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

Hao Yu, Lei He, and Sheldon X.-D. Tan

I. 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 extensively. Based on the Krylov subspace projection and congruence transformation, PRIMA [3] is widely used to generate the reduced macro-model with preserved passivity. However, the macro-model produced by PRIMA is not compact as the order is usually "too high" to achieve the specified accuracy. Furthermore, the macro-model is represented by a multiple-input-multiple-output (MIMO) transfer function, and is usually dense and inefficient to analyze for a large number of ports.

To improve upon PRIMA, a structure-preserving model reduction (SPRIM) is proposed in [4]. It 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 (G, C, L, E_s) 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, which facilitates the realization of the reduced model. However, such a simple 2×2 partition does not leverage the regularity of the substrate network. In addition, the explicit hierarchical decomposition [5], [6] is proposed to handle a large number of ports. The capacity of these methods [4]–[6] need to be improved further.

In this paper, we propose a block structure preserving model reduction (BSMOR) method, which generalizes SPRIM [4] in the sense that the G, C, L and E_s matrices are further partitioned into blocks. 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. Compared to PRIMA, BSMOR can lead to more efficient reduction under the same accuracy. In addition, BSMOR can also preserve the sparsity for reduced block matrices, which gives further efficiency boost to constructing a macro-model. The resulting macro-model consists of order-reduced blocks, each containing a subset of ports. To analyze a macro-model with a large number of ports, we further propose a bordered-block diagonal (BBD) partitioning and hierarchical clustering of reduced blocks. We call it BBDC analysis. The experiment shows that under the same accuracy, the reduction of our approach is 20X times faster than PRIMA to reduce a circuit with 1M elements, and 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 method with two differnt partioing algorithms in Sections II and III, a hierarchical bordered-block-diagonal analysis for the reduced model in Section IV, the prototype experimental results in Section,V and conclude the paper in Section VI.

Hao Yu and Lei He are with Department of Electrical Engineering, University of California, Los Angeles, CA 90095 USA (e-mail: {hy255,lhe}@ee.ucla.edu). Sheldon X.-D. Tan is with Department of Electrical Engineering, University of California, Riverside, Riverside, CA 92521 USA (e-mail:stan@ee.ucr.edu)

II. Block Structure Preserving Model Reduction

A. Preliminary

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

$$\begin{aligned} \mathcal{G}x(s) + s\mathcal{C}x(s) &= \mathcal{B}u_e(s) \\ y_e(s) &= \mathcal{B}^T x(s) \end{aligned} \tag{1}$$

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

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

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

$$\begin{aligned} y_e(s) &= H(s)u_e(s) \\ H(s) &= \mathcal{B}^T(\mathcal{G} + s\mathcal{C})^{-1}\mathcal{B}, \end{aligned} \tag{3}$$

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

$$\mathcal{K}(\mathcal{A}, \mathcal{R}, q) = span(V) = \{\mathcal{R}, \mathcal{A}\mathcal{R}, ..., \mathcal{A}^{n-1}\mathcal{R}\}, \tag{4}$$

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

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

where

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

Theorem 1: If $\mathcal{K}(\mathcal{A}, \mathcal{R}, q) \subseteq span(V)$, then $\hat{H}(s)$ has the identical expanded first q-th moments with $H(s)$.

It is called the *Grimme's projection theorem* [7]. Note that $\hat{\mathcal{G}}$, $\hat{\mathcal{C}} \in R^{q\times q}$, and $\hat{\mathcal{B}} \in R^{q\times n_p}$.

In [4], 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.

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

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

$$\mathcal{G} = \begin{bmatrix} G & E_s^T \\ -E_s & 0 \end{bmatrix}, \mathcal{C} = \begin{bmatrix} C & 0 \\ 0 & L \end{bmatrix}, \tag{8}$$

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 E_s ($\in R^{n_2\times n_1}$) is the adjacent matrix indicating the branch current flow at the inductor. Note that $n_1 + n_2 = N$.

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

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

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

$$\widetilde{\mathcal{G}} = \begin{bmatrix} \widetilde{G} & \widetilde{E}_s^T \\ -\widetilde{E}_s & 0 \end{bmatrix}, \widetilde{\mathcal{C}} = \begin{bmatrix} \widetilde{C} & 0 \\ 0 & \widetilde{L} \end{bmatrix}, \tag{10}$$

where $\widetilde{G} = {V_1}^T G V_1$, $\widetilde{E}_s = {V_2}^T E_s V_1$ and $\widetilde{C} = {V_1}^T C V_1$ and $\widetilde{L} = {V_2}^T L V_2$. Similarly, the size of $\widetilde{\mathcal{G}}$, $\widetilde{\mathcal{C}}$ ($\in R^{2q\times 2q}$), and $\widetilde{\mathcal{B}}$ ($\in R^{2q\times n_p}$) 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.

