Fast Analysis of Structured Power Grid by Triangularization Based Structure Preserving Model Order Reduction *

Hao Yu, Yiyu Shi, and Lei He
Electrical Engineering Department
University of California, Los Angeles

ABSTRACT

*In this paper, a Triangularization Based Structure preserving (**TBS**) model order reduction is proposed to verify power integrity of on-chip structured power grid. The power grid is represented by interconnected basic blocks according to current density, and basic blocks are further clustered into compact blocks, each with a unique pole distribution. Then, the system is transformed into a triangular system, where compact blocks are in its diagonal and the system poles are determined only by the diagonal blocks. Finally, projection matrices are constructed and applied for compact blocks separately. The resulting macromodel has more matched poles and is more accurate than the one using flat projection. It is also sparse and enables a two-level analysis for simulation time reduction. Compared to existing approaches, TBS in experiments achieves up to 133X and 109X speedup in macromodel building and simulation respectively, and reduces waveform error by 33X.*

Categories and Subject Descriptors: B.7.2[Hardware]: Integrated circuits – Design aids

General Terms: Algorithms, design

Keywords: Model Order Reduction, PG grid simulation

1. INTRODUCTION

Power integrity verification is an essential part to design nowadays on-chip Power/Ground (P/G) grids. Typical P/G grid circuits usually have millions of nodes and large numbers of ports. Moreover, due to heterogeneous integration of various modules, the current density becomes highly non-uniform across the chip. It is beneficial to design a structured P/G grid that is globally irregular and locally regular [1] according to the current density. This results in a P/G circuit model as a heterogeneously structured network. To ensure power integrity, specialized simulators for P/G grid are required to efficiently and accurately analyze the voltage bounce/drop using macromodels. In [2], internal sources are eliminated to obtain a macromodel with only external ports. The entire gird is partitioned at and connected by those external ports. Because elimination results in a dense macromodel, [2] applies an additional sparsification that is error-prone and inefficient. An alternative approach to obtain macromodels is to use projection based model order reduction (MOR) such as

PRIMA [3]. The reduced model by PRIMA from a projection matrix with order q can match $n = \lfloor q/p \rfloor$ block moments (p is the port number). Although PRIMA can be implemented by iterative path-tracing to efficiently solve *tree* structured P/G grids [4], it is inefficient to be applied to *mesh* structured P/G grids.

The difficulty to apply MOR in P/G grid analysis stems mainly from following reasons. The cost of Arnoldi orthonormalization is high for large sized circuits, and the moment matching using block Krylov subspace is less accurate with an increased number of ports. In addition, the reduced macromodel is dense, which slows down simulation when the port number is large [5]. To reduce orthonormalization cost for large sized circuits, HiPRIME [6] applies a partitioned PRIMA to reduce the entire circuit in a divide-and-conquer fashion. After gluing the reduced state matrices, HiPRIME performs an additional projection to further reduced the entire system. However, approaches in [3, 6] use a flat projection that leads to the loss of the block structure of the state matrices such as *sparsity* and *hierarchy*. The resulting macromodel, therefore is too dense to be efficiently factorized in the time/frequency-domain simulation.

In this paper, we propose a triangularization based structure-preserving model order reduction, in short, TBS method. As discussed in Section 2, instead of matching block moments of the transfer function, we directly match moments of output with an *excitation current vector*. As a result, the first q moments or q dominant poles of output can be matched using a projection matrix with order q, which is independent on port number. In contrast, the number of matched block moments by PRIMA decreases as the port number increases. Hence our approach has improved accuracy for circuits with large number of ports.

As discussed in Section 3, we represent the original system by interconnected *basic blocks*. The basic blocks are obtained from the current density of locally regular structures in P/G grids. We reduce each basic block *independently* with order q, determine its first q dominant poles, and obtain its corresponding projection matrix. We then carry out a dominant-pole based clustering to obtain m clusters of basic blocks, where m is decided by the nature of structured networks. Each cluster is called as *compact block* with a unique pole distribution and a projection matrix accordingly.

As discussed in Section 4, we further triangulate the system into a triangular system with m compact blocks in the diagonal. The poles of the resulting triangular system are determined only by m diagonal blocks. Projection matrices are constructed and applied for compact blocks separately. The reduced triangular system is provable to match mq poles of the original one. This is the *primary contribution* of this paper. Because PRIMA or HiPRIME can only match q poles using the same number of moments, the reduced system by TBS is more accurate, or TBS has a higher reduction efficiency under the same error bound. A recent method BSMOR [7] leverages the subblock structure in state matrices **G** and **C**. After obtaining a flat projection matrix by PRIMA, BSMOR constructs a new block-diagonal projection matrix accordingly. Its resulting macromodel matches more poles than PRIMA does and hence improves accuracy. However, in BSMOR [7] the system poles are not determined only by those

blocks in the diagonal part of $\mathbf{G}$ and $\mathbf{C}$. As a result, the pole-matching in BSMOR is not as accurate as that in TBS. In addition, same as [3], BSMOR is inefficient for large sized circuits because it orthnormalizes the entire state matrix to obtain the projection matrix.

Further as discussed in Section 5, because the projection preserves the structure, the obtained macromodel by TBS is intrinsically sparse, and does not require the LP-sparsification procedure used in [2]. In addition, the macromodel by TBS can be efficiently analyzed by a two-level relaxation analysis [8] in both frequency and time domain, where the implicit integration is used in TBS to obtain the time domain response. As a result, the reduction and simulation of macromodel by TBS are both performed at block level, their computational cost is small although the triangularization increases the system size. In contrast, the reduced model by PRIMA or HiPRIME is dense and can not be analyzed directly with relaxation. We present the experiments in Section 6, and conclude the paper in Section 7.

