate Krylov subspace method as in EKS/IEKS [19,20]. After the reduction, Monte Carlo based statistical simulation is performed on the reduced system and the statistical responses of the original systems are obtained thereafter. The varETBR only requires the simulation of the reduced circuit using any existing transient analysis method. It is insensitive to the number of variables and variation ranges in terms of computing costs and accuracy, which makes it very general and scalable. Experimental results, on a number of the IBM benchmark circuits [21] up to 1.6 million nodes, show that the varETBR can be up to 1900*X* faster than the Monte Carlo method, and is much more scalable than the StoEKS method [13,14]. varETBR can solve very large power grid networks with large numbers of random variables, large variation ranges and different variational distributions.

The rest of this paper is as follows: Section 2 presents the variational power grid models used in this paper. Section 3 reviews the extended Krylov subspace methods and fast balanced truncation methods. Our new variational analysis method varETBR is presented in Section 4. Section 5 shows the experimental results and Section 6 concludes this paper.

2. Power grid network models

2.1. Nominal power grid model

The power grid networks in this paper are modeled as RC networks with known time-variant current sources, which can be obtained by gate level logic simulations of the circuits. Fig. 1 shows the power grid models used in this paper. For a power grid (versus the ground grid), some nodes having known voltage are modeled as constant voltage sources. For C4 power grids, the known voltage nodes can be internal nodes inside the power grid. Given the current source vector, $u(t)$, the node voltages can be obtained by solving the following differential equations, which is formulated using the modified nodal analysis (MNA) approach,

$$G v(t) + C\frac{dv(t)}{dt} = Bu(t) \tag{1}$$

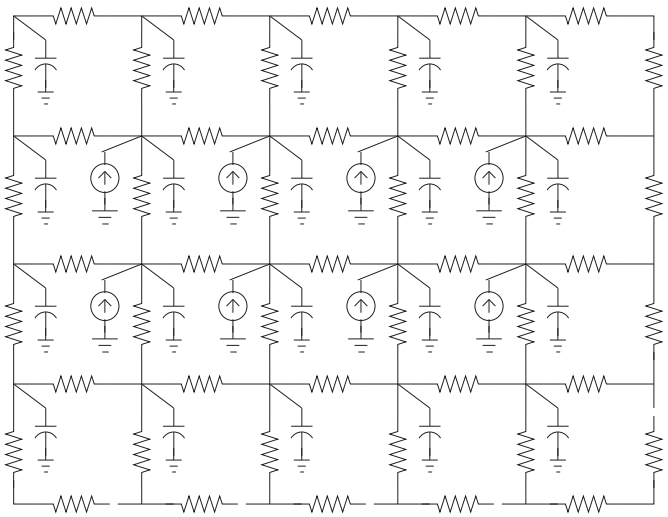


Fig. 1. The power grid model used.

where $G \in R^{n\times n}$ is the conductance matrix, $C \in R^{n\times n}$ is the matrix resulting from storage elements. $v(t)$ is the vector of time-variant node voltages and branch currents of voltage sources. $u(t)$ is the vector of independent sources, and B is the input selector matrix.

We remark that the proposed method can be directly applied to power grids modeled as RLC/RLCK circuits. But inductive effects are still most visible at board and package levels and the recent power grid networks from IBM only consists of resistance [21].

2.2. Variational model

In the presence of process variations, the G and C matrices and input currents $u(t)$ depend on variational circuit parameters, such as metal wire width, length, and metal thickness on power grids, as well as transistor parameters, such as channel length, width, gate oxide thickness, etc. Process-induced random variations can be systematic and random and can be highly partially correlated [4]. For highly correlated variations like inter-die variations, the worst case corner can be easily found by setting the parameters to their range limits (mean plus 3σ). The difficulty lies in the intra-die variations, where circuit parameters are not correlated or spatially correlated. Intra-die variations also consist of local and layout dependent deterministic components and random components. In this paper, we focus on the random variations, which are typically modeled as multivariate Gaussian processes with any spatial correlation [22].

We assume that we have a number of independent (uncorrelated) transformed orthonormal Gaussian random variables $\xi = [\xi_1, \ldots, \xi_M]$, which model the channel length, the device threshold voltage and the wire geometry variations. Therefore, the MNA equation for (1) becomes

$$G(\xi)v(t) + C(\xi)\frac{dv(t)}{dt} = Bu(t,\xi) \tag{2}$$

The spatial correlation in the intra-die variation can be processed by using the principal component analysis method (or other methods like K-L transformation or principal factor analysis, etc.) to transform the correlated variables into un-correlated variables before spectral statistical analysis [11].

Note that the input vector $u(t,\xi) = i(t,\xi) + u_0(t)$, where the current vector $i(t,\xi)$ follows the log-normal distribution and has both deterministic and random components, and the input voltage vector $u_0(t)$ is not effected by ξ. In this paper, we assume the dynamic currents (power) due to circuit switching are still modeled as deterministic currents. Therefore, we only consider

the leakage variations as they are more significant owing to their log-normal distributions. Specifically, we expand the variational G and C around their mean values and keep the first order terms as in [23,24,15].

$$G(\xi) = G_0 + G_1\xi_1 + G_2\xi_2 + \cdots + G_M\xi_M$$

$$C(\xi) = C_0 + C_1\xi_1 + C_2\xi_2 + \cdots + C_M\xi_M \quad (3)$$

We remark that the proposed method can be trivially extended to the second and higher order terms [15]. The input current variation $i(t,\xi)$ follows the log-normal distribution as leakage variations are dominant factors:

$$i(\xi) = e^{g(\xi)}, \quad g(\xi) = \mu + \sigma\xi \quad (4)$$

Note that input current variation $i(\xi)$ is not a function of time as we only model the static leakage variations for the simplicity of presentation. However, the proposed approach can be easily applied to time-variant variations with any distribution.

3. Review of fast truncated balanced realization methods

3.1. Standard truncated balanced realization methods

The truncated balanced realization (TBR) based reduction method has two steps in the reduction process: The *balancing* step transforms the states that can be controlled and observed equally. The *truncating* step then throws away the weak states, which usually leads to much smaller models. The major advantage of the TBR method is its ability to give a deterministic global bound for the approximate error and as well as provide nearly optimal models in terms of errors and model sizes.