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 the Schur's decomposition [8]: $\widetilde{\mathcal{H}}_{NA} = \widetilde{G} + s\widetilde{C} + \frac{1}{s}\widetilde{E}_s^T \widetilde{L}^{-1} \widetilde{E}_s$ where $\widetilde{\mathcal{H}}_{NA}$ 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 $\widetilde{\mathcal{H}}_{NA}$ becomes 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, but both 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.

B. Structured Block Projection

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 poles 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

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

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

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

where each 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 or the power/ground grid, called *natural basic blocks*, or determined after specific partitioning algorithms discussed later on.

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

$$\begin{aligned} V &= \begin{bmatrix} V_{1(n_1\times q)} \\ V_{2(n_2\times q)} \\ \vdots \\ V_{m(n_m\times q)} \end{bmatrix} \\ \rightarrow \widetilde{V} &= \begin{bmatrix} V_{1(n_1\times q)} & 0 & \cdots & 0 \\ 0 & V_{2(n_2\times q)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & V_{m(n_m\times q)} \end{bmatrix}. \end{aligned} \tag{13}$$

where $\widetilde{V} \in R^{N\times mq}$. We call this as *Block Structure-preserving Model Reduction* (BSMOR), where m is the number of blocks.

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

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

Elementwise, we have

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

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 $V_i = V_{i(q\times n_i)}$ to simplify notations. Note that such a $m \times m$ block projection preserves the structure and sparsity of the original $\mathcal{G}$, $\mathcal{C}$ matrices. For example, when projected by $\widetilde{V}$, the reduced $\widetilde{\mathcal{G}}$ matrix is

$$\widetilde{\mathcal{G}} = \begin{bmatrix} V_1{}^T\mathcal{G}_{1,1}V_1 & V_1{}^T\mathcal{G}_{1,2}V_2 & \dots & V_1{}^T\mathcal{G}_{1,m}V_m \\ V_2{}^T\mathcal{G}_{2,1}V_1 & V_2{}^T\mathcal{G}_{2,2}V_2 & \dots & V_2{}^T\mathcal{G}_{2,m}V_m \\ \vdots & \vdots & \ddots & \vdots \\ V_m{}^T\mathcal{G}_{m,1}V_1 & V_m{}^T\mathcal{G}_{m,2}V_2 & \dots & V_m{}^T\mathcal{G}_{m,m}V_m \end{bmatrix}. \quad (16)$$

Clearly, if the original $\mathcal{G}$ is sparse, the resulted $\widetilde{\mathcal{G}}$ is sparse as well. In the contrast, when projected by V using PRIMA, the resulted $\hat{\mathcal{G}}$ is

$$\hat{\mathcal{G}} = \sum_{i=1}^{m}\sum_{j=1}^{m} V_i{}^T\mathcal{G}_{i,j}V_j, \quad (17)$$

which is dense in general. Similar observations can be found for $\mathcal{C}$ and $\mathcal{B}$.

C. Passivity Preservation and Moment Matching

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.

Proof: Because $\widetilde{V}^T\widetilde{V} = I$, and $\mathcal{G} + \mathcal{G}^{\mathcal{T}}$, $\mathcal{C} + \mathcal{C}^{\mathcal{T}}$ are symmetric positive definite, the congruence transformation based projections: $\widetilde{V}^T\mathcal{G}\widetilde{V}$, $\widetilde{V}^T\mathcal{C}\widetilde{V}$ preserve the passivity [3].

Using such a projection matrix $\widetilde{V}$, we define a reduced-order model with the following transfer function

$$\widetilde{H}(s) = \widetilde{\mathcal{B}}^T(\widetilde{\mathcal{G}} + s\widetilde{\mathcal{C}})^{-1}\widetilde{\mathcal{B}}, \quad (18)$$

and we have

Theorem 3: Let V be a matrix that satisfies $\mathcal{K}(\mathcal{A}, \mathcal{R}, q) \subseteq span(V)$ and $\widetilde{V}$ is defined in Eq.(13). $\widetilde{H}(s)$ will match the first q moments in the expansion of $H(s)$ about s_0. If the $\mathcal{G}$, $\mathcal{C}$ matrices in the block diagonal form have m different blocks, i.e. $\mathcal{A}_{1,1} \neq \mathcal{A}_{2,2} \neq ... \neq \mathcal{A}_{m,m}$, $\widetilde{H}(s)$ will match mq poles of $H(s)$.

