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

Under Review, Please donnot Distribute

Hao Yu, Yiyu Shi, and Lei He *

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. 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, a block-diagonal structured projection matrix is constructed by stacking projection matrices for individual diagonal blocks in the triangular system. The resulting macro-model has more matched poles and is more accurate than the one using the flat projection. It is also passive and 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 macro-model building and simulation respectively, and reduces waveform error by 33X.*

1. INTRODUCTION

The 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 [1] that is globally irregular and locally regular [2] 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 macro-models. In [3], internal sources are eliminated to obtain a macro-model with only external ports. The entire gird is partitioned at and connected by those external ports. Because elimination results in a dense macro-model, [3] applies an additional sparsification procedure that is error-prone and inefficient. Alternative approach to obtain a macro-model is to use projection based model order reduction (MOR) such as PRIMA [4]. The reduced model by PRIMA by a projection matrix with order q can match $n = \lfloor q/n_p \rfloor$ block moments (n_p is the port number). PRIMA can be implemented in a fashion of iterative path-tracing to efficiently solve *tree* structured P/G grids [5]. However, it is inefficient to be directly 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 macro-model is dense, which slows down simulation when the port number is large. 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, all these approaches [4, 6] use a flat projection that leads to the loss of the structure information of the state matrices. For example, the original state matrices may be sparse, but they become dense after flat projection. The resulting macro-model, therefore is too dense to be efficiently factorized in the time/frequency-domain simulation.

A recent method BSMOR [7] leverages the sub-block structure in $\mathbf{G}$ and $\mathbf{C}$ matrices. After obtaining a flat projection matrix by PRIMA, BSMOR constructs a new block-diagonal structured projection matrix accordingly. Its projection results in a macro-model with more matched poles than PRIMA, and hence an improved accuracy. Moreover, as the projection preserves structure, the reduced macro-model is sparse and can be solved by a two-level analysis. However, [7] uses a genetic block-based structure, the system poles are not only determined by those blocks in the diagonal part of $\mathbf{G}$ and $\mathbf{C}$. As a result, the additional matched poles are not accurate. Moreover, it assumes that all blocks have same size, which is not compact and optimum as discussed in Section 3 and 4 later on. In addition, same as [4], it orthnormalizes the entire state matrices to obtain the projection matrix. As a result, it is inefficient for large sized circuits.

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 dormant 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 mo-

*Hao Yu, Yiyui Shi, and Lei He are with Department of Electrical Engineering, University of California, Los Angeles, CA 90095 USA (e-mail: {hy255,yshi,lhe}@ee.ucla.edu).
This paper is partially supported by NSF CAREER award CCR-0093273/0401682 and a UC MICRO grant sponsored by Analog Devices, Intel and Mindspeed. Address comments to lhe@ee.ucla.edu.

ments by PRIMA decreases as the port number increases. Hence our approach has improved accuracy for circuits with large number of ports.

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 network. Each cluster is called a *compact block* with a unique pole distribution and a projection matrix accordingly. Because clustering reduces the redundant block information, the block-based form in our method is more compact than that in BSMOR.

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. A block-diagonal structured projection matrix is constructed by stacking projection matrices for individual diagonal blocks in the triangular system. 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. Moreover, the mq poles are exactly matched in TBS but not in BSMOR.

In addition, as discussed in Section 5, because the projection preserves the structure, the obtained macro-model by TBS is intrinsically sparse, and does not need the LP-sparsification used in [3]. The obtained macro-model by TBS also enables a two-level analysis similar to [8] to reduce simulation time in both frequency and time domains. In contrast, the reduced model by PRIMA and HiPRIME is dense and can not be analyzed in a fashion of two-level analysis. 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 domain is

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

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, and $\mathbf{B}$ and $\mathbf{L}$ ($\in R^{N\times n_p}$) are input/output port incident matrices with n_p ports.