2. BACKGROUND

2.1 Grimme's Moment Matching Theorem

Using the modified nodal analysis (MNA), the system equation of a P/G grid in the frequency(s)-domain is

$$(\mathbf{G} + s\mathbf{C})\mathbf{x}(s) = \mathbf{BI}(s), \quad \mathbf{y}(s) = \mathbf{L}^T\mathbf{x}(s) \tag{1}$$

where $x(s)$ is the state variable vector, $\mathbf{G}$ and $\mathbf{C}$ ($\in R^{N \times N}$) are state matrices for conductance and capacitance with size N, $\mathbf{B}$ and $\mathbf{L}$ ($\in R^{N \times p}$) are input/output port incident matrices with p ports, and $\mathbf{I}(s)$ is the input current sources.

Eliminating $x(s)$ in (1) gives

$$H(s) = \mathbf{L}^T(\mathbf{G} + s\mathbf{C})^{-1}\mathbf{B}. \tag{2}$$

$H(s)$ is a multiple-input multiple-output (MIMO) transfer function. Its expanded (s_0) columns are contained in nth-order ($n = \lceil q/p \rceil$) block-Krylov subspace $\mathcal{K}(\mathbf{A}, \mathbf{R}, n)$, i.e,

$$\mathcal{K}(\mathbf{A}, \mathbf{R}, n) = span(V) = \{\mathbf{R}, \mathbf{AR}, ..., \mathbf{A}^{n-1}\mathbf{R}\}, \tag{3}$$

where two *moment generation matrices* are $\mathbf{A} = (\mathbf{G} + s_0\mathbf{C})^{-1}\mathbf{C}$, and $\mathbf{R} = (\mathbf{G} + s_0\mathbf{C})^{-1}\mathbf{B}$. Using Arnoldi method, PRIMA [3] finds a orthonormalized projection matrix V ($\in R^{N \times q}$), which columns span block-Krylov subspace $\mathcal{K}(\mathbf{A}, \mathbf{R}, n)$.

The reduced transfer function is

$$\hat{H}(s) = \hat{\mathbf{L}}^T(\hat{\mathbf{G}} + s\hat{\mathbf{C}})^{-1}\hat{\mathbf{B}}, \tag{4}$$

where

$$\hat{\mathbf{G}} = V^T\mathbf{G}V, \ \hat{\mathbf{C}} = V^T\mathbf{C}V, \ \hat{\mathbf{B}} = V^T\mathbf{B}, \ \hat{\mathbf{L}} = V^T\mathbf{L}.$$

Note that $\hat{\mathbf{G}}$ and $\hat{\mathbf{C}} \in R^{q \times q}$, and $\hat{\mathbf{B}}$ and $\hat{\mathbf{L}} \in R^{q \times p}$. As proved in [9], $\hat{H}(s)$ preserves the block moments of $H(s)$. I.e.,

THEOREM 1. *If $\mathcal{K}(\mathbf{A}, \mathbf{R}, n) \subseteq span(V)$, then the first n expanded block moments at s_0 are identical for $\hat{H}(s)$ in (4) and $H(s)$ in (2).*

2.2 Moment Matching of Output Response

According to Theorem 1, if there is only one port, i.e., a (single-input single-output) SISO system, the reduced model can match q moments. When the port number p is large, which is typical for P/G grids, the number of matched block moment n reduces and the reduced transfer function $\hat{H}(s)$ is less accurate. In this case, it is better to define an *excitation current vector*

$$\mathbf{J} = \mathbf{BI}(s)$$

similar to [10, 6, 11], and to directly match the moment of output

$$\mathbf{x}(s) = (\mathbf{G} + s\mathbf{C})^{-1}\mathbf{J}. \tag{5}$$

The new moment generation matrices become $\mathbf{A} = (\mathbf{G} + s_0\mathbf{C})^{-1}\mathbf{C}$, and $\mathbf{R} = (\mathbf{G} + s_0\mathbf{C})^{-1}\mathbf{J}$. Using the Arnoldi method, a qth-order orthonormalized projection matrix can be found to contain the new Krylov subspace $\mathcal{K}(\mathbf{A}, \mathbf{R}, q)$. As a result, the reduced output response $\hat{\mathbf{x}}$

$$\hat{\mathbf{x}}(s) = (\hat{\mathbf{G}} + s\hat{\mathbf{C}})^{-1}\hat{\mathbf{J}} \tag{6}$$

matches the first q moments of $\mathbf{x}$, and is independent of the port number p. This is because an MIMO system with right-hand-side $\mathbf{B}u$ can be transformed into superposed SISO systems with $\mathbf{J}$. The following Theorem has been proved in [11].

THEOREM 2. *Assume an MIMO system with unit-impulse current source u, and define the excitation current vector $\mathbf{J} = \mathbf{BI}(s)$, where $\mathbf{I}(s) \in R^p$ and $\mathbf{J} \in R^N$. When the q columns of projection matrix V are obtained, the reduced response at the output $\hat{\mathbf{x}}(s)$ in (6) ($\hat{\mathbf{J}} = V^T\mathbf{J}$) matches the first q moments of the original $\mathbf{x}(s)$ in (5).*

Following two systems have the same output $x(s)$

$$(\mathbf{G} + s\mathbf{C})\mathbf{x}(s) = \mathbf{B}u(s), \ (\mathbf{G} + s\mathbf{C})\mathbf{x}(s) = \mathbf{J}(s).$$

In addition, $\mathbf{J}$ can be decomposed into p non-zero excitation components

$$\mathbf{J} = \sum_{i=1}^p \mathbf{J}_i = [J_1 \quad 0 \quad ... \quad 0]^T + ... + [0 \quad ... \quad J_p \quad 0]^T.$$

Clearly for each $\mathbf{J}_i$ ($i = 1, 2, ..., p$), it is equivalent to excite an SISO system with input $\mathbf{J}_i$. The according reduced output $\hat{\mathbf{x}}_i(s)$ matches the first q moments of $\mathbf{x}_i(s)$. With superposition, it is easy to verify that $\sum_{i=1}^p \hat{\mathbf{x}}_i(s)$ matches the first q moments of $\sum_{i=1}^p \mathbf{x}_i(s)$. In contrast, PRIMA [3] matches $n(\lceil q/p \rceil)$ block moments of the transfer function with the input matrix $\mathbf{B}$. In [11], this theorem is verified by experiments and extended to inputs with non-impulse current sources by using a *generalized excitation current source* with an augmented Arnoldi orthonormalization.