Given a system in a standard state-space form

$$\dot{x}(t) = Ax(t) + Bu(t)$$

$$y(t) = Cx(t) \quad (5)$$

where $A \in R^{n\times n}$, $B \in R^{n\times p}$, $C \in R^{p\times n}$, $y(t)$, $u(t) \in R^p$. The controllable and observable Grammians are the unique symmetric positive definite solutions to the Lyapunov equations.

$$AX + XA^T + BB^T = 0$$

$$A^TY + YA + C^TC = 0 \quad (6)$$

Since the eigenvalues of product XY are invariant under similarity transformation, we can perform a similarity transformation ($A_b = T^{-1}AT$, $B_b = T^{-1}B$, $C_b = CT$) to diagonalize the product XY such that

$$T^{-1}XYT = \Sigma = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2) \quad (7)$$

where T matrix is the transformation matrix and the Hankel singular values of the system (σ_k), are arranged in a descending order. If we partition the matrices as

$$\begin{bmatrix} W_1^T \\ W_2^T \end{bmatrix} XY [V_1 \;\; V_2] = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \quad (8)$$

where $\Sigma_1 = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2)$ are the first r largest eigenvalues of Grammian product XY and W_1 and V_1 are corresponding eigenvectors. A reduced model can be obtained as follows

$$\dot{x}(t) = A_r x(t) + B_r u(t)$$

$$y(t) = C_r x(t) \quad (9)$$

where $A_r = W_1^T A W_1$, $B_r = W_1^T B$, $C_r = CV_1$. One most desired feature of the TBR method is that it has proved error bound: the error in the transfer function of the order r approximation is bounded by $2\sum_{i=r+1}^{N}\sigma_k$ [25,26]. In the TBR procedure, the computational cost is dominated by solving Lyapunov equations of complexity $O(n^3)$, which makes it too expensive to apply to large problem sizes.

3.2. Fast and approximate TBR methods

The TBR method generally suffers high computation costs, as it needs to solve expensive Lyapunov Eq. (6). To mitigate this problem, fast TBR methods [27,28] have been proposed recently, which compute the approximate Grammians. The Poor men's TBR method or PMTBR [28] was proposed for variational interconnect modeling.

Specifically, the Grammian X can also be computed in the time domain as

$$X = \int_0^\infty e^{At} BB^T e^{A^T t}\, dt \quad (10)$$

From Parseval's theorem, and the fact that the Laplace transform of e^{At} is $(sI-A)^{-1}$, the Grammian X can also be computed in the frequency domain as

$$X = \int_{-\infty}^{+\infty} (j\omega I - A)^{-1} BB^T (j\omega I - A)^{-H}\, d\omega \quad (11)$$

where superscript H denotes Hermitian transpose. Let ω_k be the k th sampling point. If we define

$$z_k = (j\omega_k I - A)^{-1} B \quad (12)$$

then based on the numerical quadrature rule, X can be approximated as [28]

$$\hat{X} = \sum w_k z_k z_k^H = ZW^2Z^H \quad (13)$$

where $Z = [z_1, z_2, \ldots, z_n]$. W a diagonal matrix with diagonal entries $w_{kk} = \sqrt{w_k}$. w_k comes from a specific numerical quadrature method. Since $\hat{X}$ is symmetric, it is orthogonally diagonalizable.

$$\hat{V}^T\hat{X}\hat{V} = \begin{bmatrix} \hat{V}_1^T \\ \hat{V}_2^T \end{bmatrix} \hat{X} [\hat{V}_1 \;\; \hat{V}_2] = \begin{bmatrix} \hat{\Sigma}_1 & 0 \\ 0 & \hat{\Sigma}_2 \end{bmatrix} \quad (14)$$

where $\hat{V}^T\hat{V} = I$. $\hat{V}$ converges to the eigenspaces of X and the dominant eigenvectors $\hat{V}_1$ can be used as the projection matrix in a model reduction approach ($A_r = \hat{V}_1^T A \hat{V}_1, B_r = \hat{V}_1^T B$).

3.3. Statistical reduction by variational TBR

In [15], PMTBR has been extended to reduce interconnect circuits with variational parameters. The idea is that the computation of Grammian in (11) can be viewed as the mean computation of $(j\omega I-A)^{-1}BB^T(j\omega I-A)^{-H}$ with respect to statistical variable ω, the frequency. If we have more statistical variable parameters, the Grammians can be still viewed as the mean computation, but over *all* the variables (including the frequency variables).

In the fast TBR framework, computing Grammian (11) is essentially an one-dimensional integral with respect to the complex frequency ω. When multiple variables with specific distributions are considered, multi-dimensional integral with respect to random variables will be computed. As in PMTBR, the Monte Carlo method was still employed in variational TBR to compute the multiple-dimensional integral.

One important observation in variational PMTBR is that number of samplings in building subspaces are much smaller than the number of general Monte Carlo samplings for achieving the same accuracy. As a result, variational PMTBR is much faster than the brute-force Monte Carlo method and its costs are much less sensitive to the number of random variables and variation ranges, which makes this method much more efficient than the existing variational or parameterized model order reduction methods [29].

4. New variational analysis method: varETBR

In this section, we detail the new proposed *varETBR* method. We first present the recently proposed ETBR method for deterministic power grid analysis based on reduction techniques.

4.1. Extended truncated balanced realization scheme

The new method is based on the recently proposed extended truncated balanced realization method [16]. We first review this method.

For a linear system in (1), we first define the frequency-domain *Response Grammian*,

$$X_r = \int_{-\infty}^{+\infty} (j\omega C+G)^{-1}Bu(j\omega)u^T(j\omega)B^T(j\omega C+G)^{-H}\,d\omega \tag{15}$$

which is different from the Grammian concepts in the traditional TBR based reduction framework. Notice that in the new Grammian definition, the input signals $u(j\omega)$ are considered. As a result, $(j\omega C+G)^{-1}Bu(j\omega)$ serves as the system response with respect to the input signal $u(j\omega)$ and resulting X_r becomes the response Grammian.

To fast compute the response Grammian X_r, we can use Monte Carlo based method to estimate the numerical value as done in [15]. Specifically, let ω_k be k th sampling point over the frequency range. If we further define

$$z_k^r = (j\omega_k C+G)^{-1}Bu(j\omega_k) \tag{16}$$

then $\hat{X}$ can be computed approximately by numerical quadrature methods

$$\hat{X}_r = \sum_k w_k z_k^r z_k^r H = Z_r W^2 Z_r^H \tag{17}$$

where Z_r is a matrix whose columns are z_k^r and W a diagonal matrix with diagonal entries $w_{kk}=\sqrt{w_k}$. w_k comes from a specific quadrature method.