Proof: Because $\mathcal{K}(\mathcal{A}, \mathcal{R}, q) \subseteq span(V) \subseteq span(\widetilde{V})$, according to *Grimme's' projection theorem* (Theorem 1), $\widetilde{H}(s)$ will match the first q moments in the expansion of $H(s)$ about s_0. Moreover, if $\mathcal{G}$, $\mathcal{C}$ matrices are in the block diagonal form, then the state matrix $\mathcal{A} = \mathcal{G}^{-1}\mathcal{C}$ is in the block-diagonal form as well. Due to the structure-preserving projection, the resulted state matrix $\widetilde{\mathcal{A}}$

$$\begin{aligned} \widetilde{\mathcal{A}} &= (\widetilde{V}^T\mathcal{G}\widetilde{V})^{-1}(\widetilde{V}^T\mathcal{C}\widetilde{V}) = diag[\mathcal{A}_{1,1}, \mathcal{A}_{2,2}, ..., \mathcal{A}_{m,m}] \\ &= \begin{bmatrix} (\widetilde{\mathcal{G}}_{1,1})^{-1}\widetilde{\mathcal{C}}_{1,1} & 0 & \dots & 0 \\ 0 & (\widetilde{\mathcal{G}}_{2,2})^{-1}\widetilde{\mathcal{C}}_{2,2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & (\widetilde{\mathcal{G}}_{m,m})^{-1}\widetilde{\mathcal{C}}_{m,m} \end{bmatrix} \end{aligned}$$

have m sub-blocks each with size $q \times q$. Note that eigenvalues of $\widetilde{\mathcal{A}}$ represent the reciprocal poles of the reduced model [3]. As a result, the partitioned projection matrix $\widetilde{V}$ leads to localized projection. In other words, the block projection matrix $\widetilde{V}_i$ is used only for the state matrix block $\mathcal{A}_{i,i}$ ($i = 1, ...m$). Each structured block projection matrix $\widetilde{V}_i$ will lead to the localized model order reduction for block i, which is represented by $(\mathcal{A}_{i,i}, \mathcal{R}_{i,i})$. Conceivably, the order reduced block $(\widetilde{\mathcal{A}}_{i,i}, \widetilde{\mathcal{R}}_{i,i})$ will match $(\mathcal{A}_{i,i}, \mathcal{R}_{i,i})$ with q poles. But the whole system consisting of the m blocks will match mq poles.

In general when $\mathcal{G}_{i,j}, \mathcal{C}_{i,j} \neq 0$, i.e., $\mathcal{A}_{i,j} \neq 0$, BSMOR obtains $(m-1)q$ times more poles than PRIMA within q iterations, where the $(m-1)q$ additional poles approximately match poles of the original system. If the coupling $\mathcal{A}_{i,j}$ between blocks $\mathcal{A}_{i,i}$ is weak, the original system can be well approximated by the block diagonal structure, and therefore, each block has a different pole distribution, and the $(m-1)q$ additional poles are close to the poles of the original system. As a result, introducing more partitions or blocks can archive the same reduction accuracy by using less iterations, which can in turn improve the reduction efficiency. On the other hand, if the coupling $\mathcal{A}_{i,j}$ between blocks $\mathcal{A}_{i,i}$ is strong, the accuracy of BSMOR depends on the partition algorithm. Note that in this case, blocks are strongly correlated. Below, we present a partitioning algorithm based on the the singular-value-decomposition and K-means clustering.

D. SVD and K-means Clustering Based Partitioning

When blocks are strongly correlated, the state matrix $(\mathcal{A})$ will be a low rank matrix. The true rank should reveal the number of uncorrelated rows in $\mathcal{A}$. Therefore, if we can find this true rank number m, and greedy group all similar row vectors of $\mathcal{A}$ into m most uncorrelated blocks, then we can obtain a partitioning that each partitioned block has weak correlation with each other, and shows a different pole distribution. Using such a partitioned V to construct the structure projection matrix $\mathcal{V}$ accordingly, and project $\mathcal{G}$, $\mathcal{C}$ matrices, the obtained mq poles can closely match poles of the original system.

Based on this observation, SVD is applied on $\mathcal{A}$ to determine the rank value m, which is used as the partition number. With the known partition number, K-means clustering method [9] can be applied to greedily add similar rows into one block, and find m disjoint blocks. As a result, the moment value, i.e., the pole distribution is similar in the same block, and is different between them. However, the dimention of $\mathcal{A}$ ($\in R^{N\times N}$) is too high to directly apply SVD and K-menas clustering. Observing that the projection matrix V composed by $(\mathcal{A}, \mathcal{R})$ as in (4) has a low-dimension $R^{q\times N}$, we apply SVD and K-means clustering to V. The overall procedure is presented in Algorithm 1.

