Block Structure Preserving Model Reduction for Linear Circuit with Large Number of Ports

Under Review, Please Do not Distribute.

Hao Yu, Lei He, Sheldon X.D. Tan†
Electrical Engineering Department
University of California, Los Angeles 90095
{hy255, lhe}@ee.ucla.edu
†Electrical Engineering Department
University of California, Riverside 92521
stan@ee.ucr.edu

ABSTRACT

We propose a block structure preserving model reduction (BSMOR), which generalizes the structure preserving model order reduction (SPRIM). The blocks can be derived based on specific applications such as block current characterization of the substrate. Increasing block numbers leads to more matched poles or moments using the same Krylov space and also increases the sparse ratio of the state matrices of the resulting macro-model. Experiment shows that BSMOR has a 20X smaller reduction time than PRIMA does under a same error bound. To efficiently analyze the resulting macro-model with a large number of ports, we further propose a bordered-block diagonal (BBD) partitioning with a bottom-up hierarchical clustering (BBDC) where the macro-model is partitioned into a number of subset-port models, each with a manageable model size. With a similar accuracy, BBDC obtains 30X speedup compared to the original macro-model.

1. INTRODUCTION

VLSI circuits contain a number of highly structured components such as bus, power ground grid and substrate. These components can be modeled by passive networks with tremendous amount of circuit elements and large numbers of ports. To analyze such network efficiently, model order reduction [1–3] has been studied and used extensively in the past. Based on the Krylov subspace projection and congruence transformation, PRIMA [3] is the one widely used to efficiently generate an order reduced macro-model with preserved passivity. However, the produced macro-model by PRIMA is not compact as the order is usually "too high" [4] to achieve the specified accuracy. Furthermore, when the macro-model is represented by a multiple-input-multiple-output (MIMO) transfer function, it is usually dense and becomes inefficient to analyze when there are a large number of ports [5].

Alternative methods include the truncated balanced realization (TBR) [4], where the singular value decomposition (SVD) is used to truncate less dominant states and achieve a more compact model. However, it may be slow as several computationally expensive numerical techniques are used to diagonalize the overall state matrix and guarantee the passivity. Recently, a structure-preserving model reduction (SPRIM) is proposed in [6]. This approach partitions the state matrix in the MNA (modified nodal analysis) form into a natural 2×2 block matrices, i.e., conductance, capacitance, inductance, and adjacent matrices. Accordingly the projection matrix is partitioned and the number of its columns is doubled. As a result, SPRIM matches the twice poles of the models by using the projection matrix given by PRIMA. In addition, the block structure of state matrices is preserved and it facilitates the realization of the reduced model. However, such a simple 2×2 partition does not leverage the regularity of the aforementioned passive networks. As to reducing the complexity introduced by large number of ports, the explicit hierarchical decomposition of the network is usually applied [5, 7, 8]. The capacity of these methods need to be improved further.

In this paper, we propose a block structure preserving model reduction (BSMOR) method, which generalizes the structure-preserving model order reduction (SPRIM) [6]. The blocks can be derived based on specific applications such as block current characterization of the substrate in this paper. We show that increasing the block number leads to more matched poles or moments using the same Krylov space. In other words, BSMOR can lead to more efficient reduction under the same accuracy. In addition, BSMOR can also preserve the sparsity for the reduced block matrices, which gives further efficiency boost to constructing a MIMO macro-model. Note that the macro-model consists of order-reduced blocks, where each reduced block contains a subset of ports. To efficiently analyze a macro-model with a large number of ports, we further propose a bordered-block diagonal (BBD) partitioning and hierarchical and bottom-up clustering of reduced blocks. We call it as BBDC analysis.

The experiment shows that under the same accuracy, the reduction of our approach is 20X times faster than PRIMA for a circuit with 1M elements. Moreover, with a similar accuracy, the BBDC analysis is 30X faster compared to analyzing the original macro-model.

The rest of the paper is organized as follows. We present BSMOR and BBDC in Sections II and III, respectively. In Section IV, we apply our method to the substrate macro-modeling and noise analysis, and discuss how to find the block structure from the characterization of the block current. We present the experimental results in Section V, and conclude the paper in Section VI. Proofs of theorems will be included in a technical report.

2. BLOCK STRUCTURE PRESERVING MODEL REDUCTION

In this section, we present a block structure preserving model reduction (BSMOR) that implicitly uses the block structure information of the matrix during the reduction. We show that by increasing the block number, we can match more poles or moments using the same Krylov subspace, which is also confirmed by our experimental results. On top of this, we introduce the concept of the *structured Krylov subspace* to summarize our contribution.

2.1 Preliminary

Consider a modified nodal formulation (MNA) of the circuit equation in the frequency domain:

$$\mathcal{G}x(s) + s\mathcal{C}x(s) = \mathcal{B}i_p(s)$$
$$v_p(s) = \mathcal{B}^T x(s) \quad (1)$$

where $x(s)$ is the state variable vector, $\mathcal{G}$ and $\mathcal{C}$ ($\in R^{N\times N}$) are state matrices. $\mathcal{B}$ ($\in R^{N\times n_p}$) is

$$\mathcal{B} = [B \quad 0]^T, \quad (2)$$

a port incident matrix. Eliminating $x(s)$ in (1) gives

$$v_p(s) = H(s)i_p(s)$$
$$H(s) = \mathcal{B}^T(\mathcal{G} + s\mathcal{C})^{-1}\mathcal{B}, \quad (3)$$

where $H(s)$ is a multiple-input multiple-output (MIMO) transfer function. PRIMA finds a projection matrix V ($\in R^{N\times qn_p}$) such that its columns span the q-th block Krylov subspace $\mathcal{K}(\mathcal{A}, \mathcal{R}, q)$, i.e.,

$$spanV = \mathcal{K}(\mathcal{A}, \mathcal{R}, q), \quad (4)$$

where $\mathcal{A} = (\mathcal{G} + s_0\mathcal{C})^{-1}\mathcal{C}$, $\mathcal{R} = (\mathcal{G} + s_0\mathcal{C})^{-1}\mathcal{B}$, and s_0 is the expansion point that ensures $\mathcal{G}+s_0\mathcal{C}$ is nonsingular. The resulting reduced transfer function is

$$\hat{H}(s) = \hat{\mathcal{B}}^T(\hat{\mathcal{G}} + s\hat{\mathcal{C}})^{-1}\hat{\mathcal{B}}, \quad (5)$$

where

$$\hat{\mathcal{G}} = V^T\mathcal{G}V, \quad \hat{\mathcal{C}} = V^T\mathcal{C}V, \quad \hat{\mathcal{B}} = V^T\hat{\mathcal{B}}, \quad (6)$$

has the identical expanded first q-th moments with $H(s)$. It is called as the *Grimme's projection theorem* [9]. Note that $\hat{\mathcal{G}}$, $\hat{\mathcal{C}}$ are $\in R^{qn_p\times qn_p}$, and $\hat{\mathcal{B}}$ is $\in R^{qn_p\times n_p}$.