The projection matrix can be obtained by singular value decomposition of Z_r. After this, we can reduce the original matrices into small ones and then perform the transient analysis on the reduced circuit matrices. The extended TBR algorithm is summarized in Algorithm 1.

Algorithm 1. ETBR: extended truncated balanced realization method

Input: Circuit of G, C, B, $u(t)$, number of samples: q
Output: Transient voltage waveforms

1. Convert all the input signals $u(t)$ into $u(s)$ using Fast Fourier Transformation (FFT).
2. Select q frequency points $s_1, s_2, \ldots, s_q$ over the frequency range
3. Compute $z_k^r=(s_kC+G)^{-1}Bu(s_k)$
4. Form the matrix $Z_r=[z_1^r, z_2^r, \ldots, z_q^r]$
5. Perform Singular Value Decomposition (SVD) on Z_r, $Z_r=V_rS_rU_r^T$
6. $\hat{G}=V_r^TGV_r$, $\hat{C}=V_r^TCV_r$, $\hat{B}=V_r^TB$
7. Perform the transient analysis on reduced system $[\hat{G},\hat{C},\hat{B}]$ to compute responses $\hat{v}(t)$
8. Compute the final transient waveforms $v(t)=V_r\hat{v}(t)$

Notice that we need the frequency response caused by input signal $u(j\omega_k)$ in (16). This can be obtained by fast Fourier transformation on the input signals in time domain. Using frequency spectrum representations for the input signals is a significant improvement over the EKS method as we avoid the explicit moment representation of the current sources, which are not accurate for currents rich in high frequency components due to the well-known problems in explicit moment matching methods [30]. Accuracy is also improved owing to the use of the fast balanced truncation method for the reduction, which has global accuracy [25,31].

Note that we use congruence transformation for the reduction process with orthogonal columns in the projection matrix (by using Arnoldi or Arnoldi-like process), the reduced system must be stable. For simulation purposes, this is sufficient. If all the observable ports are also the current source nodes, i.e. $y(t)=B^Tv(t)$, where $y(t)$ is the voltage vector at all observable ports, the reduced system is also passive. It was also shown in [31] that the fast TBR method has similar time complexity to multiple-point Krylov subspace based reduction methods. The extended TBR method also has similar computation costs as the EKS method.

4.2. The new variational ETBR method

We first start the new statistical interpretation of Grammian computation before introducing the new method.

4.2.1. Statistical interpretation of Grammian

For a linear dynamic system formulated in state space equations (MNA) in (1), if complex frequency $j\omega$ is a vector of random variables with uniform distribution in the frequency domain, then the state responses $V(j\omega)=(G+j\omega C)^{-1}Bu(\omega)$ become random variables in frequency domain. Its covariance matrix can be computed as

$$X_r = E\{V(j\omega)V(j\omega)^T\} = \int_{-\infty}^{+\infty} V(j\omega)V(j\omega)^T\,d\omega \tag{18}$$

where $E\{x\}$ stands for computing the mean of random variable x. X_r is defined in (15). The response Grammian essentially can be viewed as the covariance matrix associated with state responses. X_r can also be interpreted as the mean for function $P(j\omega)$ on evenly distributed random variables $j\omega$ over $[-\infty,+\infty]$.[1] ETBR method actually performs the principal component analysis (PCA) transformation of the mentioned random process with uniform distribution.

4.2.2. Computation of variational response Grammian

Define $P(j\omega)=V(j\omega)V(j\omega)^T$. Now suppose in addition to the frequency variable $j\omega$, $P(j\omega,\xi)$ is also the function of the random variable ξ with probability density $f(\xi)$. The new *variational* response Grammian X_{vr} can be defined as

$$X_{vr} = \int_{s_\xi}\int_{-\infty}^{+\infty} f(\xi)P(j\omega,\xi)\,d\omega\,d\xi = E\{P(j\omega,\xi)\} \tag{19}$$

where s_ξ is the domain of variable ξ with a specific distribution. Hence, X_{vr} is essentially the mean of $P(j\omega,\xi)$ with respect to both $j\omega$ and ξ. The concept can be extended to more random variables $\xi=[\xi_1,\xi_2,\ldots,\xi_n]$ and each variable ξ_i adds one more dimension of integration for the integral.

As a result, calculating the variational Grammian is equivalent to computing the multi-dimensional integral in (19), which can be computed by numerical quadrature methods. For one dimensional integration, efficient methods like Gaussian quadrature rule [32] exist. For multi-dimension integral, quadrature points are created by taking tensor products of one-dimensional quadrature points, which, unfortunately, grow exponentially with the number of variables (dimensions) and makes the integration intractable for practical problems [33].

[1] Practically, the interesting frequency range is always bounded.

Practically, established techniques like Monte Carlo or quasi Monte Carlo are more amenable for computing the integrals [32] as the computation costs are not dependent on the number of variables (integral dimensions). In this paper, we apply the standard Monte Carlo method to compute the variational Grammian X_{vr}. The Monte Carlo estimation of (19) consists of sampling N random points $x_i \in S$, where S is the domain for both frequency and other variables, from a uniform distribution, and then computing the estimate as

$$\hat{X}_{vr} = \frac{1}{N}\sum_{i=1}^{N} P(x_i) \tag{20}$$

The Monte Carlo method has a slow convergence rate $(1/\sqrt{N})$ in general although it can be improved to $(1/N)$ by quasi Monte Carlo methods. But as observed by Phillips [15], the projection subspace constructed from the sampled points actually converge much faster than the value of $\hat{X}_{vr}$. As we are concerned with the projection subspace rather than the actual numerical values of X_{vr}, we require only the drawing of a small number of samples as shown in the experimental result. The *varETBR* algorithm flow is shown in Algorithm 2.