K-means clustering classifies a given data set. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. As a result, the better choice is to place them as far away from each other as possible. The next step is to iteratively associate data points to the nearest centroid. When no point is left, the first step is completed and an initial clustering is obtained. At this point we need to recalculate k new centroids as bary-centers of the clusters resulting from the previous step. After we have these k new centroids, a new iteration binding has to be done. As a result the k centroids may change their locations and iterations stop when there is no more changes for the k centroids. This procedure minimizes the following squared error function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2, \tag{19}$$

where $||x_i^{(j)} - c_j||^2$ is the distance between data point $x_i^{(j)}$ and cluster centroid c_j.

Algorithm 1 BSMOR Partitioning 1

1.Find projection matrix by PRIMA
(1.1) *input*: $\mathcal{G}, \mathcal{C}, \mathcal{B}$;
(1.2) *form*: $\mathcal{A} = (\mathcal{G} + s_0\mathcal{C})^{-1}\mathcal{C}$, $\mathcal{R} = (\mathcal{G} + s_0\mathcal{C})^{-1}\mathcal{B}$;
(1.3) *construct*: $V = [V_1, V2, ..., V_q]$ such that $span(V) = \mathcal{K}(\mathcal{A}, \mathcal{R}, q)$, where $V \in R^{N \times mq}$;
(1.4) *output*: V.
2.SVD and K-means clustering based partitioning
(2.1) *input*: V;
(2.2) *solve*: $m = SVD(V)$;
(2.3) *cluster*: $[V_{1(n_1 \times q)}, \ldots, V_{m(n_m \times q)}]^T = kmeans(V, m)$;
(2.4) *construct*: $\mathcal{V} = diag[V_{1(n_1 \times q)}, \ldots, V_{m(n_m \times q)}]$;
(2.5) *output*: $\mathcal{V}$.
3.Structured block projection
(3.1) *input*: $\mathcal{V}$;
(3.2) *project*: $\widetilde{\mathcal{G}} = \widetilde{V}^T\mathcal{G}\widetilde{V}, \quad \widetilde{\mathcal{C}} = \widetilde{V}^T\mathcal{C}\widetilde{V}, \quad \widetilde{\mathcal{B}} = \widetilde{V}^T\mathcal{B}$;
(3.3) *output*: $\widetilde{\mathcal{G}}, \widetilde{\mathcal{C}}, \widetilde{\mathcal{B}}$.

In summary, using the partitioned V by the SVD and K-means clustering, one can construct a structured-projection matrix $\mathcal{V}$, where each block has a different pole distribution, and the reduced model can closely match mq poles of the original system. As a result, one can obtain order reduced models with higher accuracy for each structured block by using the same sized 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 a better trade-off between efficiency and model accuracy for reducing linear dynamic system models than PRIMA does.

III. Explicit Partition and Structured Block Projection

To construct the structured-projection matrix $\mathcal{V}$, we in Section II first obtain V by applying PRIMA to the entire $\mathcal{G}$, $\mathcal{C}$ matrices, and then partition V using SVD and K-means algorithm. As all operations are on the entire matrix, its computation effort could be too big for large scale circuits (although fewer iterations are needed to achieve higher accurate reduced model compared to PRIMA). To reduce the computation cost, we propose to first explicitly partition the original circuit into m interconnected blocks with $(\mathcal{G}_{i,i}, \mathcal{C}_{i,i})$, then apply PRIMA to each block and obtain the projection matrix $V_i (span(V_i) = \mathcal{K}(\mathcal{A}_{i,i}, \mathcal{R}_{i,i}, q))$ for each block, and construct the structured-projection matrix

$$\mathbf{V} = diag[V_1, V_2, ..., V_m]. \tag{20}$$

Below, we prove that the structured-projection matrix $\mathbf{V}$ constructed in this fashion guarantees q moments matching as well.