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. PRIMA [4] finds a projection matrix V ($\in R^{N\times n}$). It has dimension q and its columns span n-block ($n = \lceil q/n_p \rceil$) 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 generating matrices* are $\mathbf{A} = (\mathbf{G}+s_0\mathbf{C})^{-1}\mathbf{C}$ and $\mathbf{R} = (\mathbf{G} + s_0\mathbf{C})^{-1}\mathbf{B}$, and s_0 is the expansion point that ensures $\mathbf{G} + s_0\mathbf{C}$ is nonsingular. 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 $\hat{\mathbf{L}} \in R^{q\times n_p}$. 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 $\hat{H}(s)$ and $H(s)$.*

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 n_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* $\mathbf{J} = \mathbf{B}u(s)$ and to directly match the moment of output $\mathbf{x}(s) = (\mathbf{G} + s\mathbf{C})^{-1}\mathbf{J}$ with the input vector $\mathbf{J}$. As a result, the matched moments of the output with input $\mathbf{J}$ is q that is independent on the port number n_p. This is because an MIMO system with right-hand-side $\mathbf{B}u$ can be transformed into the superposed SISO systems with the input $\mathbf{J}$. The following Theorem has been proved.

THEOREM 2. *Assume an MIMO system with unit-impulse current source u, and define the excitation current vector $\mathbf{J} = \mathbf{B}u$, where $u \in R^p$ and $\mathbf{J} \in R^N$. When the q columns of projection matrix V are obtained, the reduced response at the output $\hat{\mathbf{x}}(s) = (\hat{\mathbf{G}}+s\hat{\mathbf{C}})^{-1}\hat{\mathbf{J}}$ ($\hat{\mathbf{J}} = V^T\mathbf{J}$) matches the first q moments of the original $\mathbf{x}(s) = (\mathbf{G} + s\mathbf{C})^{-1}\mathbf{J}$.*

Note that the following two systems have the same output $x(s)$

$$(\mathbf{G} + s\mathbf{C})\mathbf{x}(s) = \mathbf{B}u(s), \quad (\mathbf{G} + s\mathbf{C})\mathbf{x}(s) = \mathbf{J}(s). \tag{5}$$

In addition, $\mathbf{J}$ can be decomposed into several 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 $\mathbf{J}_i$, it is equivalent to excite an SISO system with input $\mathbf{J}_i$. Therefore, $\hat{\mathbf{x}}_i(s)$ 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 [4] matches the block moment of transfer function with input matrix $\mathbf{B}$.

Moreover, we have

COROLLARY 1. *With the input $\mathbf{J}$, the first q dominant poles of $\mathbf{x}(s)$ are matched by $\hat{\mathbf{x}}(s)$.*

Using excitation current vector J as input, the first q moments are identical for $\mathbf{x}(s)$ and $\hat{\mathbf{x}}(s)$. So does the first q dominant poles.

Because the typical P/G girds contains large number of ports, in this paper the MOR is performed to match the moment of output $\mathbf{x}(\mathbf{s})$ with the input $\mathbf{J} = \mathbf{B}u$, similar to [10, 6].

3. COMPACT BLOCK FORMULATION

To handle large sized P/G grids and generate an accurate and sparse macro-model, we represent the original grid in compact blocks, where the overlap of pole distribution between 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 [2, 7]. The *ith* basic block has state matrices $\mathbf{G}_{ii}$ and $\mathbf{C}_{ii}$ with size n_i. Due to the heterogeneous structure of grids, each block can have different *RC* values. 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

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_{11} & \dots & \mathbf{G}_{1m_0} \\ \vdots & \ddots & \vdots \\ \mathbf{G}_{m_0 1} & \dots & \mathbf{G}_{m_0 m_0} \end{bmatrix} \mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \dots & \mathbf{C}_{1m_0} \\ \vdots & \ddots & \vdots \\ \mathbf{C}_{m_0 1} & \dots & \mathbf{C}_{m_0 m_0} \end{bmatrix}$$

and

$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_1 \dots \mathbf{J}_{m_0} \end{bmatrix}^T, \quad \mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \dots \mathbf{x}_{m_0} \end{bmatrix}^T. \tag{6}$$

In addition, $\mathbf{G}$ and $\mathbf{C}$ can be decomposed into two levels by

$$\mathbf{G} + s\mathbf{C} = \mathbf{Y}_0(s) + \mathbf{Y}_1(s). \tag{7}$$

The diagonal part $\mathbf{Y}_0(s)$ is: $\mathbf{G}_0 + s\mathbf{C}_0$, where

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

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 passive. 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

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

To be discussed in Section 4 and 5, the two level decomposition enables structure-preserving model order reduction and two-level analysis.