Algorithm 2. varETBR: Variational extended Truncated Balanced Realization method

Input: Circuit of $G(\xi)$, $C(\xi)$, B, $u(t,\xi)$, variables $\xi=[\xi_1,\ldots,\xi_M]$, number of samples: q

Output: The variational response $v(t)$

1. Convert all the nominal input signals $u(t)$ into $u(s)$ using FFT.
2. Select q points over an M+1 dimensional space $(s,\xi_1,\ldots,\xi_M)$
3. Compute $z_k^r=(s_kC(\xi_1^k,\ldots,\xi_M^k)+G(\xi_1^k,\ldots,\xi_M^k))^{-1}Bu(s_k,\xi_1^k,\ldots,\xi_M^k))$ through Monte Carlo.
4. Form the matrix $Z_r=[z_1^r,z_2^r,\ldots,z_q^r]$
5. Perform SVD on Z_r,Z_r=$V_rS_rU_r^T$
6. $\hat{G}(\xi)=V_r^TG(\xi)V_r$, $\hat{C}((\xi))=V_r^TC(\xi)V_r$, $\hat{B}=V_r^TB$
7. Perform the Monte Carlo simulation on $\hat{G}(\xi)\hat{v}(t)+\hat{C}(\xi)\frac{d\hat{v}(t)}{dt}=\hat{B}u(t,\xi)$
8. Obtain the variational response $v(t)=V_r\hat{v}(t)$.
9. End

Where $\hat{G}(\xi)=V_r^TG(\xi)V_r$ and $\hat{C}((\xi))=V_r^TC(\xi)V_r$ stand for

$$\hat{G}(\xi)=V_r^TG_0V_r+V_r^TG_1V_r\xi_1+V_r^TG_2V_r\xi_2+\cdots+V_r^TG_MV_r\xi_M \tag{21}$$

$$\hat{C}(\xi)=V_r^TC_0V_r+V_r^TC_1V_r\xi_1+V_r^TC_2V_r\xi_2+\cdots+V_r^TC_MV_r\xi_M \tag{22}$$

The algorithm starts with the given power grid network and the number of samplings q, which are used for building the projection subspace. Then it computes the variational response $z_k^r=(s_kC(\xi_1^k,\ldots,\xi_M^k)+G(\xi_1^k,\ldots,\xi_M^k))^{-1}Bu(s_k,\xi_1^k,\ldots,\xi_M^k))$ randomly.

Table 1
Power Grid (PG) benchmarks.

Name	#Nodes	#V Sources	#I Sources
ibmpg1	30 638	14 308	10 774
ibmpg2	127 238	330	37 926
ibmpg3	851 584	955	201 054
ibmpg4	953 583	962	276 976
ibmpg5	1 079 310	539 087	540 800
ibmpg6	1 670 494	836 239	761 484

Then we perform the SVD on $Z_r=[z_1^r,z_2^r,\ldots,z_q^r]$ to construct the projection matrix. After the reduction, we perform the Monte Carlo based statistical analysis to obtain the variational responses from $v(t)=V_r\hat{v}(t)$.

We remark that in both Algorithms 1 and 2, we perform Monte-Carlo like random sampling to obtain q frequency sampling points over the M+1 dimensional space for given frequency range and parameter spaces (for Algorithm 1, sampling is on the given frequency range only). We note that the MC based sampling method is also used in the PMTBR method [15].

Compared with existing approaches, varETBR offers several advantages and features. First, varETBR only uses Monte Carlo sampling, it is easy to implement and is very general for dealing with different variation distributions and large variation ranges. It is also more amenable for parallel computing as each sampling in frequency domain can be done in parallel. Second, it is vary scalable for solving large networks with large number of variables as reduction is performed. Third, varETBR is more accurate over

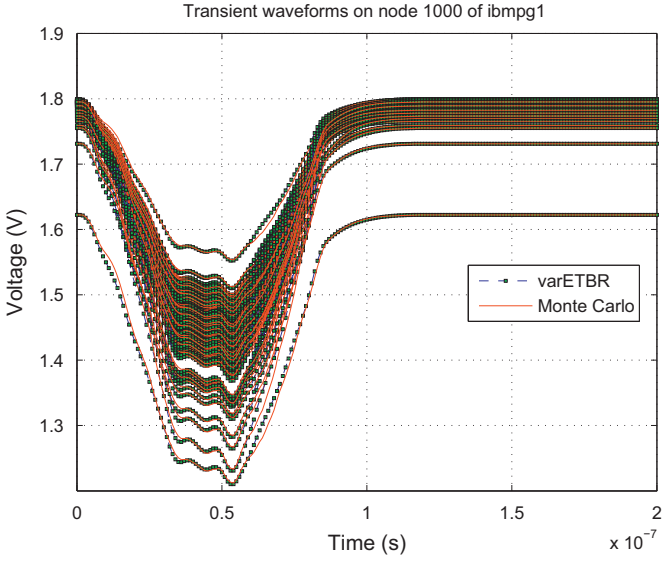


Fig. 2. Transient waveform at the 1000th node (n1_20583_11663) of *ibmpg1* (p=10, 100 samples).

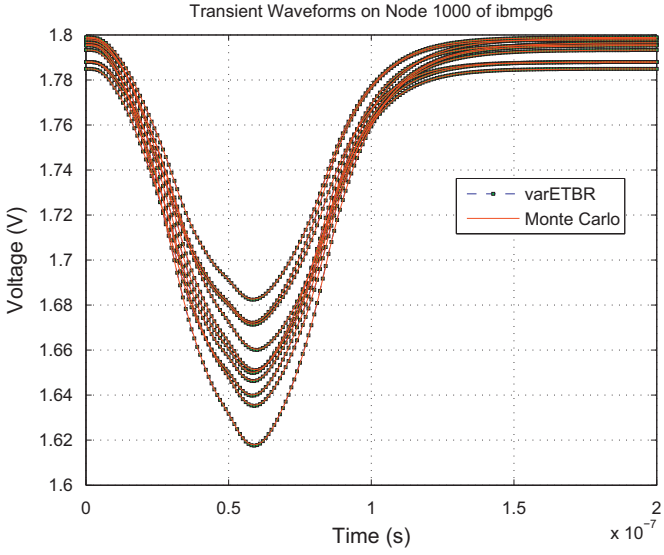


Fig. 3. Transient waveform at the 1000th node (n3_16800_9178400) of *ibmpg6* (p=10, 10 samples).

wide band frequency ranges as it samples over frequency band (compared with the less accurate moment-matching based EKS method). Last, it avoids the explicit moment representation of the input signals, leading to more accurate results than the EKS method when signals are rich in high frequency components.

5. Experimental results

The proposed *varETBR* algorithm has been implemented using MATLAB and tested on an Intel quad-core workstation with 16 GB memory under Linux environment.

All the benchmarks are real PG circuits from IBM provided by [21], but the circuits in [21] are resistor-only circuits. For transient analysis, we need to add capacitors and transient input waveforms. As a result, we modified the benchmark circuits. First we added one grounded capacitor on each node with a random value in the magnitude of pF. Second we replaced the DC current sources by a piecewise linear signal in the benchmark. The values of these signals are also randomly generated based on their original values in the DC benchmarks. We implemented a parser using Python to transform the SPICE format benchmarks into MATLAB format.