Moreover, we have

COROLLARY 1. *The first q dominant poles of $\mathbf{x}(s)$ in (5) are matched by $\hat{\mathbf{x}}(s)$ in (6).*

Poles are calculated from the eigen-decomposition of the *order reduced* moment matrix $\widetilde{\mathbf{A}} = \widetilde{\mathbf{G}}^{-1}\widetilde{\mathbf{C}}$ ($\in R^{q \times q}$). With an input of excitation current vector $\mathbf{J}$, the first q moments are identical for $\mathbf{x}(s)$ and $\hat{\mathbf{x}}(s)$. So do the first q dominant poles. In this paper, the reduction is performed to match the moment of output $\mathbf{x(s)}$.

3. COMPACT BLOCK FORMULATION

To handle large sized P/G grids and generate an accurate and sparse macromodel, we represent the original grid in compact blocks, where the overlap of pole distribution between compact blocks is minimized.

3.1 Two-level Organization of Basic Block

The original P/G grids can be partitioned into m_0 *basic blocks*, where dense grid with small pitch is used for a region with high current density, and sparse grid with large pitch is used for a region with low current density [1, 7]. The *ith* basic block has state matrices $\mathbf{g}_{ii}$ and $\mathbf{c}_{ii}$. Due to the heterogeneous structure of grids, each block can have different RC time constant. Moreover, $\mathbf{g}_{ii}$ and $\mathbf{c}_{ii}$ are interconnected by the *coupling block* $\mathbf{g}_{ij}$ and $\mathbf{c}_{ij}$ ($i \neq j$), respectively. The resulting block-based state matrices are

$$\mathbf{G} = \begin{bmatrix} \mathbf{g}_{11} & \cdots & \mathbf{g}_{1m_0} \\ \vdots & \ddots & \vdots \\ \mathbf{g}_{m_01} & \cdots & \mathbf{g}_{m_0m_0} \end{bmatrix} \mathbf{C} = \begin{bmatrix} \mathbf{c}_{11} & \cdots & \mathbf{c}_{1m_0} \\ \vdots & \ddots & \vdots \\ \mathbf{c}_{m_01} & \cdots & \mathbf{c}_{m_0m_0} \end{bmatrix}$$

and

$$\mathbf{J} = [\mathbf{J}_1 \ldots \mathbf{J}_{m_0}]^T, \quad \mathbf{x} = [\mathbf{x}_1 \ldots \mathbf{x}_{m_0}]^T. \tag{7}$$

In addition, **G** and **C** can be decomposed into the following *two-level representation* containing diagonal part $\mathbf{Y}_0(s)$ and off-diagonal part $\mathbf{Y}_1(s)$, where

$$\mathbf{Y}_0(s) + \mathbf{Y}_1(s) = \mathbf{G} + s\mathbf{C}. \tag{8}$$

Clearly, $\mathbf{Y}_0(s) = \mathbf{G}_0 + s\mathbf{C}_0$ with

$$\mathbf{G}_0 = diag[\mathbf{g}_{11}, ..., \mathbf{g}_{m_0 m_0}], \quad \mathbf{C}_0 = diag[\mathbf{c}_{11}, ..., \mathbf{c}_{m_0 m_0}].$$

Note that each block matrix $\mathbf{g}_{ii}$ or $\mathbf{c}_{ii}$ is symmetric positive definite (s.p.d), i.e., each basic block is stable. The off-diagonal part $(\mathbf{Y}_1)_{ij}$ is composed by the coupling block $\mathbf{g}_{ij} + s\mathbf{c}_{ij}$ $(i \neq j)$. Its entries are usually smaller than those in basic blocks in the diagonal. Accordingly, the moment generation matrices for each basic block are

$$(\mathbf{A}_0)_i = (\mathbf{g}_{ii} + s_0 \mathbf{c}_{ii})^{-1} \mathbf{c}_{ii}, \quad (\mathbf{R}_0)_i = (\mathbf{g}_{ii} + s_0 \mathbf{c}_{ii})^{-1} \mathbf{J}_i.$$

This two-level decomposition facilitates structure-preserving model order reduction and two-level analysis in Sections 4 and 5.

3.2 Clustering

The behavior of each basic block can be approximately determined by its q dominant poles, i.e., the first q most dominant eigen-values $(\lambda_1 \leq ... \leq \lambda_q)$. However, the basic block representation in [1, 7] is not compact. There are many basic blocks with similar time-constants as well as many basic blocks with quite dissimilar time-constants. To obtain a more compact block representation, we propose a bottom-up clustering algorithm based on the dominant poles.

Let basic block i have a q-dominant-pole set

$$\Lambda_i = \{\lambda_1 \leq ... \leq \lambda_q\},$$

we define its *pole distance* from another basic block j

$$d(\Lambda_i, \Lambda_j) = \max\{\min\{|\lambda_m - \lambda_n| : \lambda_n \in \Lambda_j\} : \lambda_m \in \Lambda_i\}.$$

The two basic blocks have a similar pole distribution and are clustered if $d(\Lambda'_i, \Lambda_i) < \epsilon$, where ϵ is a small value specified by the user. After clustering basic block i ($\mathbf{g}_{ii}, \mathbf{c}_{ii}$) with j ($\mathbf{g}_{jj}, \mathbf{c}_{jj}$) and their interconnection ($\mathbf{g}_{ij}, \mathbf{c}_{ij}$), the q-dominant-pole set of the clustered block becomes

$$\Lambda'_i = \{\lambda'_1 \leq ... \leq \lambda'_q\}.$$

More basic blocks can be merged into this cluster if they have a similar pole distribution as the cluster. On the other hand, a basic block itself is a cluster if it does not share a similar pole distribution with other blocks.