A. General Interconnected Block Structure

For a interconnected block structure, we assume that the input source vector for ith-block is

$$u_{ci}(s) = u_{ei}(s) + \sum_{j \in 1,\ldots,m}^{j \neq i} X_{ij} y_{cj}(s), \tag{21}$$

where $u_{ei}(s)$ is the external sources at ith block, $X_{ij}(s) = X_{g_{ij}} + sX_{cij}$ ($X_{ij} \in R^{n_i \times n_j}$) is the branch admittance matrix that connects ith and jth block, and $y_{cj}(s)$ is the output vector at jth block. Moreover, the transfer function of ith block is

$$\begin{aligned} y_{ci}(s) &= h_i(s) u_{ci}(s) \\ h_i(s) &= b_i^T (\mathcal{G}_{i,i} + s\mathcal{C}_{i,i})^{-1} b_i. \end{aligned} \tag{22}$$

where b_i is an extended port matrix ($\in R^{n_i \times n_i}$) to connect boundary ports at ith block. Note that

$$\mathcal{G}_{i,j} = b_i^T X_{g_{i,j}} b_j, \quad \mathcal{C}_{i,j} = b_i^T X_{ci,j} b_j. \tag{23}$$

Define the block connection matrix

$$\mathbf{x} = \begin{bmatrix} 0 & X_{1,2} & \ldots & X_{1,m} \\ X_{2,1} & 0 & \ldots & X_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{m,1} & X_{m,2} & \ldots & 0 \end{bmatrix}; \tag{24}$$

the block transfer function

$$\mathbf{H}_0(s) = diag\{h_1(s), ..., h_m(s)\} = \mathbf{B}_0^T (\mathbf{G}_0 + s\mathbf{C}_0)^{-1} \mathbf{B}_0, \tag{25}$$

where

$$\begin{aligned} \mathbf{B}_0 &= diag\{b_1, ..., b_m\}, \\ \mathbf{G}_0 &= diag\{\mathcal{G}_{1,1}, ..., \mathcal{G}_{m,m}\}, \\ \mathbf{C}_0 &= diag\{\mathcal{C}_{1,1}, ..., \mathcal{C}_{m,m}\}; \end{aligned} \tag{26}$$

the block connection input/output vector

$$u_c(s) = [u_{c1}(s), ..., u_{cm}(s)], \quad y_c(s) = [y_{c1}(s), ..., y_{cm}(s)]; \tag{27}$$

and the block external input/output vector

$$\mathbf{F} = [F_1, ..., F_m]. \tag{28}$$

The system equation, therefore, can be rewritten as

$$\begin{aligned} u_c(s) &= \mathbf{F} u_e(s) + \mathbf{X}(s) y_c(s), \\ y_c(s) &= \mathbf{H}_0(s) u_c(s), \\ y_e(s) &= \mathbf{F}^T y_c(s). \end{aligned} \tag{29}$$

Its transfer function is

$$\mathbf{H}(s) = \mathbf{F}^T (I - \mathbf{H}_0(s)\mathbf{X}(s))^{-1} \mathbf{H}_0(s) \mathbf{F}, \tag{30}$$

which can be realized by the following state space

$$\begin{aligned}\mathbf{G} &= \mathbf{G}_0 + \mathbf{B}_0^T\mathbf{X_g}\mathbf{B}_0,\\ \mathbf{C} &= \mathbf{C}_0 + \mathbf{B}_0^T\mathbf{X_c}\mathbf{B}_0,\\ \mathbf{B} &= \mathbf{B}_0^T\mathbf{F}. \end{aligned} \tag{31}$$

It is easy to check that: $\mathbf{G} = \mathcal{G}, \mathbf{C} = \mathcal{C}$, and $\mathbf{B} = \mathcal{B}$, i.e., the interconnected representation $(\mathbf{G}, \mathbf{C}, \mathbf{B})$ has the same state space with the original system $(\mathcal{G}, \mathcal{C}, \mathcal{B})$. Note that the rationale to use the presentation in the interconnected block structure is to obtain following relation for the Krylov subspace between the diagonal block structure and the original system.

Theorem 4: Define the state matrices for the block diagonal structure: $\mathbf{A}_0 = (\mathbf{G}_0 + s_0\mathbf{C}_0)^{-1}\mathbf{C}_0$, $\mathbf{R}_0 = (\mathbf{G}_0 + s_0\mathbf{C}_0)^{-1}\mathbf{B}_0$, and for the general interconnected structure: $\mathbf{A} = (\mathbf{G} + s_0\mathbf{C})^{-1}\mathbf{C}$, $\mathbf{R} = (\mathbf{G} + s_0\mathbf{C})^{-1}\mathbf{B}$. Then the corresponding Krylov subspaces satisfy:

$$\mathcal{K}(\mathbf{A}, \mathbf{R}, q) \subseteq \mathcal{K}(\mathbf{A}_0, \mathbf{R}_0, q). \tag{32}$$