The summary of our transient PG benchmarks is shown in Table 1. We use MNA formulation to set up the circuit matrices. To efficiently solve PG circuits with 1.6 million nodes in MATLAB, an external linear solver package UMFPACK [34] is used, which is linked with Matlab using Matlab mexFunction.

We will compare varETBR with the Monte Carlo method, first in accuracy and then in CPU times. In all the test cases, the

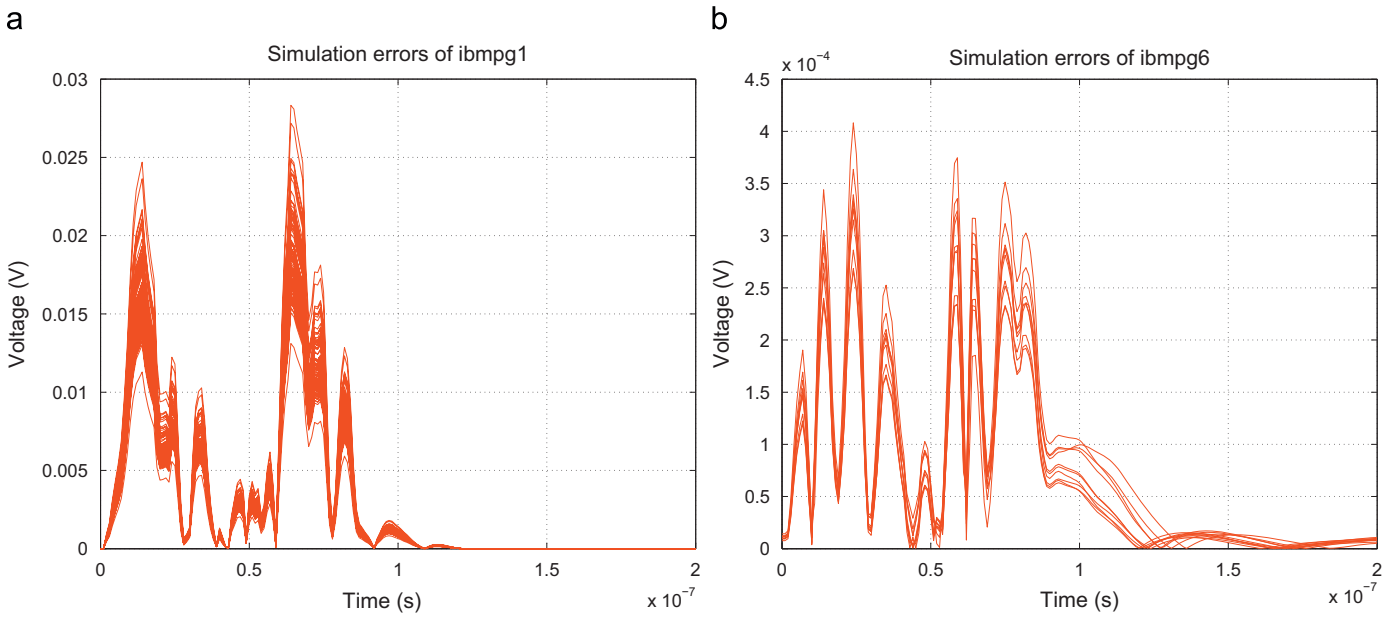


Fig. 4. Simulation errors of *ibmpg1* and *ibmpg6*. (a) Simulation errors of *ibmpg1* (100 samples). (b) Simulation errors of *ibmpg6* (10 samples).

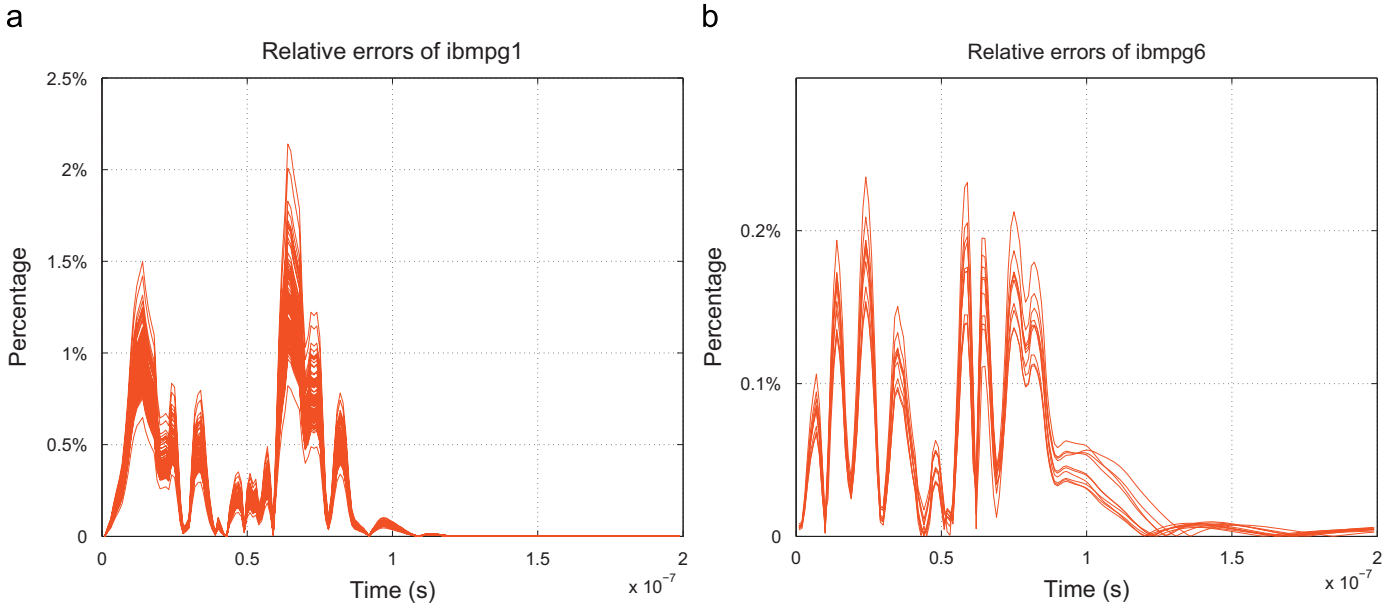


Fig. 5. Relative errors of *ibmpg1* and *ibmpg6*. (a) Relative errors of *ibmpg1* (100 samples). (b) Relative errors of *ibmpg6* (10 samples).