For clustering purpose, the first q dominant poles for a basic block is obtained by model order reduction. A qth-order projection matrix V_i is found for basic block i by

$$span(V_i) = \mathcal{K}((\mathbf{A}_0)_i, (\mathbf{R}_0)_i, q) \quad i = 1, ..., m_0. \tag{9}$$

It results in a order reduced $(\widetilde{\mathbf{A}}_0)_i$ ($\in R^{q \times q}$), whose reciprocal eigen-values are poles of the reduced system, and match the first q dominant poles of the original system according to Corollary 1. The cost of eigen-decomposition is inexpensive if the size of reduced model is small. Because the excitation current vector is used during the moment matching of the output, the size q of the reduced model with desired accuracy is small even when the original basic block contains large number of ports. In contrast, the block moment matching by PRIMA may result in a larger cost of eigen-decomposition.

The clustering obtains m clusters of basic blocks, where m is decided by the nature of P/G grids and ϵ. We call cluster as a *compact block* in this paper. It results in an interconnected compact block representation, where the sets of q-dominant poles for compact blocks have minimum overlap between them. Therefore, different from [1, 7], our method reduces the redundant information because fewer number of compact blocks are needed to represent the structured system.

4. TBS MODEL ORDER REDUCTION

Although clustering results in m blocks, each with a unique pole distribution, the poles of the entire grids are not determined only by those diagonal blocks. In this section, we discuss how to form the upper triangular system $(\mathcal{G}, \mathcal{C})$ that are equivalent to the original system $(\mathbf{G}, \mathbf{C})$, and the system poles of $(\mathcal{G}, \mathcal{C})$ are determined only by its diagonal blocks [12]. This enables block structured projection that can match more poles than the flat projection.

4.1 Triangularization

With respect to the following **G** after the clustering discussed in Section 3.2

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} & \ldots & \mathbf{G}_{1m} \\ \mathbf{G}_{21} & \mathbf{G}_{22} & \ldots & \mathbf{G}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{G}_{m1} & \mathbf{G}_{m2} & \ldots & \mathbf{G}_{mm} \end{bmatrix}, \tag{10}$$

the triangularization is to introduce a replica of **G**, and move those lower triangular blocks $\mathbf{G}_{ij}$ $(i < j)$ to the upper triangular parts at $\mathcal{G}_{i,m+j}$. This results in an *upper triangular* state matrix $\mathcal{G}$

$$\mathcal{G} = \left[\begin{array}{cccc|cccc} \mathbf{G}_{11} & \mathbf{G}_{12} & \ldots & \mathbf{G}_{1m} & 0 & 0 & \ldots & 0 \\ 0 & \mathbf{G}_{22} & \ldots & \mathbf{G}_{2m} & \mathbf{G}_{21} & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \mathbf{G}_{mm} & \mathbf{G}_{m1} & \mathbf{G}_{m2} & \ldots & 0 \\ \hline & & \mathbf{0} & & & & \mathbf{G} & \end{array} \right]. \tag{11}$$

$\mathcal{C}$ can be transformed in a similar fashion. In addition, the new state variable x is

$$x = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \ldots & \mathbf{x}_m & \mathbf{x} \end{bmatrix}^T,$$

where $\mathbf{x}$ is defined in (7), and the port matrix $\mathcal{B}$ and $\mathcal{L}$ have similar structures as x. The resulting *triangular system* equation is

$$(\mathcal{G} + s\mathcal{C})x(s) = \mathcal{J}, \ y(s) = \mathcal{L}^T x(s). \tag{12}$$

It is easy to verify that the solution $x(s)$ from (12) is the same as $\mathbf{x}(s)$ from (1).

Below, we prove that the new triangular system is stable.

THEOREM 3. *The upper block triangular system $(\mathcal{G}, \mathcal{C})$ is stable.*

Proof: The eigen-values of the triangular system are given by the product of determinants of diagonal blocks

$$|\mathcal{G}| = \prod_{i=1}^{m+1} |(\mathcal{G}_0)_i| = |(\mathbf{G}_0)_1|...|(\mathbf{G}_0)_m||\mathbf{G}|.$$

Because each block $(\mathbf{G}_0)_i$ $(1 \leq i \leq m)$ and **G** are positive definite, $\mathcal{G}$ is positive definite as well. The same procedure can be used to prove that $\mathcal{C}$ is positive definite. Therefore, $\mathcal{G} + \mathcal{G}^T$ and $\mathcal{C} + \mathcal{C}^T$ are both s.p.d, and hence the triangular system is stable.

Note that directly solving (12) involves a similar cost to solve (1) as the replica block at the lower-right corner needs to be factorized first. In addition, the dimension of the triangular system is increased. However, because the reduction in TBS is performed at the block level, the orthonormalization cost is small. Moreover, as shown below, its benefits can be further appreciated after a structure-preserving model order reduction, where the state variable of each reduced block can be solved independently with q matched poles.