In [6], a structure-preserving reduced model order reduction technique, SPRIM, is proposed. The primary observation is that instead of using the Krylov subspace $\mathcal{K}(\mathcal{A}, \mathcal{R}, q)$ for the projection matrix $\widetilde{V}$, one can use any projection matrix such that the space spanned by the column in $\widetilde{V}$ contains the q-th block Krylov subspace. i.e.

$$\mathcal{K}(\mathcal{A}, \mathcal{R}, q) \subseteq \widetilde{V} \quad (7)$$

In SPRIM, a 2×2 partition is naturally used as a linear state matrix in the MNA form shows a 2×2 block structure

$$\mathcal{G} = \begin{bmatrix} G & A^T \\ -A & 0 \end{bmatrix}, \mathcal{C} = \begin{bmatrix} C & 0 \\ 0 & L \end{bmatrix}, \quad (8)$$

where G ($\in R^{n_1\times n_1}$), C ($\in R^{n_1\times n_1}$), L ($\in R^{n_2\times n_2}$) are conductance, capacitance and inductance matrix, and A ($\in R^{n_2\times n_1}$) is the adjacent matrix indicating the branch current flow at the inductor. Note that $n_1 + n_2 = N$.

Therefore, a structured projection vector $\widetilde{V}$ is constructed by partitioning the projection vector V obtained from the q-th PRIMA iteration

$$V = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \rightarrow \widetilde{V} = \begin{bmatrix} V_1 & 0 \\ 0 & V_2 \end{bmatrix}. \quad (9)$$

where V_1 is $\in R^{n_1\times qn_p}$, V_2 is $\in R^{n_2\times qn_p}$, and hence $\widetilde{V}$ is $\in R^{N\times 2qn_p}$. As a result, the number of columns in $\widetilde{V}$ is twice of that in V. Accordingly the new reduced state matrices are

$$\widetilde{\mathcal{G}} = \begin{bmatrix} \widetilde{G} & \widetilde{A}^T \\ -\widetilde{A} & 0 \end{bmatrix}, \widetilde{\mathcal{C}} = \begin{bmatrix} \widetilde{C} & 0 \\ 0 & \widetilde{L}, \end{bmatrix} \quad (10)$$

where $\widetilde{G} = {V_1}^T G V_1$, $\widetilde{A} = {V_2}^T A V_1$ and $\widetilde{C} = {V_1}^T C V_1$ and $\widetilde{L} = {V_2}^T L V_2$. Similarly, the size of $\widetilde{\mathcal{G}}$, $\widetilde{\mathcal{C}}$ ($\in R^{2qn_p\times 2qn_p}$), and $\widetilde{\mathcal{B}}$ ($\in R^{2qn_p\times n_p}$) is twice than that of $\hat{\mathcal{G}}$, $\hat{\mathcal{C}}$, and $\hat{\mathcal{B}}$ reduced by using V. Therefore, the moments of the reduced model with state matrices: $\widetilde{\mathcal{G}}$ and $\widetilde{\mathcal{C}}$ are twice than those of the reduced model with state matrices: $\hat{\mathcal{G}}$ and $\hat{\mathcal{C}}$. In other words, the reduced model by $\widetilde{V}$ matches $2q$ moments of the original model instead of q moments as the reduced model by V.

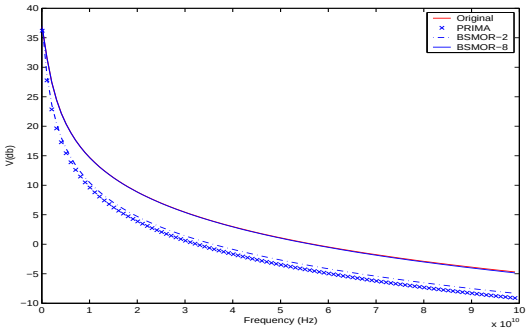

Figure 1: Frequency responses of the BSMOR, PRIMA, and original model at one port of a uniform mesh (256x256) after 10 iterations.

Since the reduced model is written in the first order form in (10), the model reduced by SPRIM is twice larger than that produced by PRIMA. But the reduced model produced by SPRIM preserves the structure of the original model and can be further reduced into the second-order form using node elimination base on Schur's decomposition [10]: $\widetilde{\mathcal{G}}_{NA} = \widetilde{G} + s\widetilde{C} + \frac{1}{s}\widetilde{A}^T\widetilde{L}^{-1}\widetilde{A}$ where $\widetilde{\mathcal{G}}_{NA}$ is the reduced state matrix in NA form, which has the same size of the reduced matrix by using V. But the difference is that each element in $\widetilde{\mathcal{G}}_{na}$ become second-order rational function of s instead of first-order polynomial of s.

Hence SPRIM algorithm essentially consists of two reduction steps: the first step is the structure-preserving projection-based reduction and the second step is block node elimination based on Schur's decomposition. As a result, SPRIM can match more poles than PRIMA, which uses V as the projection matrix, and they result in a same *size* of the reduced model.

If we just look at the first step, SPRIM simply matches more moments by using more columns in the projection matrix $\widetilde{V}$, thus produces larger reduced state matrices in the first-order form.

2.2 BSMOR Method

SPRIM essentially is based on a 2×2 partitioning of the state matrices. If we use more partitions (*each partition called a block*), we can add more columns into the project matrix $\widetilde{V}$, thus match more moments given the same Krylov space $\mathcal{K}(\mathcal{A}, \mathcal{R}, q)$.

Specifically, we assume that the conductance matrix $\mathcal{G}$ can be distinguished in m blocks

$$\mathcal{G} = \begin{bmatrix} \mathcal{G}_{1,1(n_1\times n_1)} & \mathcal{G}_{1,2(n_1\times n_2)} & \cdots & \mathcal{G}_{1,m(n_1\times n_m)} \\ \mathcal{G}_{2,1(n_2\times n_1)} & \mathcal{G}_{2,2(n_2\times n_2)} & \cdots & \mathcal{G}_{2,m(n_2\times n_m)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{G}_{m,1(n_m\times n_1)} & \mathcal{G}_{m,2(n_m\times n_2)} & \cdots & \mathcal{G}_{m,m(n_m\times n_m)} \end{bmatrix}, \quad (11)$$

where each block has the size n_k ($\sum_{k=1}^{m} n_k = N$). A similar block structure can be found for $\mathcal{C}$ matrix. Then, B becomes

$$\mathcal{B} = [\mathcal{B}_{1(n_1\times n_p)}, \quad \mathcal{B}_{2(n_2\times n_p)}, \quad \cdots \quad \mathcal{B}_{m(n_m\times n_p)}]^T \quad (12)$$

where each basic block contains user specified n_{p_k} ports ($n_p = \sum_{k=1}^{m} n_{p_k}$). Note that these blocks can be derived based on specific applications such as block current characterization of the substrate as discussed in Section 4.