number of samples used for forming the subspace in varETBR are 50, based on our experience. The reduced order is set to $p=10$, which is sufficiently accurate in practice. Here we set the variation range, the ratio of the maximum variation value to the nominal value, to 10% and set the number of variables to 6 (2 for G, 2 for C and 2 for i). $G(\xi)$ and $C(\xi)$ follow Gaussian distribution. $i(t,\xi)$, which models the leakage variations [7], follows log-normal distribution.

varETBR is essentially a kind of reduced Monte Carlo method. It inherits the merits of Monte Carlo methods, which are less sensitive to the number of variables and can reflect the real distribution very accurately for a sufficient number of samples. But the main disadvantage of Monte Carlo is that it is too slow to simulate on large scale circuits. varETBR first reduces the size of circuits to a small number while maintaining sufficient accuracy. Thus, varETBR can do Monte Carlo simulation on the reduced circuits very fast. Note that the reduction process is done only once during the simulation process.

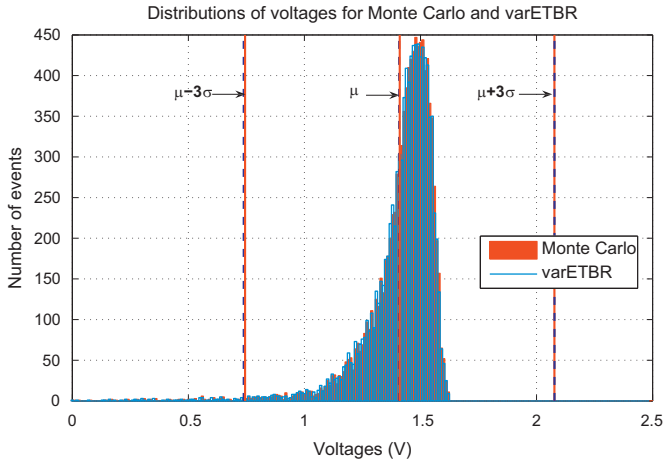


Fig. 6. Voltage distribution at the 1000th node of *ibmpg1* (10 000 samples) when $t=50$ ns.

Table 2
CPU times (s) comparison of varETBR and Monte Carlo ($q=50$, $p=10$).

Test Ckts	varETBR (s)		Monte Carlo
	Red. (s)	Sim. (s)	Sim. (s)
ibmpg1 (100)	23	14	739
ibmpg1 (10 000)	23	1335	70 719
ibmpg2 (10)	115	1.4	536
ibmpg3 (10)	1879	1.5	4973
ibmpg4 (10)	2130	1.3	5275
ibmpg5 (10)	1439	1.3	5130
ibmpg6 (10)	1957	1.5	6774

Table 3
Projected CPU times (s) comparison of varETBR and Monte Carlo ($q=50$, $p=10$, 10 000 samples).

Test Ckts	varETBR (s)	Monte Carlo (s)	Speedup
ibmpg1	1358	70 719	53*X*
ibmpg2	1515	53 600	354*X*
ibmpg3	3379	497 300	1472*X*
ibmpg4	3430	527 500	1538*X*
ibmpg5	2739	513 000	1873*X*
ibmpg6	3457	677 400	1960*X*

To verify the accuracy of our varETBR method, we show the results of simulations on *ibmpg* 1 (100 samples) and *ibmpg* 6 (10 samples). Figs. 2 and 3 show the results of varETBR and the pure Monte Carlo method at the 1000th node (named n1_20583_11663 in SPICE format) of *ibmpg* 1 and at the 1000th node (named n3_16800_9178400 in SPICE format) of *ibmpg* 6, respectively. The circuit equations in Monte Carlo are solved by MATLAB.

The absolute errors and relative errors of *ibmpg* 1 and *ibmpg* 6 are shown in Figs. 4 and 5. We can briefly see that errors are very small and our varETBR is very accurate. Note that the errors are not only influenced by the variations but also depends on the reduced order. To increase the accuracy, we may increase the reduced order. In our tests, we set the reduced order to $p=10$ for all the benchmarks.

Next we do accuracy comparison with Monte Carlo on the probability distributions including means and variances. Fig. 6 shows the voltage distributions of both varETBR and original Monte Carlo at the 1000th node of *ibmpg* 1 when $t=50$ ns (200 time steps between 0 and 200 ns in total). We can also refer to simulation waveforms on $t=50$ ns in Fig. 2. Note that the results do not follow Gaussian distribution as $G(\xi)$ and $C(\xi)$ follow Gaussian distribution and $i(t,\xi)$ follows log-normal distribution. From Fig. 6, we can see that not only are the means and the variances of varETBR and Monte Carlo almost the same, but so are their probability distributions.

Finally, we compare the CPU times of varETBR and the pure Monte Carlo method. To verify the efficiency of varETBR on both CPU time and memory, we do not need to run simulations many times for both varETBR and Monte Carlo. We will run 10 or 100 samples for each benchmark to show the efficiency of varETBR since we already showed its accuracy. Although we only run a small number of samples, the speedup will be the same. Table 2 shows the actual CPU times of both varETBR (including FFT costs) and Monte Carlo on the given set of circuits. The number of sampling points in reduction is $q=50$. The reduction order is $p=10$. Table 3 shows the projected CPU times of varETBR (one-time reduction plus 10 000 simulations) and Monte Carlo (10 000 samples).

In varETBR, circuit model becomes much smaller after reduction and we only need to performa the reduction once. Therefore, the total time is much faster than Monte Carlo (up to 1960*X*). Basically, the bigger the original circuit size is, the faster the simulation will be for varETBR. Compared to the Monte-Carlo method, the reduction time is negligible compared to the total simulation time.

Note that we run random simulation 10 000 times for *ibmpg1*, as shown in Table 2, to show the efficiency of our varETBR in practice.

It can be seen that varETBR is very scalable. It is, in practice, almost independent of the variation range and numbers of variables. One possible reason is that varETBR already captures

Table 4
Relative errors for the mean of max voltage drop of varETBR compared with Monte Carlo on the 2000th node of *ibmpg1* ($q=50$, $p=10$, 10 000 samples) for different variation ranges and different numbers of variables.