4.2 mq-pole Matching

After the clustering in Section 3.2, we can also obtain a set of projection matrices $[V_{1(n_1 \times q)}, \ldots, V_{m(n_m \times q)}]$, one for each diagonal compact block with size n_i. The projection matrix V_{m+1} for
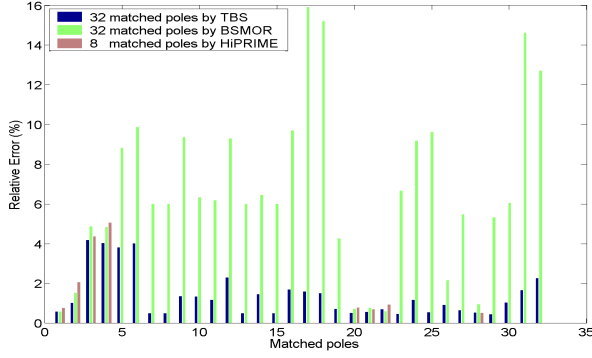
Figure 1: Pole matching comparison: mq **poles matched by TBS and BSMOR, and** q **poles matched by HiPRIME.**
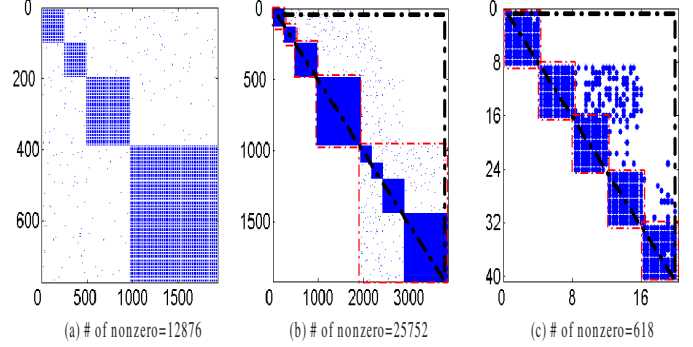

Figure 2: Nonzero-entry pattern of conductance matrices: (a) original system (b) triangular system (c) reduced system by TBS. (a)-(c) have different dimensions, but (b)-(c) have the same triangular structure and same diagonal block structure.

the replica block can be obtained from a qth-order orthonormalization or practically constructed from

$$V_{m+1} = [V_1, ..., V_m] \quad (\in R^{N \times q}). \quad (13)$$

Furthermore, instead of constructing a flat projection matrix

$$V = [V_1, ..., V_m, V_{m+1}], \quad (\in R^{2N \times q}), \quad (14)$$

we reconstruct a block-diagonal structured projection matrix $\mathcal{V}$:

$$\mathcal{V} = diag[V_{1(n_1 \times q)}, ..., V_{m(n_m \times q)}, V_{m+1(N \times q)}] \quad (15)$$

with $\mathcal{V} \in R^{2N \times (m+1)q}$, $\sum_{i=1}^m n_i = N$. Note that $\mathcal{V}^T \mathcal{V} = I$, i.e., each column of $\widetilde{V}$ is still linearly independent and hence the total column-rank is increased by a factor of the block number m. With the use of $\mathcal{V}$ to project $\mathcal{G}$, $\mathcal{C}$ and $\mathcal{B}$ matrices at block level, respectively, we can obtain the order reduced state matrices

$$\widetilde{\mathcal{G}} = \mathcal{V}^T \mathcal{G} \mathcal{V}, \quad \widetilde{\mathcal{C}} = \mathcal{V}^T \mathcal{C} \mathcal{V}, \quad \widetilde{\mathcal{J}} = \mathcal{V}^T \mathcal{J}.$$

Especially, the diagonal blocks in reduced $\widetilde{\mathcal{G}}$ and $\widetilde{\mathcal{C}}$ are called *reduced blocks*.

The reduced $\widetilde{\mathcal{G}}$ matrix preserves the upper block triangular structure

$$\widetilde{\mathcal{G}} = \begin{bmatrix} \widetilde{\mathcal{G}}_A & \widetilde{\mathcal{G}}_B \\ \mathbf{0} & \widetilde{\mathcal{G}}_D \end{bmatrix}, \quad (16)$$

where

$$\widetilde{\mathcal{G}}_A = \begin{bmatrix} V_1^T \mathbf{G}_{11} V_1 & V_1^T \mathbf{G}_{12} V_2 & \ldots & V_1^T \mathbf{G}_{1m} V_m \\ 0 & V_2^T \mathbf{G}_{22} V_2 & \ldots & V_2^T \mathbf{G}_{2m} V_m \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & V_m^T \mathbf{G}_{mm} V_m \end{bmatrix}$$

$$\widetilde{\mathcal{G}}_B = \begin{bmatrix} 0 & 0 & \ldots & 0 \\ V_1^T \mathbf{G}_{12} V_2 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ V_m^T \mathbf{G}_{m1} V_1 & V_m^T \mathbf{G}_{m2} V_2 & \ldots & 0 \end{bmatrix}$$

$$\widetilde{\mathcal{G}}_D = V_{m+1}^T \mathbf{G} V_{m+1}. \quad (17)$$

Since BSMOR does not use triangularization, its system poles are not determined by those diagonal blocks. Therefore, its reduced macromodel does not exactly have mq poles matching (See experiments in Fig. 1 of Section 6). In contrast, TBS can exactly match mq poles as discussed below.

THEOREM 4. *For the state matrices $\mathcal{G}$ and $\mathcal{C}$ in the upper triangular block form, if there is no overlap between eigen-values of the reduced blocks $(\widetilde{\mathbf{G}}_{ii}, \widetilde{\mathbf{C}}_{ii})$ $(\in R^{q \times q})$, i.e.,*

$$|(\widetilde{\mathbf{G}}_{00})_1 + s(\widetilde{\mathbf{C}}_{00})_1| \cup ... \cup |(\widetilde{\mathbf{G}}_{00})_m + s(\widetilde{\mathbf{C}}_{00})_m| = Null, \quad (18)$$

the reduced system $(\widetilde{\mathcal{G}} + s\widetilde{\mathcal{C}})$ exactly matches mq poles of the original system $(\mathcal{G} + s\mathcal{C})$.