Accordingly, we further partition the projection matrix V obtained from PRIMA according to the block structure in state matrices from (11)

$$V = \begin{bmatrix} V_{1(n_p \times n_1)} \\ V_{2(n_p \times n_2)} \\ \vdots \\ V_{m(n_p \times n_m)} \end{bmatrix}$$

$$\rightarrow \widetilde{V} = \begin{bmatrix} V_{1(n_p \times n_1)} & 0 & \dots & 0 \\ 0 & V_{2(n_p \times n_2)} & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & V_{m(n_p \times n_m)} \end{bmatrix}. \quad (13)$$

We call this as an $m \times m$ *Block Structure preserving Model Reduction* (BSMOR), where m is the number of blocks.

We can obtain the order reduced state matrices by projecting $\widetilde{V}$:

$$\widetilde{\mathcal{G}} = (\widetilde{V})^T \mathcal{G} \widetilde{V}, \quad \widetilde{\mathcal{C}} = (\widetilde{V})^T \mathcal{C} \widetilde{V}, \quad \widetilde{\mathcal{B}} = (\widetilde{V})^T \mathcal{B}. \quad (14)$$

Element wise, we have

$$\widetilde{\mathcal{G}}_{i,j} = {V_i}^T \mathcal{G}_{i,j} V_j \quad \widetilde{\mathcal{C}}_{i,j} = {V_i}^T \mathcal{C}_{i,j} V_j \quad \widetilde{\mathcal{B}}_i = {V_i}^T \mathcal{B}_i \quad (15)$$

where $\widetilde{\mathcal{G}}_{i,j}$ represents the blocks at i block row and j block column in reduced matrix $\widetilde{\mathcal{G}}$. So do $\widetilde{\mathcal{C}}_{i,j}$ and $\widetilde{\mathcal{B}}_i$. Let $V_i = V_{i(n_p \times n_i)}$ to simplify notations. Using such a matrix $\widetilde{V}$, we define a reduced-order model with the following transfer function

$$\widetilde{H}(s) = \widetilde{\mathcal{B}}^T (\widetilde{\mathcal{G}} + s\widetilde{\mathcal{C}})^{-1} \widetilde{\mathcal{B}}. \quad (16)$$

As a result, we have the following theorem regarding the block structure preserving model reduction

THEOREM 1. *Let $\widetilde{V}$ be a matrix that satisfies $\mathcal{K}(\mathcal{A}, \mathcal{R}, q) \subseteq span(\widetilde{V})$ and $\widetilde{V}$ is defined in Eq.(13). $\widetilde{H}(s)$ will match the first mq moments in the expansion of $H(s)$ about s_0.*

This result is the natural extension of 2×2 case given by SPRIM. If the number of columns in V is k, then the number of columns in $\widetilde{V}$ is mk. As a result, $\widetilde{\mathcal{G}}$ is m times larger than $\mathcal{G}$. Conceivably, $\widetilde{H}(s)$ has m times more eigenvalues than that of $\hat{H}(s)$. Based on the Grimme's projection theorem, $\widetilde{H}(s)$ should match m times more moments than $\hat{H}(s)$.

Similar to SPRIM, the reduced model of passive network obtained by Krylov-subspace projection preserves passivity:

THEOREM 2. *The reduced order model $\widetilde{H}(s)$ by BSMOR is passive.*

Based on the Theorem 1, one important observation is that, introducing more partitions or blocks can archive the same reduction accuracy by using smaller Krylov subspace, which can in turn improve the reduction efficiency. On the other hand, we observes that the partitioned projection matrix $\widetilde{V}$ leads to localized projection as shown by (15). In other words, the block projection matrix $\widetilde{V}_i$ is used only for matrix blocks $\mathcal{G}_{i,x}$ and $\mathcal{G}_{x,i}$, $(x = 1, ...m)$. In this sense, Krylov subspace given by $\widetilde{V}$ becomes a *structured* Krylov subspace in $\widetilde{V}$.

Each structured block projection matrix $\widetilde{V}_i$ will lead to the localized model order reduction for block i, which is represented by $\mathcal{G}_{x,i}$ and $\mathcal{G}_{i,x}$ matrix blocks $(x = 1, ...m)$. Conceivably, the order reduced block $\widetilde{\mathcal{G}}_{i,x}$ and $\widetilde{\mathcal{G}}_{x,i}$ will match $\mathcal{G}_{i,x}$ and $\mathcal{G}_{x,i}$ to the first q moments. But the whole system consisting of the m blocks will match mq moments.

In summary, by introducing the structured Krylov subspace, one can obtain order reduced models with more accurate for each structure block by using the same Krylov subspace base vectors, or get the same order reduced model (same accuracy) using a smaller Krylov subspace. Therefore, BSMOR provides much more flexibility and trade-off between efficiency and model accuracy for reducing linear dynamic system models than PRIMA.

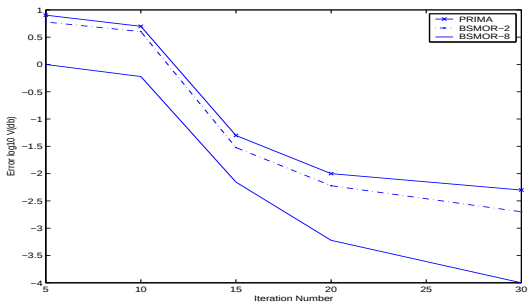

Figure 2: Maximum errors of the frequency response of the BSMOR and PRIMA for increasing order models of a uniform mesh (256x256) up to 20GHz.

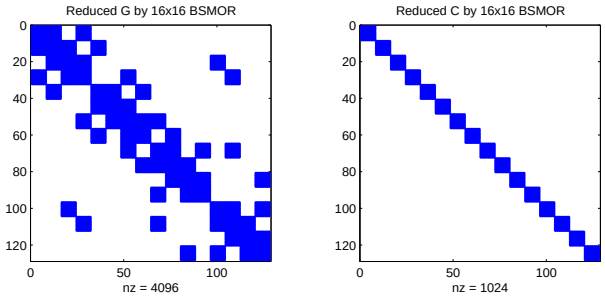

Figure 3: Non-zero patterns for G, C matrices of a uniform RC-mesh (256x256) after a 16×16 BSMOR reduction with 8 iterations, where NZ is the number of non-zero.

For a 256x256 RC-mesh (320K circuit elements), Fig. 1 compares frequency responses at one port between the original circuit and reduced models by PRIMA, 2×2 BSMOR, and 8×8 BSMOR. Clearly, with 10th iteration the 8×8 BSMOR is identical to the original circuit response but PRIMA and 2×2 BSMOR are still not converged. Fig. 2 further compares the maximum error of frequency responses by PRIMA, 2×2, and 8×8 BSMOR vs. the iteration number during the reduction. In the same iteration, it shows that using more partitions (block number) to construct the projection matrix can have better accuracy than using less partitions as PRIMA does. In other words, BSMOR can generate more compact model with improved pole matching ability.