#Variables	Variation range			
	var=10% (%)	*var*=30% (%)	*var*=50% (%)	*var*=100% (%)
$M=6$	0.16	0.08	0.17	0.21
$M=9$	0.16	0.25	0.08	0.23
$M=12$	0.25	0.07	0.07	0.28
$M=15$	0.15	0.06	0.05	0.06

Table 5
Relative errors for the variance of max voltage drop of varETBR compared with Monte Carlo on the 2000th node of *ibmpg1* (q=50, p=10, 10 000 samples) for different variation ranges and different numbers of variables.

#Variables	Variation range			
	var=10% (%)	*var*=30% (%)	*var*=50% (%)	*var*=100% (%)
M=6	0.27	1.54	1.38	1.73
M=9	0.25	0.67	1.32	1.27
M=12	0.42	0.07	0.68	1.41
M=15	0.18	1.11	0.67	2.14

Table 6
CPU times (s) comparison of StoEKS and varETBR (q=50, p=10) with 10 000 samples for different numbers of variables.

Test Ckts	M=5		M=7		M=9	
	StoEKS	varETBR	StoEKS	varETBR	StoEKS	varETBR
ibmpg1	165	1315	572	1338	3748	1326
ibmpg2	1458	1387	–	1351	–	1377

the most dominant subspaces even for small number of samples (50 in our case) as explained in Section 4.2.

When we increase the variation range and the number of variables, the accuracy of varETBR is almost unchanged. Tables 4 and 5 shows that the mean and variance comparison between the two methods for 10 K Monte Carlo runs, where we increase the number of variables from 6 to 15 and the variation range from 10% to 100%. The tables show that varETBR is very insensitive to the number of variables and variation range for a given circuit *ibmpg1*, where simulations are run on 10 000 samples for both varETBR (q=50, p=10) and Monte Carlo.

The variation range *var* is the ratio of the maximum variation value to the nominal value. So "*var*=100%" means the maximum variation value may be as large as the nominal value.

From Tables 4 and 5, we observe that varETBR is basically insensitive to the number of variables and the variation range. Here we use the same sampling size (q=50) and reduced order (p=10) for all of the different combinations between number of variables and variation range. And the computation cost of varETBR is the almost same for different number of variables and different variation ranges. This actually is consistent with the observation in PMTBR [31]. One explanation for the insensitivities or nice feature of the new method is that the subspace obtained even with small number of samplings contains the dominant responses Grammian subspaces for the wide parameter and frequency ranges.

Finally, to demonstrate the efficiency of varETBR, we compare it with one recently proposed similar approach, *StoEKS* method, which employs Krylov subspace reduction with orthogonal polynomials in [13] on the same suite of IBM circuit.

Table 6 shows the comparison results where '-' means out of memory error. StoEKS can only finish smaller circuits *ibmpg1* (30 k) and *ibmpg2* (120 k), while varETBR can go through all the benchmarks (up to 1.6 M nodes) easily. The CPU time of StoEKS increases rapidly and could not complete computations as variables count increases. For varETBR, CPU time is independent of number of variables and only depends on the reduced order and number of samples used in the reduced Monte Carlo simulation. Here we select reduced order p=10 and 10 000 samples that are sufficient in practice to obtain the accurate probability distribution.

6. Conclusion

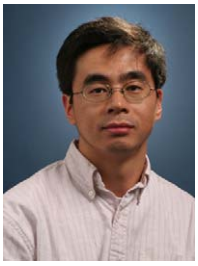
In this paper, we have proposed a new scalable statistical power grid analysis approach based on extended truncated balanced realization reduction techniques. The new method, called *varETBR*, performs reduction on the original system using variation-bearing subspaces before Monte Carlo statistical transient simulation. But different from the variational Poor man's TBR method, both system and input source variations are considered for generating the projection subspace by sampling variational response Grammians to perform the reduction. As a result, *varETBR* can reduce systems with many terminals like power grid networks while preserving variational information. After the reduction, Monte Carlo based statistical simulation is performed on the reduced system to obtain the statistical responses of the original system. Experimental results show that the varETBR can be 1900*X* faster than the Monte Carlo method and can be scalable to solve very large power grid networks with large numbers of random variables and variation ranges. varETBR is also much more scalable than the StoEKS [13] on the IBM benchmark circuits.

References

[1] D. Li, S.X.-D. Tan, G. Chen, X. Zeng, Statistical analysis of on-chip power grid networks by variational extended truncated balanced realization method, in: Proceedings of the Asia South Pacific Design Automation Conference (ASPDAC), 2009, pp. 272–277.
[2] R. Rutenbar, Next-generation design and EDA challenges, in: Proceedings of the Asia South Pacific Design Automation Conference (ASPDAC), 2007, keynote speech.
[3] S. Nassif, Model to hardware correlation for nm-scale technologies, in: Proceedings of the IEEE International Workshop on Behavioral Modeling and Simulation (BMAS), 2007, keynote speech.
[4] C. Chiang, J. Kawa, Design for Manufacturability, Springer, Berlin, 2007.
[5] S. Nassif, Delay variability: sources, impact and trends, in: Proceedings of the IEEE International Solid-State Circuits Conference, San Francisco, CA, 2000, pp. 368–369.
[6] S. Pant, D. Blaauw, Static timing analysis considering power supply variations, in: International Conference on Computer-Aided Design, 2005, pp. 365–371.
[7] I.A. Ferzli, F.N. Najm, Statistical estimation of leakage-induced power grid voltage drop considering within-die process variations, in: Proceedings of the IEEE/ACM Design Automation Conference (DAC), 2003, pp. 865–859.
[8] I.A. Ferzli, F.N. Najm, Statistical verification of power grids considering process-induced leakage current variations, in: Proceedings of the International Conference on Computer Aided Design (ICCAD), 2003, pp. 770–777.
[9] A. Srivastava, R. Bai, D. Blaauw, D. Sylvester, Modeling and analysis of leakage power considering within-die process variations, in: Proceedings of the International Symposium on Low Power Electronics and Design (ISLPED), 2002, pp. 64–67.
[10] S. Pant, D. Blaauw, V. Zolotov, S. Sundareswaran, R. Panda, A stochastic approach to power grid analysis, in: Proceedings of the IEEE/ACM Design Automation Conference (DAC), 2004, pp. 171–176.
[11] P. Ghanta, S. Vrudhula, R. Panda, J. Wang, Stochastic power grid analysis considering process variations, in: Proceedings of the European Design and Test Conference (DATE), vol. 2, 2005, pp. 964–969.
[12] P. Ghanta, S. Vrudhula, S. Bhardwaj, Stochasic variational analysis of large power grids considering intra-die correlations, in: Proceedings of the IEEE/ACM Design Automation Conference (DAC), 2006, pp. 211–216.
[13] N. Mi, S. X.-D. Tan, P. Liu, J. Cui, Y. Cai, X. Hong, Stochastic extended Krylov subspace method for variational analysis of on-chip power grid networks, in: Proceedings of the International Conference on Computer Aided Design (ICCAD), 2007, pp. 48–53.
[14] N. Mi, S.X.-D. Tan, Y. Cai, X. Hong, Fast variational analysis of on-chip power grids by stochastic extended Krylov subspace method, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 27 (11) (2008) 1996–2006.
[15] J. Phillips, Variational interconnect analysis via PMTBR, in: Proceedings of the International Conference on Computer Aided Design (ICCAD), 2004, pp. 872–879.
[16] D. Li, S.X.-D. Tan, B. McGaughy, ETBR: extended truncated balanced realization method for on-chip power grid network analysis, in: Proceedings of the European Design and Test Conference (DATE), 2008, pp. 432–437.
[17] D. Li, S.X.-D. Tan, E.H. Pacheco, M. Tirumala, Fast analysis of on-chip power grid circuits by extended truncated balanced realization method, IEICE