3.2 Clustering

To obtain a more compact block representation we propose a bottom-up clustering algorithm based on the dominant poles. The system timing response for each basic block can be approximately determined by its q dominant poles, i.e., the first q most dominant eigen-values or poles $(\lambda_1 \leq ... \leq \lambda_q)$. Poles are calculated from the eigen-decomposition of the *order reduced* moment matrix $\widetilde{\mathbf{A}} = \widetilde{\mathbf{G}}^{-1}\widetilde{\mathbf{C}}$ $(\in R^{q\times q})$. Note that when the excitation current vector is used for the moment matching of the output, the size q of the reduced model with the desired accuracy can be much smaller than the size of the original model. As a result, the cost of eigen-decomposition of reduced model is not high.

Precisely, for m_0 basic blocks, we calculate the first q dominant poles for each basic block by reducing it independently and finding its projection matrix V_i accordingly

$$span(V_i) = \mathcal{K}((\mathbf{A}_0)_i, (\mathbf{R}_0)_i, q) \quad i = 1, ..., m_0. \tag{8}$$

According to Theorem 2, using V_i, the reduced $\hat{\mathbf{x}}_i$ matches the first q moments of $\mathbf{x}_i$ with input $\mathbf{J}_i$. $\hat{\mathbf{x}}_i$ hence also matches the first q dominant poles of $\mathbf{x}_i$ according to Corollary 1.

Assume block i has $(\mathbf{G}_{ii}, \mathbf{C}_{ii}, \mathbf{J}_i)$. Its q-dominant-pole set is

$$\Lambda_i = eigen[(\widetilde{\mathbf{A}}_0)_i] = \{\lambda_1 \leq ... \leq \lambda_q\}$$

After merging block i with another block j and their interconnection $(\mathbf{G}_{ij}, \mathbf{C}_{ij})$, its q-dominant-pole set becomes

$$\Lambda_i' = eigen[(\widetilde{\mathbf{A}}_0)_i'] = \{\lambda_1' \leq ... \leq \lambda_q'\}$$

where $(\mathbf{A}_0)_i'$ is the new moment generation matrix for merged block.

Moreover, we define the *pole distance*. If Λ_i and Λ_j are two dominant-pole sets, $\lambda_m \in \Lambda_i$ and $\lambda_n \in \Lambda_j$, then the pole distance $d(\Lambda_i, \Lambda_j)$ is

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

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. 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. The clustering obtains m clusters of basic blocks, where m is decided by the structure of P/G grids and ϵ. We call cluster as a *compact block* in this paper. Accordingly, we can obtain a set of projection matrices: $V = [V_{1(n_1\times q)}, \dots, V_{m(n_m\times q)}]$, one for each compact block.

This interconnected compact block representation reduces the complexity of the original basic block representation as fewer number of blocks are need to represent the original system. Moreover, because the set of the first q dominant poles of each clustered block has minimum overlap. Note that because the original structured power grid shows heterogeneous structure that each region can have various *RC* values, the clustering algorithm will not converge to one entire circuit. This has been verified by experiments.

4. TBS MODEL ORDER REDUCTION

Although clustering results in m blocks each has the unique pole distribution, the poles of the entire grids are not only determined 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 [11]. With an additional block structured projection, the reduced blocks can match more poles than the flat projection.

4.1 Triangularization

The triangularization is based on introducing a replica block of $(\mathbf{G}, \mathbf{C})$, and moving those lower triangular blocks of $(\mathbf{G}_{ij}, \mathbf{C}_{ij})$ $(i < j)$ to the upper triangular parts at $(\mathcal{G}_{i,m+j}, \mathcal{C}_{i,m+j})$. The resulting *triangular system* has a *upper triangular* state

matrix $\mathcal{G}$

$$\mathcal{G} = \left[\begin{array}{cccc|cccc} \mathbf{G}_{11} & \mathbf{G}_{12} & \dots & \mathbf{G}_{1m} & 0 & 0 & \dots & 0 \\ 0 & \mathbf{G}_{22} & \dots & \mathbf{G}_{2m} & \mathbf{G}_{21} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{G}_{mm} & \mathbf{G}_{m1} & \mathbf{G}_{m2} & \dots & 0 \\ \hline \multicolumn{4}{c|}{\mathbf{0}} & \multicolumn{4}{c}{\mathbf{G}} \end{array}\right] \quad (9)$$