Moreover, due to the structured construction of $\widetilde{V}$ by (13), BSMOR preserves the structure and sparsity of $\widetilde{\mathcal{G}}$, $\widetilde{\mathcal{C}}$ matrices even after the reduction. For example, for the 256x256 RC-mesh above, Fig. 3 shows the structure of these two state matrices before and after a 16×16 BSMOR reduction. The $\widetilde{\mathcal{G}}$, $\widetilde{\mathcal{C}}$ matrices show 72% and 93% sparsification ratio, respectively. It is another advantage to use BSMOR other than PRIMA, as PRIMA generates a fully dense state matrices after the reduction. Moreover, the sparsification ratio increases when increasing the block number. It is not surprising as conceptually when a block contains only one element, the "reduced" state matrices become exactly the same as the original sparse state matrices.

3. BORDERED-BLOCK DIAGONAL PARTITIONING WITH HIERARCHICAL CLUSTERING

In this section, we first describe the presentation of the flat macro-model generated by the reduced state matrices from Section 2.2. To efficiently handle the flat macro-model with large

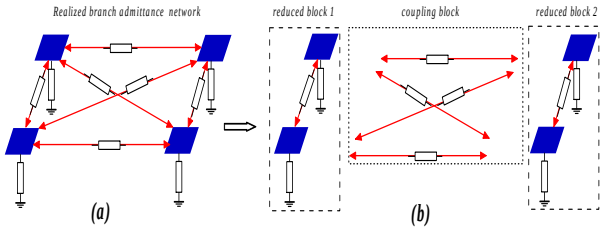

Figure 4: An example of 4-port admittance with 2 reduced blocks. (a) realization in branch admittance network; (b) represented by 2 reduced blocks with an additional coupling block.

number of ports, we present the *bordered-block diagonal* (BBD) partitioning to solve each block individually. Moreover, we discuss a hierarchical clustering method to further improve the efficiency.

3.1 Flat Macro-model

It is usually inconvenient to directly stamp back the reduced $\widetilde{\mathcal{G}}, \widetilde{\mathcal{C}}$ matrices. Moreover, for the frequency-dependent application in the analog/RF simulation like the substrate noise analysis, an Y-parameter based multiple port macro-model is widely used instead. An $n_p \times n_p$ MIMO admittance matrix $Y'(s)$ can be obtained by taking the eigen-decomposition of $\widetilde{A} = (\widetilde{\mathcal{G}} + s_0\widetilde{\mathcal{C}})^{-1}\widetilde{\mathcal{C}}$

$$Y'(s) = \begin{bmatrix} Y'_{1,1} & \cdots & Y'_{1,n_p} \\ \vdots & \ddots & \vdots \\ Y'_{n_p,1} & \cdots & Y'_{n_p,n_p} \end{bmatrix}, \tag{17}$$

with

$$Y'_{i,j} = c^{i,j} + \sum_{n=1}^{q} \frac{k_n^{i,j}}{s - p_n}, \tag{18}$$

where k_n and p_n are the residues and poles. Note that eigenvalues of $\widetilde{A}^{(q)}$ represent the reciprocal poles of $Y'(s)$ [3]. Due to the preserved sparsity, the eigen-decomposition becomes more efficient when using the $\widetilde{\mathcal{G}}$ and $\widetilde{\mathcal{C}}$ from the BSMOR other than using those from PRIMA. Furthermore, as the reduction preserves the structure, it results in additional preservations: *i*) the reciprocity of the network is also preserved, i.e., the $Y'(s)$ is symmetrical. In contrast, PRIMA does not preserve this property; *ii*) the block structure is preserved as well. It means the reduced block can be distinguished by a subset of ports specified before BSMOR. Due to the preserved block structure , we can further apply an additional port-partitioning, precisely, bottom-up port clustering to handle the large number of ports as discussed later on.

Note that the runtime and memory requirement to solve a linear system is primarily determined by the size, sparsity, and structure of the matrix. As shown in (17), the reduced model is represented by a $n_p \times n_p$ admittance matrix. Each entry represents the coupling between a pair of two ports and there are $O(n_p^2)$ of such couplings. As a result, the model reduction results in a large admittance matrix such that the efficiency of the reduced model degrades and the available memory of computing resources becomes insufficient as the sparse matrix solution becomes unavailable. In the following, we further discuss a partitioned solution to handle the admittance matrix with large number of ports. Using partitioning, the large coupled network is divided into subnetworks with manageable size and solved by blocks individually [11]. Moreover, partitioning can also be employed when network consists of repetitive identical subnetworks so that only one equation needs to be stored.

To partition a given network, we need first realize the admittance matrix $Y'(s)$. We give the realization theorem below

THEOREM 3. *If the nodal admittance matrix $Y'(s)$ has reciprocity, it can be realized by a branch admittance network using following transformation:*

$$Y_{ii} = \sum_{j=1}^{n_p} Y'_{ij}, \qquad Y_{ij} = -Y'_{ij}. \tag{19}$$

With such a nodal-to-branch transformation, the flat macro-model consists of m order reduced blocks, where each *reduced block* contains n_{p_k} ports with ground and coupling branch admittances. There are also exist coupling branch admittances between any pair of reduced blocks. A realized branch admittance network for a 4-port admittance matrix is shown in Fig. 4 (a). To partition the branch admittance network Y, one natural approach is to reserve each reduced block, and pack all the coupling branch admittances into one block, called as *couping block*, that connects all reduced blocks. An example of such a partitioning (or representation of the macro-model from BSMOR) is shown in Fig. 4 (b) for a 4-port admittance matrix.

3.2 Bordered-Block Diagonal Matrix

For the kth reduced block, we have

$$\mathbf{Y}_k v_k = i_k + \widetilde{i}_k, \tag{20}$$

where

$$(\mathbf{Y}_k)_{ii} = \sum_{j=1}^{n_{p_k}} Y'_{ij}, \qquad (\mathbf{Y}_k)_{ij} = -Y'_{ij} \quad (j \in n_{p_k}), \tag{21}$$

and v_k, i_k are the port voltage and current vectors, where i_k is part of i_p: $i_k = i_p(\ldots \underbrace{i_{k1} \ldots \ldots}_{n_{p_k}})$. Moreover, $\widetilde{i}_k$ is the *correlation current* from the other reduced block through the coupling block.