Proof: Because the original $\mathcal{G}$ and $\mathcal{C}$ are in the upper triangular form, and the projection by $\mathcal{V}$ preserves the structure, the reduced $\widetilde{\mathcal{G}}$ and $\widetilde{\mathcal{C}}$ are in the upper triangular block form as well. For an upper triangular block system $\widetilde{\mathcal{G}} + s\widetilde{\mathcal{C}}$, its poles (eigen-values) are the roots of its determinant $|\widetilde{\mathcal{G}} + s\widetilde{\mathcal{C}}|$, which are determined only by the diagonal blocks

$$|\widetilde{\mathcal{G}} + s\widetilde{\mathcal{C}}| = \prod_{i=1}^m |\widetilde{\mathbf{G}}_{ii} + s\widetilde{\mathbf{C}}_{ii}|.$$

Note that eigenvalues of $|\widetilde{\mathcal{G}} + s\widetilde{\mathcal{C}}|$ represent the reciprocal poles of the reduced model [3]. For the reduced block $\widetilde{\mathbf{G}}_{ii} + s\widetilde{\mathbf{C}}_{ii}$ with input $\mathcal{J}_i$, its output $\widetilde{x}_i$ matches q moments and the first q domain poles of the output x_i for block $\mathbf{G}_{ii} + s\mathbf{C}_{ii}$ in the triangular system. Since the entire system consists of m compact blocks, each with unique pole distribution, the reduced model by TBS can match mq poles. Note that the redundant poles obtained from the replica block are not counted here. With more matched poles, TBS is more accurate than HiPRIME and BSMOR. This will be verified by experiments in Section 6.

5. TWO LEVEL ANALYSIS

Because the projection in TBS preserves the structure, the reduced state matrices are sparse if the original ones are sparse. In contrast, when projected by flat projection V in PRIMA and HiPRIME, the resulted $\hat{\mathbf{G}}$ is

$$\hat{\mathbf{G}} = \sum_{i=1}^{m+1} \sum_{j=1}^{m+1} V_i^T \mathbf{G}_{ij} V_j, \quad (19)$$

which loses the structure in general, and the reduced state matrices are dense. This slows down simulation when $\hat{\mathbf{G}}$ and $\hat{\mathbf{C}}$ are stamped back to MNA.

Due to the structure-preserving, the reduced triangular system by TBS can be further analyzed efficiently either by a direct backward substitution or a two-level relaxation analysis [8]. As the two-level analysis enables the parallelized solution and can be extended to the hierarchical analysis, it is used in this paper to obtain the solution in both frequency and time domains. As a result, the state variable of each reduced block can be solved independently with matched q poles.

Using the two-level representation discussed in Section 3.1, the system equation for the reduced model is
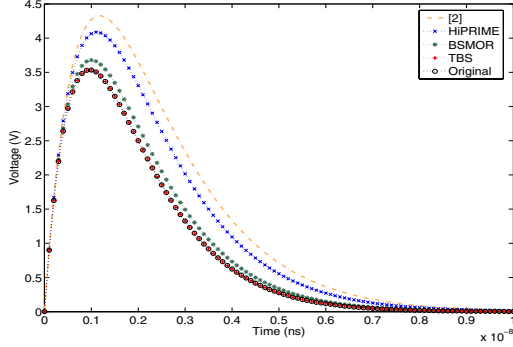
$$\widetilde{\mathcal{Y}} x = \widetilde{b}. \quad (20)$$

Figure 3: Comparison of time-domain responses between the method in [2], HiPRIME, BSMOR, TBS and the original. TBS is identical to the original.

Figure 4: Comparison of frequency-domain responses between HiPRIME, BSMOR, TBS and the original. TBS is identical to the original.

In frequency domain at a frequency point s, (20) becomes

$$\widetilde{\mathcal{Y}} = \widetilde{\mathcal{G}} + s\widetilde{\mathcal{C}} = \widetilde{\mathcal{Y}}_0(s) + \widetilde{\mathcal{Y}}_1(s), \quad \widetilde{b} = \widetilde{\mathcal{J}}(s),$$

and in time domain at a time instant t with time step h, (20) becomes

$$\widetilde{\mathcal{Y}} = \widetilde{\mathcal{G}} + \frac{1}{h}\widetilde{\mathcal{C}} = \widetilde{\mathcal{Y}}_0(h) + \widetilde{\mathcal{Y}}_1(h), \quad \widetilde{b} = \frac{1}{h}\widetilde{\mathcal{C}}x(t-h) + \widetilde{\mathcal{J}}(t).$$

Note that the time step h can be different for each reduced block according to its dominant-pole (λ_1).

The state vector x can be solved for each block in a fashion of the two-level relaxation analysis [8], where

$$x = P^{(0)} - PQ \tag{21}$$

with

$$P^{(0)} = (\widetilde{\mathcal{Y}}_0)^{-1}\widetilde{b}, \quad P = (\widetilde{\mathcal{Y}}_0)^{-1}\widetilde{\mathcal{Y}}_1, \quad Q = (I+P)^{-1}P^{(0)}. \tag{22}$$

To avoid explicit inversion, LU or Cholesky factorization is applied to $\widetilde{\mathcal{Y}}_0$ and $I + (\widetilde{\mathcal{Y}}_0)^{-1}\widetilde{\mathcal{Y}}_1$. As $\widetilde{\mathcal{Y}}_0$ shows the block diagonal form, each reduced block matrix is first solved independently with LU/Cholesky factorization and substitution at the bottom level. The results from each reduced block are then used further to solve the coupling block at the top level, and the final x_k of each reduced block is updated. In addition, because the reduced $\widetilde{\mathcal{Y}}$ has preserved block triangular structure, an implicit Back-Euler integration with the relaxation can stably converge [8].