$\mathcal{C}$ has the similar structure as $\mathcal{G}$. The port matrix $\mathcal{B}$ and state variable x are

$$\mathcal{J} = \begin{bmatrix} \mathbf{J}_1 & \mathbf{J}_2 & \dots & \mathbf{J}_m & \mathbf{J} \end{bmatrix}^T,\ x = \begin{bmatrix} x_1 & x_2 & \dots & x_m & \mathbf{x} \end{bmatrix}^T.$$

where $\mathbf{J}$ and $\mathbf{x}$ are defined in (6).

The resulting triangular system equation is

$$(\mathcal{G} + s\mathcal{C})x(s) = \mathcal{J} \quad (10)$$

It is easy to verify that the solution x(s) from (10) is the same as $\mathbf{x}$(s) from (1).

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

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

Proof: The eigen-values of the triangular system is 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 $(\mathbf{G}_0)_i$ $(1 \le i \le m)$ and $\mathbf{G}$ 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 passive.

Note that directly solving (10) involves a similar cost to solve (1) as the replica block at the lower-right corner needs to be factorized first. As shown below, its benefits can be 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 clustering in Section 3.2, we can also obtain a set of projection matrices: $\{V_1, ..., V_m, V_{m+1}\}$, where V_i $(1 \le i \le m)$ is constructed for each block. Without using orthonormalization for replica block, V_{m+1} is obtained by

$$V_{m+1} = [V_1, ..., V_m] \quad (\in R^{N \times q}) \quad (11)$$

Furthermore, instead of constructing a flat projection matrix

$$V = [V_1, ..., V_m, V_{m+1}], \quad (\in R^{2N \times q}) \quad (12)$$

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

$$\mathcal{V} = diag[V_{1(n_1 \times q)}, ..., V_{m(n_m \times q)}, V_{m+1(N \times q)}] \quad (13)$$

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 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

$$\widetilde{\mathcal{G}} = \begin{bmatrix} \widetilde{\mathcal{G}}_A & \widetilde{\mathcal{G}}_B \\ \mathbf{0} & \widetilde{\mathcal{G}}_D \end{bmatrix}, \quad (14)$$

where

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

$$\widetilde{\mathcal{G}}_B = \begin{bmatrix} 0 & 0 & \dots & 0 \\ \mathcal{V}_{11}^T\mathbf{G}_{12}\mathcal{V}_{22} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{V}_{mm}^T\mathbf{G}_{m1}\mathcal{V}_{11} & \mathcal{V}_{mm}^T\mathbf{G}_{m2}\mathcal{V}_{22}^T & \dots & 0 \end{bmatrix}$$

$$\widetilde{\mathcal{G}}_D = \mathcal{V}_{m+1,m+1}^T\mathbf{G}\mathcal{V}_{m+1,m+1}. \quad (15)$$

Since BSMOR does not use triangularization, its system poles are not determined by those diagonal blocks. Therefore, its reduced macro-model does not exactly have mq poles matching (See Fig. 1 in 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 eigenvalues of the reduced blocks $(\widetilde{\mathbf{G}}_{ii}, \widetilde{\mathbf{C}}_{ii})$ $(\in R^{q \times q})$, i.e.,*

$$|(\widetilde{\mathbf{G}}_{00})_1 + s(\widetilde{\mathbf{C}}_{00})_1| \cup ... \cup |(\widetilde{\mathbf{G}}_{00})_m + s(\widetilde{\mathbf{C}}_{00})_m| = Null, \quad (16)$$

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 a 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 [4]. 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 shown 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

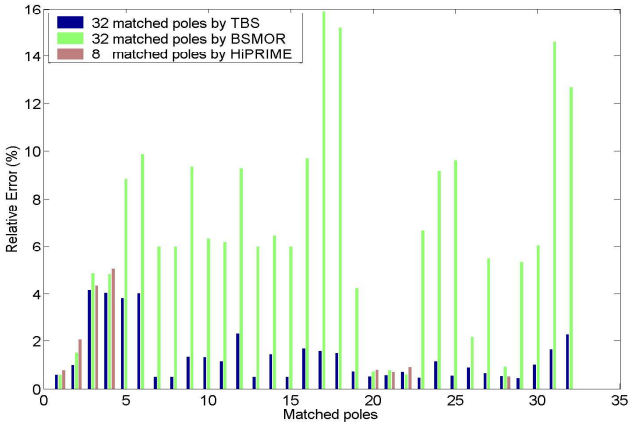


Figure 1: Pole matching comparison: mq poles matched by TBS and BSMOR, and q poles matched by HiPRIME.