The branch equation for the coupling block is

$$(\mathbf{Y}_0)^{-1} i_0 = v_0, \tag{22}$$

where $\mathbf{Y}_0$ is the branch admittance matrix describing the branches in the coupling block. It is a diagonal matrix such that its inversion is easily obtained as $1/(\mathbf{Y}_0)_{ii}$. Note that its size depends on the number of couplings among reduced blocks, and it can be efficiently implemented with the sparse matrix data structure. v_0 and i_0 are branch voltage and current vectors. They relate to the port voltage/current vectors v_k/i_k at kth block by

$$\widetilde{i}_k = C_{k0} i_0, \qquad v_0 = -\sum_{k=1}^{m} (C_{k0})^T v_k, \tag{23}$$

where C_{k0} is the cut matrix composed by $\{0, 1, -1\}$ to indicate the direction of branch currents between kth reduced block and the coupling block. For example, the C_{k0} for reduced blocks in Fig. 4 are

```
C(1,0):                    C(2,0):
 1    1    0    0           -1    0   -1    0
 0    0    1    1            0   -1    0   -1
```

Combine (21) - (23), we have the following hybrid matrix equation

$$\begin{bmatrix} \mathbf{Y}_1 & 0 & \ldots & 0 & C_{10} \\ 0 & \mathbf{Y}_2 & \ldots & 0 & C_{20} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & \mathbf{Y}_{m^{(l)}} & C_{m0} \\ (C_{10})^T & (C_{20})^T & \ldots & (C_{m0})^T & -(\mathbf{Y}_0)^{-1} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \\ i_0 \end{bmatrix} = \begin{bmatrix} i_1 \\ i_2 \\ \vdots \\ i_m \\ 0 \end{bmatrix}.$$

This hybrid matrix shows a *bordered-block-diagonal* (BBD) structure. It enables the following algorithm (Algorithm 1) to solve each reduced block individually without using the explicit inversion. Each reduced block matrix is fist solved individually with LU factorization and substitution (1.1-1.5), the results from each reduced block are then used further to solve the coupling block (2.1-2.4), and the final v_k of each reduced block is updated (3.1-3.4) with the result from the coupling current i_0.

Algorithm 1 Analysis of bordered-block-diagonal (BBD) matrix

1.Solve Y_k individually
for every k *in* m **do**
(1.1) input: Y_k, C_{k0}, i_k;
(1.2) factor: $Y_k = L_k U_k$;
(1.3) solve: $L_k \Phi_k = C_{k0}$ for Φ_k, $\Psi_k U_k = (C_{k0})^T$ for Ψ_k, and $L_k \xi_k = i_k$ for ξ_k;
(1.4) form: $F_k = \Phi_k^T \Psi_k$, and $G_k = \Psi_k^T \xi_k$
(1.5) output: F_k, G_k.
end for
2.Solve Y_0
(2.1) input: Y_0, F_k, G_k;
(2.2) form: $F = Y_0^{-1} + \sum_{k=1}^{m} F_k$, $G = \sum_{k=1}^{m} G_k$;
(2.3) solve: $F i_0 = G$ for i_0;
(2.4) output: i_0.
3.Update Y_k individually
for every k *in* m **do**
(3.1) input: i_0, Φ_k, ξ_k, U_k;
(3.2) form: $\xi_k = \xi_k - \Phi_k i_0$;
(3.3) solve: $U_k v_k = \xi_k$ for v_k;
(3.4) output: v_k.
end for

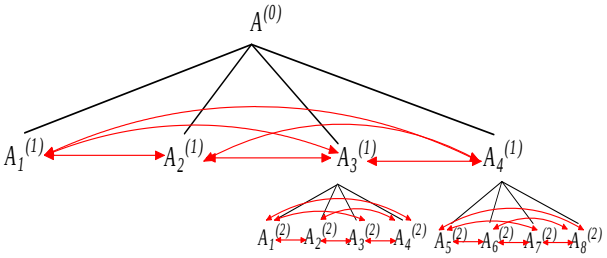

Figure 5: The hierarchical tree structure of clustered blocks.

Typically, LU factorization requires $n^3/3$ multiplications and back/forward substitution requires $n^2/2$ multiplications. The computational cost of Algorithm 1 is therefore, $\sum_{k=1}^{m}(n_{p_k}^3/3 + n_{p_k}^2/2) + (n_0^3/3 + n_0^2/2)$, where n_{p_k} is the port number (reduced block size) of each reduced block, and n_0 is the size of the coupling block. Note that if the parallel execution is used, the *summation* becomes the *maximum* execution time among blocks. To reduce the computational cost even without using the parallel execution, we need control the cost of from both the individual reduced block and the coupling block as discussed below.

3.3 Hierarchical Clustering

As the factorization cost decreases with the size of the reduced block, apparently the computation cost will be small when the network is directly partitioned based on the reduced basic block from BSMOR. However, the size of $\mathbf{Y}_0$ increases with the reduced block number, and it will increase the computation cost. To wisely arrange this trade-off, a hierarchical tree structure is used as shown in Fig. 5. In this tree, each node represents an *abstract block*. There are links connecting each pair of correlated blocks, representing inter-block couplings. The tree is constructed by iteratively *clustering* the reduced blocks from the bottom. The degree and the level is chosen to bound the size of the coupling block below a threshold. At the leaf level, a cluster of reduced blocks are siblings of a parent node, an abstract block. A *cluster-coupling block* is introduced to model the coupling between siblings. There is no direct coupling between abstract blocks not in a same cluster, but their coupling is modeled by cluster-coupling blocks for parent nodes. Therefore, we can maintain a constant link number (couplings) at each tree level. Note that the following *merge operation* is operated when two blocks k and l are clustered

$$i_{new} = [i_k, i_l], \quad v_{new} = [v_k, v_l],$$

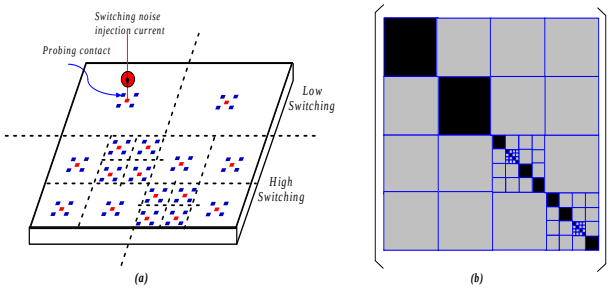

Figure 6: (a) The non-uniform substrate mesh network characterized by the switching current density; (b) The corresponding block structure of conductance/capacitance matrices.

and

$$(\mathbf{Y}_{new})_{ii} = \sum_{j \in n_{p_k} \cup n_{p_l}} Y'_{ij}, (\mathbf{Y}_{new})_{ij} = -Y'_{ij}.$$

At the bottom level, we solve each clustered block using Algorithm 1. It would be inefficient to calculate v_k directly on the higher levels since the block size get larger and larger. Fortunately this is not necessary, because one can use the already calculated v_k of the children, same as to attach the voltage sources to the coupling block at parent node. To do this we need to update i_0 from $(l-1)$th level to lth level by

$$v_0^{(l-1)} = -\sum_{k=1}^{m^{(l-1)}} (C_{k0}^{(l)})^T v_k^{(l-1)}, \quad i_0^{(l-1)} = Y_0^{(l)} v_0^{(l-1)}, \tag{24}$$

and then solve v_k at lth level by (3.1)-(3.4). Moreover, with the hierarchical tree structure, v_k is recursively updated by a bottom-up depth-first traversal of the tree, described below:

```
HPtree(Ai){
    if (Ai is leaf)
        return;
    for (each child k of Ai) {
        HPtree(Ai.k);
    }
    Update(Ai);
}
```

where we assume that the cut matrices and block branch admittance are pre-computed and stored hierarchically. Note that the factorization cost of large matrix at the top level is large. We further apply an error-bounded sparsification technique similar as [7] to the branch admittance matrix. As the sparsification is performed at the top level, this error is bounded. For simplicity of presentation, we call BBD analysis with hierarchical clustering as BBDC analysis.