Proof: It is equivalent to prove that for any $v \in \mathcal{K}(\mathbf{A}, \mathbf{R}, q)$, it has $v \in \mathcal{K}(\mathbf{A}_0, \mathbf{R}_0, q)$ as well. Note that if $v \in \mathcal{K}(\mathbf{A}, \mathbf{R}, q)$, it means that there exists a q-dimensional column vector w_i:

$$\begin{aligned} vw_0 &= \mathbf{R}, \quad vw_1 = -\mathbf{A}\mathbf{R}, ...,\\ vw_i &= (-1)^i\mathbf{A}^i\mathbf{R}, \quad vw_{q-1} = (-1)^{q-1}\mathbf{A}^{q-1}\mathbf{R}. \end{aligned}$$

Sum the left-hand-side and right-hand-side respectively,

$$\begin{aligned} vw &= (I + s\mathbf{A})^{-1}\mathbf{R} = (\mathbf{G} + s\mathbf{C})^{-1}\mathbf{B}\\ &= (\mathbf{G}_0 + s\mathbf{C} + 0 + \mathbf{B}_0^T\mathbf{X}\mathbf{B}_0)^{-1}\mathbf{B}_0\mathbf{F} \end{aligned} \tag{33}$$

where $w = \sum_{i=0}^{q-1}(-1)^i w_i$.

Therefore,

$$vw' = (\mathbf{G}_0 + s\mathbf{C}_0)^{-1}\mathbf{B}_0^T, \tag{34}$$

where $w' = w(\mathbf{F} - \mathbf{X}\mathbf{B}_0 vw)^{-1}$. As a result,

$$\begin{aligned} vw'_0 &= \mathbf{R}, \quad vw'_1 = -\mathbf{A}_0\mathbf{R}_0, ...,\\ vw'_i &= (-1)^i\mathbf{A}_0^i\mathbf{R}_0, \quad vw'_{q-1} = (-1)^{q-1}\mathbf{A}_0^{q-1}\mathbf{R}_0 \end{aligned}$$

i.e., $v \in \mathcal{K}(\mathbf{A}_0, \mathbf{R}_0, q)$.

Based on Theorem 3, we have the following Corollary:

Corollary 1: Let V_i satisfy $\mathcal{K}(\mathbf{A}_{i,i}, \mathbf{R}_i, q) \subseteq span(V_i)$ and $\mathbf{V} = diag[V_1, V_2, ..., V_m]$. The reduced system $\widetilde{H}(s)$ projected by $\mathcal{V}$ will match the first q moments in the expansion of $\mathbf{H}(s)$ about s_0.

Proof: Because $\mathcal{K}(\mathbf{A}, \mathbf{R}, q) \subset \mathcal{K}(\mathbf{A}_0, \mathbf{R}_0, q) \subseteq span((\mathcal{V}))$, the q moments matching is easy to see by *Grimme's projection theorem*.

B. Incremental Explicit Partitioning

Similarly, if the original system has a block diagonal structure, i.e., $\mathbf{X} = 0$, the above structured-projection can match mq poles as well. In general, it needs to devise a proper initial partitioning such that the resulting model can closely approximate mq poles matching. Below, we present an *incrementally partitioning based on k-dominant-time-constant* algorithm to achieve the requirement. Note that the concept of *dominant time constant* is introduced in [10], where the system timing response is approximated by the first k dominant-time-constant, i.e., the first k most dominant eigen-values (poles) $\lambda_1, ..., \lambda_k$ of state matrix $\mathcal{A} = \mathcal{G}^{-1}\mathcal{C}$.

The outlined procedure is presented in Algorithm 2. The input is a group of m_0 natural basic blocks that are determined by the specific application such as the block current characterization for the substrate or P/G grid. Each basic block has state space $(\mathcal{G}_{i,i}, \mathcal{C}_{i,i}, b_i)$, and a projection matrix V_i is found by PRIMA such that $\mathcal{K}(\mathcal{A}_{i,i}, \mathcal{R}_{i,i}, q) = span(V_i)$. The clustering begins with the first basic block by merging its connected neighboring blocks. The state matrices of the resulted block are denoted by $(\mathcal{G}'_{i,i}, \mathcal{C}'_{i,i}, b'_i)$, and its corresponding projection matrix is V'_i. SVD is applied to V_i and V'_i to obtain their first k most dominant eigen-values (poles) $\lambda_1, ..., \lambda_k$ and $\lambda'_1, ..., \lambda'_k$. As a result, the merging operation will terminate if

$$||\frac{\sum_i^k \lambda_i/\lambda'_i}{k} - 1|| < \epsilon \tag{35}$$

where ϵ is a small value specified by the user. This termination criteria implies the condition that the k most dominant eigen-values (poles) will not change much when including more interconnected blocks together. The clustering further applies to the rest basic block till a converged partitioning V' is obtained. Consequently, a structured-projection matrix $\mathbf{V}$ is constructed to project the original system matrices $(\mathcal{G}, \mathcal{C}, \mathcal{B})$.