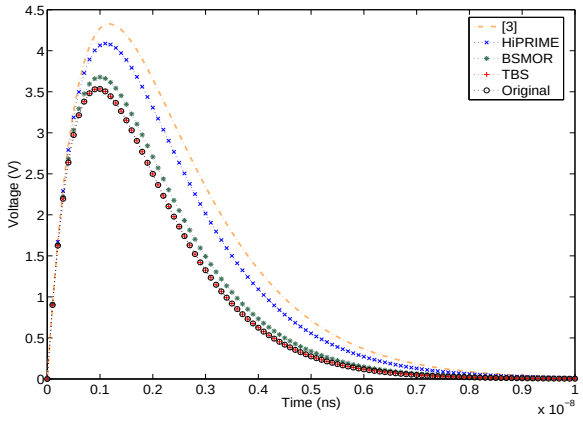


Figure 3: Comparison of time-domain responses between HiPRIME, BSMOR, [3], TBS and the original. TBS is identical to the original.

PRIMA and HiPRIME, the resulted $\hat{\mathbf{G}}$ is

$$\hat{\mathbf{G}} = \sum_{i=1}^{m+1}\sum_{j=1}^{m+1} {V_i}^T \mathbf{G}_{ij} V_j, \tag{17}$$

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 analysis similar to [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.

Consider the system equation for the reduced model

$$\widetilde{\mathcal{Y}} x = \widetilde{b}. \tag{18}$$

In frequency domain at a frequency point s, (18) 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, (18) 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 two level analysis similar to [8].

$$x = P^{(0)} - PQ \tag{19}$$

where

$$P^{(0)} = (\widetilde{\mathcal{Y}}_0)^{-1}\widetilde{b}, \quad P = (\widetilde{\mathcal{Y}}_0)^{-1}\widetilde{\mathcal{Y}}_1, \quad Q = (I+P)^{-1}P^{(0)}. \tag{20}$$

To avoid explicit inversion, LU or Cholesky factorization needs to be 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.

6. EXPERIMENTS

We implemented the TBS on a Linux workstation (P4 2.66GHz, 1Gb RAM). The RC mesh structures of the P/G grid are generated from realistic 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 HiPRIME [6], BSMOR [7] and [3]. 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 [3], 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 [3]. In the comparison of the macro-model 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 macro-models for TBS and [3] have the similar size and sparsification ratio.

6.1 A Non-uniform Structured RC Mesh

We use a non-uniform RC mesh (size 1M) with 32 same sized basic blocks and 32 unit-impulse current sources located at centers of basic blocks. Each basic block has a different magnitude of RC values. The number of connections between any 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 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 non-zero pattern of the conductance matrix before triangularization in Fig. 2 (a), after trian-

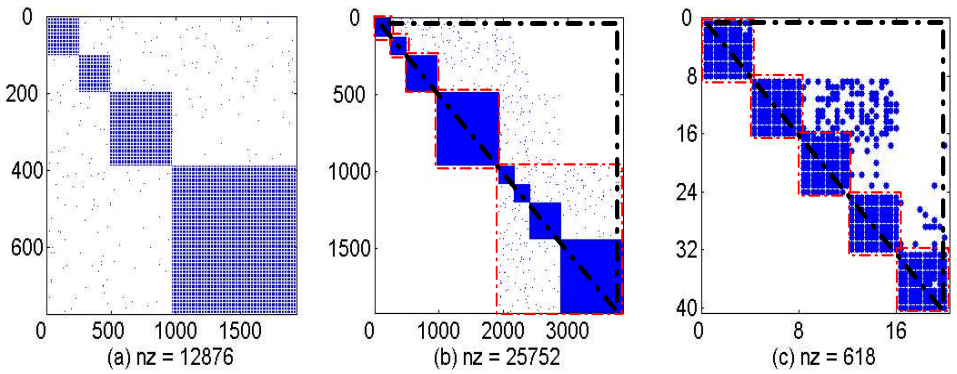


Figure 2: Nonzero (nz) pattern of conductance matrices: (a) original system (b) triangular system (c) reduced system by TBS. (a)-(c) have different dimensions, but (b)-(c) have the same triangular structure and same diagonal block structure.