4. APPLICATION

In this section, we discuss the application of BSMOR and BBDC analysis to the substrate macro-modeling and noise analysis. The substrate outside of active/contact areas can be treated as a uniformly doped layer, where an electrostatic Maxwell's equation is:

$$\epsilon \frac{\partial}{\partial t}(\nabla \cdot E) + \frac{1}{\rho}(\nabla \cdot E) = 0. \tag{25}$$

The Eddy current term (the primary cause of substrate loss) can be ignored if the substrate is highly doped, where the conduction current is dominant. Note that (25) can be discretized in differential form using finite-difference [12] or integral form using boundary element (BEM) methods. Because the BEM method needs

to find a numerically stable multi-layer Green's function [13], it is not trivial to be constructed in general when the layout geometry becomes arbitrary. In this paper, the finite-difference based discretization is used to generate the RC mesh/grid as the substrate circuit model. As the electric field varies nonlinearly as a function of the distance, the finite-difference method approximates this variation as a piecewise constant function by carefully choosing the pitch of the mesh according to the current density, i.e., the strength of the electrical field.

For leading-edge integrated circuits, the count of gate is typically in millions. The number of possible locations to place contacts of sensitive analog/RF circuits is large as well. Therefore, a flat multi-port description of each individual substrate noise injector and receptors is impractical. We assume that the chip is partitioned into smaller circuit, i.e., blocks based on the switching current density. As a result, within a block all noise current injections can be clustered into one independent current source at one single injection port. Such a block maximum current spectrum envelope is studied in [14, 15] to characterize the injection noise sources in a bottom-up fashion. The noise current injected by the gate G at frequency f_p is denoted $i_G(f_m)$, and $f_m = m \times f_0$ $(m = 0, 1, 2, ...M)$, where f_0 is the clock frequency and M is the sampling bound. Then, the total noise current of c_N gates in kth block is

$$i_{Ck} = \sum_{k=1}^{C_N} i_{Gk}(f_m), \tag{26}$$

and by a library-based characterization of the primary input transition vp, the block current envelope spectrum is found by

$$i_k^{max}(f_m) = \max_{vp} |i_{Ck}(f_m)|. \tag{27}$$

Therefore, if there are m characterized blocks, each block would contain n_{p_k} user specified ports, including one input port representing the injecting current noise source according to the above block current assumption, and $(n_{p_k} - 1)$ output ports representing all possible contact locations for analog/RF modules. There are total n_p $(n_p = \sum_{k=1}^{m} n_{p_k})$ specified ports. The port current vector i_p becomes

$$i_p = [\underbrace{i_1^{max}...0}_{n_{p_1}} \underbrace{i_k^{max}...0}_{n_{p_k}} \underbrace{i_m^{max}...0}_{n_{p_m}}], \tag{28}$$

where all ignored entries are zeros standing for probing output ports. Note that the propagated noise is observed from v_p.

However, with the use of the power management technique like the clock gating, the $i_{Ck}(f_m)$ can be very non-uniform for each block across the chip. For the block with the high current density, the electric field tends to vary largely, and a finer grids are necessary for the accurate approximation. Otherwise, coarse grid is used instead. For example, the substrate plane in Fig. 6 (a) have 4 parts with different switching current densities and it results in a non-uniform mesh structure.

For the RC mesh/grid, we have

$$\mathcal{G} = G, \quad \mathcal{C} = C, \quad \mathcal{B} = B. \tag{29}$$

As a result, they demonstrate a block structure according to the block current density. For example, Fig. 6 (b) shows such a block structure for the block current distribution in Fig. 6 (a).

After specifying the block structure according to the switching current density with the specified port information, the BSMOR can be applied to obtain a order-reduced MIMO macro-model. When the port number is large, the BBDC can be further applied to solve each reduced block individually

Note that with a given spectrum of the maximum current envelope for injection sources, and an efficient macro-model of the substrate, we can calculate the maximum noise spectrum efficiently at the victim. Because the G, C matrices are symmetric and positive definite (s.p.d.), the substrate RC-network shows the monotonity. Therefore, it is obvious to observe that the spectrum of the maximum noise voltage in the frequency domain can be obtained by examining each probing output port. Its impact on analog/RF victim, therefore, can be obtained by adding an equivalent noise voltage source. The time-domain voltage profile can be obtained by IFFT (inverse fast Fourier transformation) with sufficient sampling points.

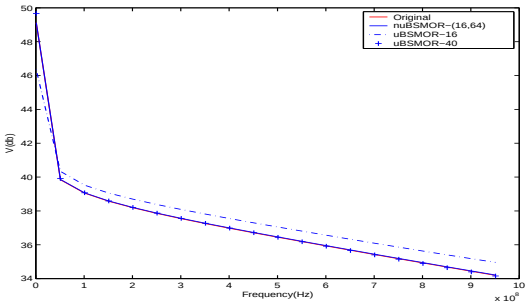

Figure 7: Frequency responses of the BSMOR by the non-uniform partition, uniform partition, and original model at one port of a nonuniform mesh (64x64-64x64-256x256-256x256).

5. EXPERIMENT

We implement the BSMOR and BBDC analysis on a Linux workstation (P4 2.66GHz, 1G RAM). The mesh structures of the substrate are generated from the typical mixed signal circuit application. In this section, we first investigate the accuracy of BSMOR and BBDC analysis, then study their scalabilities by increasing the circuit size and number of ports. As an example, we also present the noise map for a 256-contact array injected by a frequency-varying ring oscillator at dc and 10GHz.

5.1 Accuracy Comparison

We present the result of a reduced non-uniform mesh composed by 4 submeshes with different sizes (64x64-64x64-256x256-256x256). As shown in Fig. 7, after 10 iterations, the responses are visually identical for the original model, the reduced model from BSMOR by a non-uniform partition with two block size (16, 64) (resulting in 16 blocks), and the one by a uniform partition with the block size 16 (resulting in 40 blocks). But the one by a uniform partition with the block size 40 (resulting in 16 blocks) does not converge. It shows that the accurate reduced model needs to be generated from a projection matrix with the partition according to the structure of the original matrix, rather than a general 2×2 partition as SPRIM does, Moreover, the reduction time of BSMOR by the non-uniform partition is similar to the one by the uniform partition with the block size 40, and is 4X (4.17s vs. 20.38s) faster than using the uniform partition with the block size 16.