Algorithm 2 BSMOR Partitioning 2

1.Incremental partitioning
(1.1) *input*: $\mathcal{G}_{i,i}, \mathcal{C}_{i,i}, b_i, (i = 1, 2, ..., m_0)$;
I=1,i=0;
while $\overline{\lambda} < \epsilon$ **do**
(1.1) *form*: $\mathcal{A}_I^{i,i} = (\mathcal{G}_{I,I}^i + s_0\mathcal{C}_{I,I}^i)^{-1}\mathcal{C}_{I,I}^i$, $\mathcal{R}_I^{i,i} = (\mathcal{G}_{I,I}^i + s_0\mathcal{C}_{I,I}^i)^{-1}b_I^i$;
(1.2) *construct*: V_I^i such that $span(V_I^i) = \mathcal{K}(\mathcal{A}_I^{i,i}, \mathcal{R}_I^{i,i}, q)$;
(1.3) *calculate*: $[\lambda_1^i, ..., \lambda_k^i] = SVD(V^i, k)$;
(1.4) *cluster*: $(\mathcal{G}_{I,I}^{i+1}, \mathcal{C}_{I,I}^{i+1}, b_I^{i+1}) = merge[(\mathcal{G}_{I,I}^i, \mathcal{C}_{I,I}^i, b_I^i), (\mathcal{G}_{I+1,I+1}^i, \mathcal{C}_{I+1,I+1}^i, b_{I+1}^i)]$;
(1.5) *calculate*: $\overline{\lambda} = ||\frac{\sum_j^k \lambda_j^{i+1}/\lambda_j^i}{k} - 1||$;
end while
(1.6) *construct*: $\mathbf{V} = diag[V'_{1(n_1\times q)}, \ldots, V'_{m(n_m\times q)}]$;
(1.7) *output*: $\mathbf{V}$.
3.Structured block projection
(2.1) *input*: $\mathbf{V}$;
(2.2) *project*: $\widetilde{\mathcal{G}} = \mathbf{V}^T\mathcal{G}\mathbf{V}, \quad \widetilde{\mathcal{C}} = \mathbf{V}^T\mathcal{C}\mathbf{V}, \quad \widetilde{\mathcal{B}} = \mathbf{V}^T\mathcal{B}$;
(2.3) *output*: $\widetilde{\mathcal{G}}, \widetilde{\mathcal{C}}, \widetilde{\mathcal{B}}$.

Note that if

$$\mathcal{G} + s\mathcal{C} = \mathbf{Y}'(s) + \mathbf{X}'(s)\mathbf{B}_0, \tag{36}$$

where

$$\begin{aligned}\mathbf{Y}'(s) &= diag\{(\mathcal{G}_{1,1}+s\mathcal{C}_{1,1}),...,(\mathcal{G}_{m,m}+s\mathcal{C}_{m,m})\}\\ \mathbf{X}'(s) &= \mathbf{B}_0\mathbf{X}(s),\end{aligned} \tag{37}$$

we have

$$\begin{aligned}\widetilde{\mathcal{G}}+s\widetilde{\mathcal{C}} &= \mathcal{V}^T\mathbf{Y}'\mathcal{V}+\mathcal{V}^T\mathbf{X}'\mathcal{V}\mathcal{V}^T\mathbf{B}_0\mathcal{V}\\ &= \widetilde{\mathbf{Y}}'(s)+\widetilde{\mathbf{X}}'(s)\widetilde{\mathbf{B}}_0,\end{aligned} \tag{38}$$

where $\widetilde{\mathbf{Y}}'$ and $\widetilde{\mathbf{X}}'$ both $\in R^{mq\times mq}$.

IV. Two Level Partitioned Analysis of Reduced Model

Because the reduction preserves the structure, it preserves: (i) the sparsity of state matrices; In contrast, the projected matrices by PRIMA are fully dense. (ii) the reciprocity of the network which enables efficient realization of the network; In contrast, PRIMA does not preserve this property. (iii) the block structure; It means that the reduced block can be distinguished by a subset of ports specified during BSMOR. Due to the preserved block structure , a *bordered-block diagonal* (BBD) partitioned solution can be applied to solve each block individually in both frequency and time domain.

Note that the runtime and memory requirement to solve a linear system are primarily determined by the size, sparsity, and structure of the matrix. Using partitioning, the large coupled network is divided into sub-blocks with manageable size and solved in 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.