node (N)	port (n_p)	order (q)	TBS (m=4)	HiPRIME	BSMOR (m=4)	[3]
768	12	8	5.03e-7	9.09e-6	4.87e-6	5.54e-6
$7.68K$	80	40	1.84e-6	2.31e-5	7.93e-6	1.21e-5
$76.8K$	120	60	3.02e-5	6.82e-4	1.91e-4	1.31e-2
$768K$	200	100	1.27e-4	9.67e-3	4.23e-3	6.01e-2
$7.68M$	1200	200	3.01e-3	9.97e-2	5.10e-2	0.11

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

gularization 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 non-zero pattern, which verifies that TBS preserves the triangular 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 [3] 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 in average. This is because the system poles of triangular are determined by its diagonal blocks. With a structure-preserving model order reduction, the reduced triangular system by TBS can exactly match mq poles of the original system. In contrast, the reduction in BSMOR does not have the triangular structure, and hence its approximated mq poles are less accurate than those obtained by TBS.

Fig. 3 compares the time-domain response at one port for HiPRIME, BSMOR, [3], 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 shows up to 36% error due to much

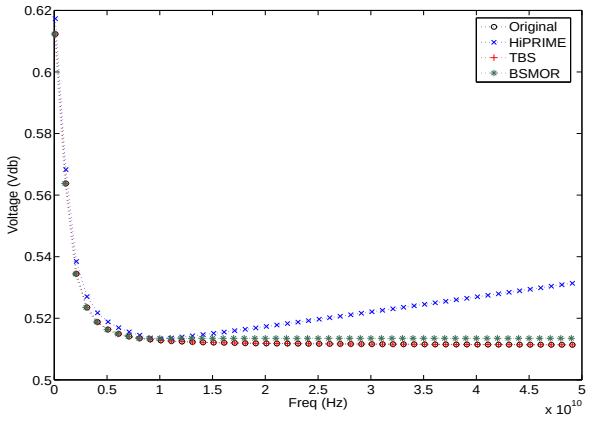


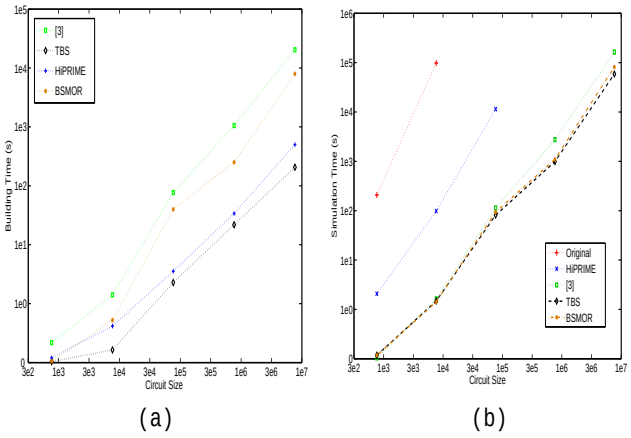
Figure 4: Comparison of frequency-domain responses between HiPRIME, BSMOR, TBS and the original. TBS is identical to the original.

fewer matched poles, and [3] shows up to 64% error due to the sparsfication. As mentioned before, the projection matrix constructed by BSMOR uses 4 uniform block each with the same size. As a result, it is not optimum to match poles and results in up to 23% error. Fig. 4 further shows the frequency-domain response under an impulse input. Using the 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 shows non-negligible deviation beyond 10GHz.

6.2 Scalability Study

We first study the runtime time scalability of reduced macro-model by HiPRIME, BSMOR, the method from [3] and TBS. The runtime time here includes both the macro-model building time and macro-model simulation time (time-domain). All reduced state matrices are constructed the similar accuracy.