In Fig. 8, we compare frequency responses of the flat macro-model, partitioned macro-model without consideration of the correlation update from the coupling block (2.1-3.4 in Algorithm 1), and partitioned macro-model with consideration of the correlation update. Clearly, shown in Fig. 8 for a 256x256 RC-mesh (320K circuit elements) with 16 ports, the partitioned model with correlation update is as accurate as the flat macro-model, but the partitioned model without the correlation update has the non-negligible error at the high frequency region.

5.2 Scalability Study

We first study the efficiency of the reduction convergence by BSMOR and PRIMA. Different block numbers are used according to the different circuit size. We set an error bound as shown in Table 5.2, defined by the maximum error of the frequency response at one port up to 20GHz. We then perform reductions of BSMOR and PRIMA by increasing their iterations until that their accuracies meet the bound. As shown in Table 5.2, BSMOR uses less iterations (≤ 8) to meet the error bound than PRIMA does. As a result, the reduction time of BSMOR is also smaller than

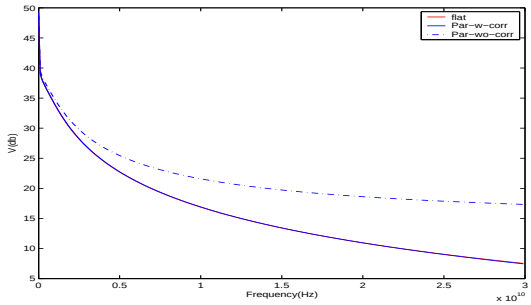

Figure 8: Frequency responses of the flat and partitioned models of a uniform mesh (256x256).

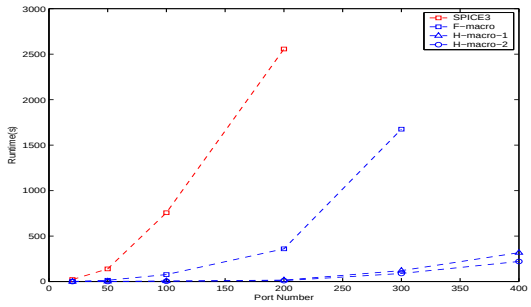

Figure 9: The scalability trend of simulation time for the original model, flat macro-model, partitioned models with different hierarchical levels.

that of PRIMA. For example, for a largest mesh circuit with 1M elements, BSMOR achieves 20X (240.22s vs. 4982.76s) speedup under the error bound 1e-4. Note that a relative small block number (64) is chosen for the largest circuit (1M) here. This is due to the fact that BSMOR needs additional steps to construct the projection matrix, and it results in a little bit larger state matrix that introduce the cost of matrix-vector multiplication. Hence the increase of the speedup is slowed if we choose large block number. In general, the result shows that with more partitions to construct a project matrix, BSMOR can match more poles than PRIMA does and hence the reduction time can be significantly reduced under the same accuracy.

We further study the simulation time scalability of the partitioned macro-model by BBDC in Table 5.2. PRIMA is used to generate the flat macro-model, BSMOR is used to generate the partitioned macro-model with hierarchy, and different block numbers are used to generate the macro-model according to the port number. Each reduced block contains 10 ports. The original, flat and partitioned models are all simulated in frequency domain up to 20GHz. The maximum error of the frequency response (relative to the original model) up to 20GHz at a selected port is used for comparison. We observe that when the port number is less than 50 ports the simulation time of the partitioned macro-model is up to 30X times faster than the flat macro-model with a similar accuracy. This speedup comes from two aspects: i) the cost of the eigen-decomposition to construct flat macro-model is reduced by BSMOR as the sparsity of reduced state matrices is reserved; On the other hand, PRIMA produces a dense reduced state matrices that are computation expensive during the eigen-decomposition; ii) the partitioned solution further reduces the simulation time as no expensive computation is involved for the large system matrix.

To achieve a similar efficiency for the circuits with the large number of ports ($\geq$ 100), we further use the hierarchical clustering (degree 10) with the sparsification (5% error bound) to control the size and sparsity of the coupling blocks. For 1-level and 2-level hierarchical solution, we sparsify the admittance matrices at bottom level, and second level, respectively. Since the

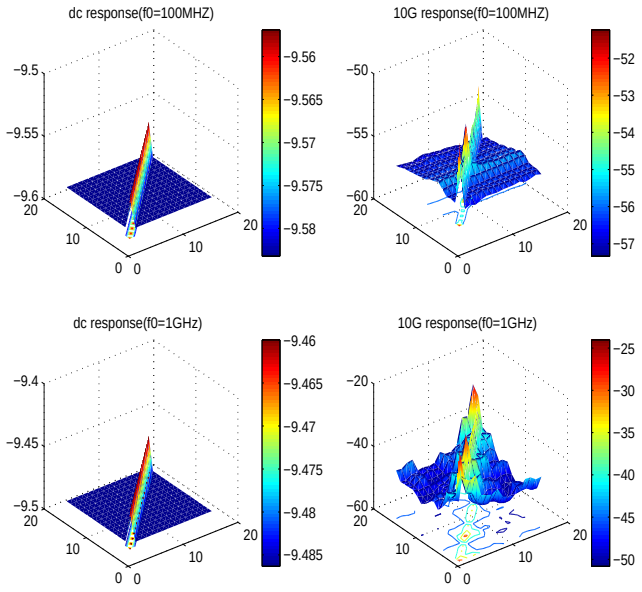

Figure 10: A noise map with 16x16 contacts array injected by frequency-controllable ring oscillators at f_0=100MHz and f_0=1GHz.

error at local matrix can propagate up, we find the solution by sparsification at 1-level partition is less accurate than that at 2-level partition. Moreover, we find that the flat macro-model can not be completed for a 400-port circuit. A clear scalability trend is shown in Fig. 9. We find that the simulation time of the flat macro-model grows up quickly. It shows the similar trend as the original model. This is due to the fact that the dense matrix structure degrades the overall performance when compared to the original larger but sparser matrix. In contrast, with the use of the BBDC analysis, the simulation time grows much slower than the flat macro-model.