6. EXPERIMENTS

We implemented the TBS and experimented on a Linux workstation (P4 2.66GHz, 1Gb RAM). The RC mesh structures of the P/G grid are generated from industrial applications. In this section, we first verify that TBS preserves triangular structure (sparsity) and matches mq poles, and then compare its accuracy and runtime with the method in [2], HiPRIME [6] and BSMOR [7]. The excitation current sources (unit-impulse) are explicitly considered in all MOR based methods to avoid block moment matching. The clustered block structure obtained from TBS is used as the partition for HiPRIME and [2], and the same block number is used for BSMOR but each block has the same size. Back-Euler method is used for time-domain simulation, and two-level analysis is applied for TBS, BSMOR and [2]. In the comparison of the macromodel building and simulation time, all reduced models have similar accuracy. In the comparison of the waveform error, all MOR methods use the same number of matched moments, and macromodels for TBS and [2] have the similar size and sparsification ratio.
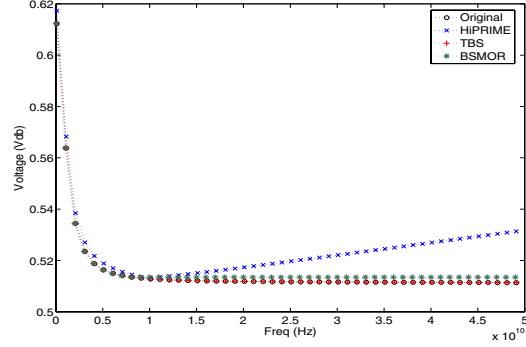
6.1 Accuracy Comparison

We use a non-uniform RC mesh (size $2K \times 2K$) with 32 same sized basic blocks and 32 unit-impulse current sources located at centers of basic blocks. Each basic block has a different RC time constant. The number of connections between a pair of basic blocks are also different. HiPRIME, BSMOR and TBS all use $q = 8$ moments to generate the reduced model. The clustering algorithm found 4 clusters, each with 4, 4, 8, 16 basic blocks respectively. As a result, TBS constructs a block structured projection using 4 blocks with the aforementioned sizes. In contrast, BSMOR constructs a block structured projection using 4 blocks with same size.

Fig. 2 shows the nonzero-entry pattern of the conductance matrix before triangularization in Fig. 2 (a), after triangularization in Fig. 2 (b), and after the TBS reduction ($m = 4, q = 8$) in Fig. 2 (c). Fig. 2 (b) and (c) have the similar block triangular structure, which verifies that TBS preserves the block structure. Due to the intrinsic sparsity by TBS, the reduced model has a 40.1% sparsification ratio. In contrast, HiPRIME generates a fully dense state matrices after the reduction and the sparsity in the reduced model by [2] is obtained by an additional LP-based sparsification.

To compare pole-matching, we choose one observation port that is not at the source node. The relative errors are calculated as the magnitude difference of poles between the reduced and original models. As shown by Fig. 1, HiPRIME can only approximate 8 poles of the original model, but TBS and BSMOR can approximate 32 poles due to increased column rank in the projection matrix. Moreover, for poles matched by both TBS and BSMOR, TBS is about 6X more accurate on average. As discussed in Section 4.2, this is because the poles of triangularized system in TBS are determined only by its diagonal blocks.

Fig. 3 compares the time-domain response at one port for HiPRIME, BSMOR, [2], TBS and the original under a unit-impulse input. The time-domain waveform error is counted as the relative deviation at peak voltage. The reduced model by TBS is visually identical to the original model, but HiPRIME has up to 36% error due to much fewer matched poles, and [2] has up to 64% error due to the sparsification. As mentioned before, the projection matrix constructed by BSMOR uses 4 blocks with the same size. As a result, it is not exact in matching poles and results in up to 23% error. Fig. 4 further presents the frequency-domain response under an impulse input. Using same number of moments, we observe that the reduced model by TBS is identical to the original up to 50GHz, but the one by BSMOR or HiPRIME has non-negligible deviations beyond 10GHz.

6.2 Scalability Study

Table 1 compares the accuracy scalability of reduced macromodel by [2], HiPRIME, BSMOR and TBS. All reduced models by MOR use the same number of moments. The standard deviation of waveform differences between the reduced and the original models is used as the measure of error. We use higher

ckt	node (N)	port (p)	order (q)	[2]	HiPRIME	BSMOR	TBS
ckt1	1K	48	8	5.54e-6	9.09e-6	4.87e-6	5.03e-7
ckt2	10K	320	40	1.21e-5	2.31e-5	7.93e-6	1.84e-6
ckt3	100K	480	60	1.31e-2	6.82e-4	1.91e-4	3.02e-5
ckt4	1M	800	100	6.01e-2	9.67e-3	4.23e-3	1.27e-4
ckt5	7.68M	4800	200	0.11	9.93e-2	5.10e-2	3.01e-3
ckt6	7.68M	6.14M	300	NA	NA	NA	5.04e-3

Table 1: Time-domain waveform error of reduced models by the method in [2], HiPRIME, BSMOR and TBS under the same order (number of matched moments).

ckt	[2]		HiPRIME		BSMOR		TBS	
	build	sim	build	sim	build	sim	build	sim
ckt1	0.44s	0.08s	0.15s	1.02s	0.12s	0.08s	0.09s	0.08s
ckt2	2.19s	1.24s	0.54s	1min:42s	0.63s	1.18s	0.11s	1.02s
ckt3	1min:17s	1min:51s	5.76s	2hr:48min:20s	1min:2s	1min:38s	1.62s	1min:32s
ckt4	34min:58s	21min:32s	47.3s	> 2day	4min:54s	11min:42s	20.7s	11min:23s
ckt5	4hr:43min:18s	1day:5hr:11min	2min:42s	> 2day	1hr:45min	1day:1hr:36min	2min:8s	1day:18min
ckt6	NA	NA	NA	NA	NA	NA	6min:16s	1day:1hr:29min