Fig. 5 (a) compares the macro-model building time. As [3] needs the additional LP-based sparsification, it is inefficient for large sized P/G grids. For example, for a RC-mesh with



(a) (b)

Figure 5: Comparison of runtime under the similar accuracy. (a) macro-model building time (log-scale) comparison; (b) macro-model time-domain simulation time (log-scale) comparison.

size of $7.68M$, the method in [3] needs $4hrs : 42mins : 38s$ to build a reduced macro-model with size $1K$ and sparsity 30%, but TBS only spends $2mins : 8s$ (133X speedup) to build the similar sized macro-model. Moreover, TBS also has 54X speedup than BSMOR ($1hr : 45mins : 30s$) 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. Fig. 5 (b) further compares the simulation time, where we also increase the port number when increasing the circuit size. Because HiPRIME still uses flat projection, it results in a dense macro-model that 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 macro-models. On the other hand, BSMOR, [3] and TBS enable the two-level analysis to handle larger circuits with sizes up to $7.68M$ and 1200 ports in similar runtimes. For a circuit with size $(76.8K)^2$ and 120 ports, TBS achieves 109X runtime speedup compared to HiPRIME.

In Table 1, we further study the accuracy scalability of reduced macro-model by HiPRIME, BSMOR, [3] 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 model is used as the measure of error. We use higher order reduced model (by 4X) as the base if the waveform of the original model is unavailable. We find that the accuracy of [3] degrades when a large sparsity ratio is needed, where LP optimization can not preserve accuracy. On the other hand, using moment matching based projection with preserved sparsity, TBS generates a macro-model with higher accuracy. For example, it has a 38X higher accuracy than [3] when reducing a $7.68M$ circuit to a $(1K)$ macro-model with 32% sparsity. For the same circuit, TBS is 17X more accurate than BSMOR due to the exactly mq-pole matching, and is also 33X more accurate than HiPRIME due to more matches poles.

7. CONCLUSIONS

In this paper, we have proposed an accurate and efficient TBS model order reduction method to verify the power integrity for large sized P/G grids in the time-domain. Using triangularization, we show that the original system is passively transformed into a form with upper triangular block structure, where system poles are determined only by m diagonal blocks, where 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 original system. Experiments show that the waveform error is reduced 33X compared to the flat projection method like PRIMA and HiPRIME, and 17X compared to BSMOR using user specified partition. Moreover, with a two-level organization the reduction and analysis in TBS can be performed for each block independently. Therefore, it reduces both macro-model building and simulation time. TBS is up to 54X faster to build macro-models than BSMOR, and up to 109X to simulate macro-models in time-domain than HiPRIME. In addition, as TBS preserves sparsity, it is up to 133X faster to build macro-models than [3].

8. REFERENCES

[1] S. Boyd, L. Vandenberghe, A. E. Gamal, and S. Yun, "Design of robust global power and ground networks," in *Proc. ACM Int. Symp. on Physical Design (ISPD)*, 2001.

[2] J. Singh and S. Sapatnekar, "Congestion-aware topology optimization of structured power/ground networks," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, pp. 683–695, 2005.

[3] M. Zhao, R. Panda, S.Sapatnekar, and D. Blaauw, "Hierarchical analysis of power distribution networks," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, pp. 159–168, 2002.

[4] A. Odabasioglu, M. Celik, and L. Pileggi, "PRIMA: Passive reduced-order interconnect macro-modeling algorithm," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, pp. 645–654, 1998.

[5] H. Su, K.Gala, and S. Sapatnekar, "Analysis and optimization of structured power/ground networks," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, pp. 1533–1544, 2003.

[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 Computer-Aided Design of Integrated Circuits and Systems*, vol. 26, no. 6, pp. 797–806, 2005.

[7] H. Yu, L. He, and S. Tan, "Block structure preserving model reduction," in *IEEE International Workshop on Behavioral Modeling and Simulation (BMAS)*, 2005.

[8] G.Guardabassi and A. Sangiovanni-Vincentelli, "A two level algorithm for tearing," *IEEE Trans. on Circuits and Systems*, pp. 783–791, 1976.

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

[10] J. M. Wang and T. V. Nguyen, "Extended Krylov subspace method for reduced order analysis of linear circuits with multiple sources," in *Proc. ACM/IEEE*

Design Automation Conf. (DAC), 2000.

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