5.3 Map of Substrate Noise Spectrum

We then apply the partitioned macro-model to generate a map of substrate noise spectrum. The injection current of a frequency-varying ring oscillations is characterized at $f_0 = 100MHz, 1GHz$. The maximum currents are characterized in time domain and then FFT (2048 samplings) is used to obtain the current envelope in frequency domain. The substrate considered here is a $3mm \times 3mm$ plane with a $200um$ thick p-type substrate ($\sigma = 0.1[\Omega cm]^{-1}$). We assume that the contacts are in a 16×16 array, and all the noise-current injection sources (ring oscillators) are placed diagonally in the array. The original substrate circuit is a 256x256 RC-mesh with 320K elements, and we apply 32×32 BSMOR to obtain a 256-port macro-model, representing a 16×16 contact array. The reduction time is about 120s. A 2-level hierarchical partition is used to generate a port-matrix response within 90s. Fig. 10 shows the map of the noise envelope (voltage bounce magnitude) at dc and $10GHz$. Clearly, reducing the central clock frequency from 1GHz to 100MHz can reduce 25db peak noise at the high frequency (10GHz), but the noise envelope at dc is not reduced. Moreover, the substrate noise coupling is localized at dc but it can diffuse across the contact array at 10GHz. As we assume a high conductivity substrate, the use of the guard ring is effective for this type of substrate. A p^+-guard ring is used for the isolation with the conductivity $\sigma = 100.0[\Omega cm]^{-1}$. We model the effect of this isolation by changing the surrounding resistance of the contact for each ring oscillator. As shown in Fig. 11, by using a guard ring at 10GHz for $f_0 = 1GHz$, the substrate noise is confined around the injection sources at the diagonal of the contact array.

Ckt	elements	err-bound	BSMOR			PRIMA	
			block#	iter#	time	iter#	time
mesh1	1K	1e-8	2x2	4	0.03s	10	0.09s
mesh2	10K	1e-8	8x8	6	0.07s	20	0.28s
mesh3	80K	1e-6	16x16	6	0.42s	30	3.82s
mesh4	160K	1e-6	16x16	6	5.14s	40	46.98s
mesh5	320K	1e-4	32x32	6	10.27s	60	104.62s
mesh6	1M	1e-4	64x64	8	240.22s	80	4982.76s

Table 1: Comparison of the reduction time of BSMOR and PRIMA under the same accuracy up to 20GHz.

Ckt	Port#	SPICE3	flat macro-model		H-partitioned-macro-model			
					1-level		2-level	
		time	time	error	time	error	time	error
mesh2	20p	22.12s	1.23s	1e-6	0.04s	1e-6	0.04s	1e-6
mesh3	50p	139.80s	14.83s	3e-6	1.53s	4e-6	0.72s	3e-6
mesh4	100p	757.12s	76.98s	3e-6	5.13s	1e-3	3.92s	5e-4
mesh5	200p	2556.54s	360.39s	3e-6	16.09s	2e-2	10.27s	3e-3
mesh6	300p	NA	1674.98s	NA	120.03s	NA	89.23s	NA
mesh6	400p	NA	NA	NA	317.34s	NA	220.87s	NA

Table 2: Simulation efficiency comparison by the original model, flat macro-model (PRIMA), partitioned macro-model with hierarchy (BSMOR).

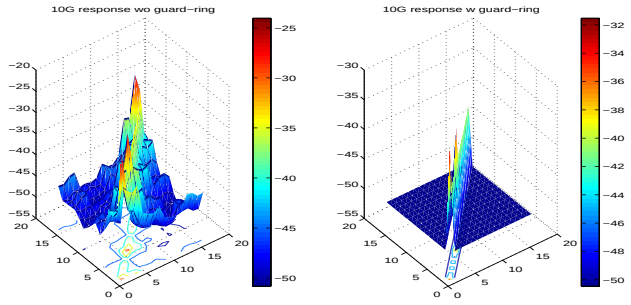

Figure 11: A noise map at high frequency 10GHz (f_0=1GHz) with/without guard rings.

6. CONCLUSION

In this paper, we have proposed a block structure preserving model reduction (BSMOR), which generalizes the structure preserving model order reduction (SPRIM) [6]. We found that increasing block number leads to more matched poles or moments than PRIMA using the same iteration. It in turn improves the model reduction efficiency compared to PRIMA under the same error bound. For a circuit with 1M elements, BSMOR has a 20X smaller reduction time than PRIMA does. As BSMOR preserves the structure of state matrices, it generates *sparse* reduced state matrices. For a circuit with 320K elements, the reduced state matrices ($\mathcal{G}$, $\mathcal{C}$) has 72% and 93% sparsification ratio after a 16×16 BSMOR reduction. It leads to an efficient construction of a MIMO macro-model when using eigen-decomposition. To be able to handle the resulting macro-model with large number of ports, we further used bordered-block diagonal partition with hierarchical clustering (BBDC) to decompose the macro-model into blocks with the manageable size. The experiment shows that BBDC reduces 30X simulation time than the original macro-model. In the future, we plan to study how to find the optimum block number for BSMOR to generate a order-reduced state matrix that is sparse yet small.

7. REFERENCES

[1] P. Feldmann and R. W. Freund, "Efficient linear circuit analysis by pade approximation via the lanczos process," *IEEE Trans. on CAD*, vol. 14, pp. 639–649, May 1995.

[2] 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*, vol. 16, pp. 734–744, July 1998.

[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] J.R.Phillips and L.M.Silveira, "Poor man's TBR: A simple model reduction scheme," *IEEE Trans. on CAD*, pp. 43–55, 2005.

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

[6] R. W. Freund, "Sprim: Structure-preserving reduced-order interconnect macro-modeling," in *IEEE/ACM ICCAD*, 2004.

[7] M. Zhao, R. Panda, and et.al., "Hierarchical analysis of power distribution networks," *IEEE Trans. on CAD*, pp. 159–168, 2002.

[8] S. X.-D. Tan, "A general s-domain hierarchical network reduction algorithm," in *IEEE/ACM ICCAD*, 2003.

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

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

[11] R.A.Rohrer, "Circuit partitioning simplified," *IEEE Trans. on CAS*, pp. 2–5, 1988.

[12] B. Stanisic, N.K.Verghese, R.A.Rutenbar, L.R.Carley, and D.J.Allstot, "Addressing substrate coupling in mixed-mode ICs: simulation and power distribution synthesis," *IEEE J. Solid-State Circuits*, pp. 226 – 238, 1994.

[13] A.M.Niknejad, R.Gharpurey, and R.G.Meyer, "Numerically stable green function for modeling and analysis of substrate coupling in integrated circuits," *IEEE Trans. on CAD*, pp. 305 – 315, 1998.

[14] E.Charbon, P.Miliozzi, L.Carloni, A.Ferrari, and A.Sangiovanni-Vincentelli, "Modeling digital substrate noise injection in mixed-signal IC's," *IEEE Trans. on CAD*, pp. 301 – 310, 1999.

[15] A. Nardi, H.B.Zeng, J.L.Garrett, L.Daniel, and A.Sangiovanni-Vincentelli, "A methodology for the computation of an upper bound on noise current spectrum of CMOS switching activity," in *IEEE/ACM ICCAD*, 2002.