Table 2: Comparison of runtime under the similar accuracy of the method in [2], HiPRIME, BSMOR and TBS. The runtime includes macromodel building and simulation time, respectively.

order reduced model (by 4X) as the baseline for comparison if the waveform of the original model is unavailable. We find that the accuracy of [2] degrades when a large sparsity ratio is needed, because LP-based sparsification can not preserve accuracy. On the other hand, using moment matching based projection with preserved sparsity, TBS generates a macromodel with higher accuracy. For example, it has a 38X higher accuracy than [2] when reducing a 7.68M circuit with 4800 ports to a (1K) macromodel with 32% sparsity. For the same circuit, TBS is 17X more accurate than BSMOR due to the exact mq-pole matching, and is also 33X more accurate than HiPRIME due to more matched poles. Because [2] and BSMOR are inefficient to build macromodels and HiPRIME is inefficient to simulate macromodels, only TBS can handle a 7.68M circuit with 6.14M ports for less 1% waveform error.

Table 2 compares the runtime scalability of reduced macromodel by [2], HiPRIME, BSMOR and TBS. The runtime time here includes both the macromodel building time and macromodel simulation time (time-domain). The same circuits in Table 1 are used (but reduced state matrices are constructed with the similar accuracy). As for the the macromodel building time, [2] needs the additional LP-based sparsification, which is inefficient for large sized P/G grids. For example, for a RC-mesh with 7.68M nodes, the method in [2] needs $4hr : 43min : 18s$ to build a reduced macromodel with 1K nodes and sparsity 30%, but TBS only spends $2min : 8s$ (133X speedup) to build the similar sized macromodel. Moreover, TBS also has 54X speedup than BSMOR (1$hr : 45min$) because orthonormalization is applied to each block independently in TBS. HiPRIME orthnormalizes each block independently, but its building time is still larger than TBS. This is due to that a higher order (4X) is required to generate a reduced model with similar accuracy as TBS. Moreover, as for the simulation time, because HiPRIME still uses flat projection, it results in a dense macromodel, loses the structure information and can not be analyzed hierarchically. Therefore, it becomes inefficient to be used for time-domain simulation. As a result, its simulation time is much larger than the other macromodels. On the other hand, [2], BSMOR and TBS enable the two-level analysis. For a circuit with 100K nodes and 480 ports, TBS achieves 109X runtime speedup compared to HiPRIME. In addition, for the circuit with 7.68M nodes and 6.14M ports, only TBS can handle it with $6min : 16s$ to build and $1day : 1hr : 29min$ to simulate.

7. CONCLUSIONS

In this paper, we have proposed an accurate and efficient TBS model order reduction method to verify integrity of for large sized P/G grids in the time-domain. Using triangularization, we show that the original system is stably transformed into a form with up-per triangular block structure, where system poles are determined only by m diagonal blocks, and m is decided by the nature of the structured network. With an efficient dominant-pole based clustering and a block structured projection, the reduced triangular system can match mq poles of the original system. Experiments show that the waveform error is reduced 33X compared to the flat projection method by HiPRIME. Moreover, with a two-level block representation, the reduction and analysis in TBS can be performed for each block independently. Therefore, it reduces both macromodel building and simulation time. TBS is up to 54X faster to build macromodels than BSMOR, and up to 109X to simulate macromodels in time-domain than HiPRIME. In addition, as TBS preserves sparsity, it is up to 133X faster to build macromodels than [2].

8. REFERENCES

[1] J. Singh and S. Sapatnekar, "Congestion-aware topology optimization of structured power/ground networks," *IEEE Trans. on CAD*, pp. 683–695, 2005.

[2] M. Zhao, R. Panda, S.Sapatnekar, and D. Blaauw, "Hierarchical analysis of power distribution networks," *IEEE Trans. on CAD*, pp. 159–168, 2002.

[3] A. Odabasioglu, M. Celik, and L. Pileggi, "PRIMA: Passive reduced-order interconnect macro-modeling algorithm," *IEEE Trans. on CAD*, pp. 645–654, 1998.

[4] H. Su, K.Gala, and S. Sapatnekar, "Analysis and optimization of structured power/ground networks," *IEEE Trans. on CAD*, pp. 1533–1544, 2003.

[5] P. Feldmann and F. Liu, "Sparse and efficient reduced order modeling of linear sub-circuits with large number of terminals," in *Proc. IEEE/ACM ICCAD*, 2004.

[6] Y. Lee, Y. Cao, T. Chen, J. Wang, and C. Chen, "HiPRIME: Hierarchical and passivity preserved interconnect macromodeling engine for RLKC power delivery," *IEEE Trans. on CAD*, vol. 26, no. 6, pp. 797–806, 2005.

[7] H. Yu, L. He, and S. Tan, "Block structure preserving model reduction," in *Proc. IEEE BMAS*, 2005.

[8] J. White and A. Sangiovanni-Vincetelli, *Relaxation Techniques for the Simulation of VLSI Circuits.* Kluwer Academic Publishers, 1987.

[9] E.J.Grimme, *Krylov projection methods for model reduction (Ph. D Thesis).* Univ. of Illinois at Urbana-Champaign, 1997.

[10] K. J. Kerns and A. T. Yang, "Stable and efficient reduction of large, multiport RC network by pole analysis via congruence transformations," *IEEE Trans. on CAD*, pp. 734–744, 1998.

[11] Y. Shi, H. Yu, and L. He, "SAMSON: A generalized second-order Arnoldi method for multi-source network reduction," in *Proc. ACM ISPD*, 2006.

[12] G. H. Golub and C. F. V. Loan, *Matrix Computations.* Baltimore, MD: The Johns Hopkins University Press, 3 ed., 1989.