Transactions on Fundamentals of Electronics, Communications and Computer Science (IEICE) E92A (12) (2009) 3061–3069.
[18] S.X.-D. Tan, L. He, Advanced Model Order Reduction Techniques in VLSI Design, Cambridge University Press, Cambridge, 2007.
[19] J.M. Wang, T.V. Nguyen, Extended Krylov subspace method for reduced order analysis of linear circuit with multiple sources, in: Proceedings of the Design Automation Conference (DAC), 2000, pp. 247–252.
[20] Y. Lee, Y. Cao, T. Chen, J. Wang, C. Chen, HiPRIME: Hierarchical and passivity preserved interconnect macromodeling engine for RLKC power delivery, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 24 (6) (2005) 797–806.
[21] S.R. Nassif, Power grid analysis benchmarks, in: Proceedings of the Asia South Pacific Design Automation Conference (ASPDAC), 2008, pp. 376–381.
[22] A.B. Kahng, DFM tools and methodologies for 65 nm and below, in: Proceedings of the Asia South Pacific Design Automation Conference (ASPDAC), 2006, tutorial.
[23] Y. Liu, L.T. Pileggi, A.J. Strojwas, Model order-reduction of rc(l) interconnect including variational analysis, in: DAC '99: Proceedings of the 36th ACM/IEEE Conference on Design Automation, 1999, pp. 201–206.
[24] L. Daniel, O.C. Siong, L.S. Chay, K.H. Lee, J. White, Multi-parameter moment-matching model-reduction approach for generating geometrically parameterized interconnect performance models, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 23 (5) (2004) 678–693.
[25] B. Moore, Principal component analysis in linear systems: controllability, and observability, and model reduction, IEEE Transactions on Automatic Control 26 (1) (1981) 17–32.
[26] K. Glover, All optimal Hankel-norm approximations of linear multi-variable systems and their L_∞ error bounds, International Journal on Control 36 (1984) 1115–1193.
[27] K. Willcox, J. Peraire, Balanced model reduction via the proper orthogonal decomposition, AIAA Journal 40 (11) (2002).
[28] J.R. Phillips, L.M. Silveira, Poor man's TBR: a simple model reduction scheme, in: Proceedings of the European Design and Test Conference (DATE), 2004, pp. 938–943.
[29] Z. Zhu, J. Phillips, Random sampling of moment graph: a stochastic Krylov-reduction algorithm, in: Proceedings of the European Design and Test Conference (DATE), 2007, pp. 1502–1507.
[30] L.T. Pillage, R.A. Rohrer, C. Visweswariah, Electronic Circuit and System Simulation Methods, McGraw-Hill, New York, 1994.
[31] J.R. Phillips, L.M. Silveira, Poor man's TBR: a simple model reduction scheme, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 24 (1) (2005) 43–55.
[32] E. Suli, D. Mayers, An Introduction to Numerical Analysis, Cambridge University Press, Cambridge, 2006.
[33] R.W. Shonkwiler, L. Lefton, An Introduction to Parallel and Vector Scientific Computing, Cambridge University Press, Cambridge, 2006.
[34] Umfpack 〈http://www.cise.ufl.edu/research/sparse/umfpack/〉.

Duo Li received his B.S. and M.S. degrees in Computer Science from Northeastern University, Shenyang and Tsinghua University, Beijing, in 2003 and 2006, respectively. Now he is a Ph.D. candidate in Electrical Engineering at the University of California, Riverside. His current research interests include model order reduction, fast simulation for power grid networks, thermal modeling and thermal simulation.



Sheldon X.-D. Tan (S'96-M'99-SM'06) received his B.S. and M.S. degrees in Electrical Engineering from Fudan University, Shanghai, China in 1992 and 1995, respectively, and the Ph.D. degree in Electrical and Computer Engineering from the University of Iowa, Iowa City, in 1999.

He is an Associate Professor in the Department of Electrical Engineering, University of California, Riverside. He was a faculty member in the Electrical Engineering Department of Fudan University from 1995 to 1996. His research interests include modeling and simulation of analog/RF/mixed-signal and interconnect circuits, analysis and optimization of high performance power and clock distribution networks, architecture level thermal, power, modeling and simulation for multi-core microprocessors and embedded system designs based on FPGA platforms. He also co-authored books "Symbolic Analysis and Reduction of VLSI Circuits" by Springer/Kluwer 2005 and "Advanced Model Order Reduction Techniques" for VLSI Designs, by Cambridge University Press 2007. Dr. Tan now is serving as an Associate Editor for three journals: ACM Transaction on Design Automation of Electronic Systems (TODAE), Integration, The VLSI Journal, and Journal of VLSI Design.

Dr. Tan received Outstanding Oversea Investigator Collaboration Award from the National Natural Science Foundation of China (NSFC) in 2008. He received NSF CAREER Award in 2004. Dr. Tan received the Best Paper Award from 2007 IEEE International Conference on Computer Design (ICCD'07), a Best Paper Award Nomination from 2005 and 2009 IEEE/ACM Design Automation Conference, the Best Paper Award from 1999 IEEE/ACM Design Automation Conference. He served as a technical program committee member for ASPDAC, BMAS, ASPDAC, ISQED, ICCAD.