Consider the system equation for the reduced model

$$\widetilde{\mathbf{Y}}x=\widetilde{\mathbf{b}} \tag{39}$$

In frequency domain at a frequency point s

$$\begin{aligned}\widetilde{\mathbf{Y}} &= \widetilde{\mathcal{G}}+s\widetilde{\mathcal{C}}=\widetilde{\mathbf{Y}}'(s)+\widetilde{\mathbf{X}}'(s)\widetilde{\mathbf{B}}_0\\ \widetilde{\mathbf{b}} &= \widetilde{\mathcal{B}}u_e(s).\end{aligned} \tag{40}$$

Note that in time domain at a time instant t with time step h

$$\begin{aligned}\widetilde{\mathbf{Y}} &= \widetilde{\mathcal{G}}+\frac{1}{h}\widetilde{\mathcal{C}}=\widetilde{\mathbf{Y}}'(h)+\widetilde{\mathbf{X}}'(h)\widetilde{\mathbf{B}}_0\\ \widetilde{\mathbf{b}} &= \frac{1}{h}\widetilde{\mathcal{C}}x(t-h)+\widetilde{\mathcal{B}}u_e(t).\end{aligned} \tag{41}$$

The state vector can be solved by Sherman-Morrison-Woodbury formula [12]

$$\begin{aligned}x &= P^{(0)}-Pq\\ P^{(0)} &= (\widetilde{\mathbf{Y}}')^{-1}\widetilde{\mathbf{b}},\quad P=(\widetilde{\mathbf{Y}}')^{-1}\widetilde{\mathbf{X}}'\\ q &= (I+\widetilde{\mathbf{B}}_0P)^{-1}\widetilde{\mathbf{B}}_0P^{(0)}.\end{aligned} \tag{42}$$

To avoid explicit inversion, LU-factorization needs to applied $\widetilde{\mathbf{Y}}'$ and $I+\widetilde{\mathbf{B}}_0(\widetilde{\mathbf{Y}}')^{-1}\widetilde{\mathbf{X}}'$. A two level analysis is proposed in [13] and outlined in Algorithm 3. Note that as $\widetilde{\mathbf{Y}}$ shows the block diagonal form, each reduced block matrix is fist solved individually with LU/Cholesky factorization and substitution (1.1-1.4) at the bottom level, the results from each reduced block are then used further to solve the coupling block (2.1-2.4) at the top level, and the final x_k of each reduced block is updated (3.1).

Algorithm 3 Two Level Partitioned Analysis

1.Solve bottom level individually
for every $k\ in\ m$ **do**
(1.1) *input*: $\widetilde{\mathbf{b}}_k, \widetilde{\mathbf{X}}_k'^{(i)}, \widetilde{\mathbf{Y}}_k'$;
(1.2) *factor*: LU/Cholesky factor $\widetilde{\mathbf{Y}}_k'$;
$\widetilde{\mathbf{X}}_k'^{(0)}=\widetilde{\mathbf{b}}_k$;
for every $i\ in\ q$ **do**
(1.3) *solve*: back-substitution $\widetilde{\mathbf{Y}}_k'P_k^{(i)}=\widetilde{\mathbf{X}}_k'^{(i)}$;
end for
end for
(1.4) *output*: $P^{(0)}, P=[P^{(1)},...,P^{(n)}]$.
2.Solve top level
(2.1) *input*: $\widetilde{\mathbf{B}}_0, P, P^{(0)}$;
(2.2) *factor*: LU/Cholesky factor $I+\widetilde{\mathbf{B}}_0P$;
(2.3) *solve*: back-substitution $(I+\widetilde{\mathbf{B}}_0P)q=\widetilde{\mathbf{B}}_0P^{(0)}$;
(2.4) *output*: q.
3.Update bottom level individually
(3.1) *output*: $x=P^{(0)}-Pq$.

Typically, LU factorization requires $n^3/3$ multiplications and back/forward substitution requires $n^2/2$ multiplications. The computational cost of Algorithm 3 is therefore, ... Moreover, this algorithm can be extended in a hierarchical fashion if the multi-level partitioning is used initially.

References

[1] L. T. Pillage and R. A. Rohrer, "Asymptotic waveform evaluation for timing analysis," *IEEE Trans. on CAD*, pp. 352–366, April 1990.
[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] R. W. Freund, "Sprim: Structure-preserving reduced-order interconnect macro-modeling," in *IEEE/ACM ICCAD*, 2004.
[5] M. Zhao, R. Panda, and et.al., "Hierarchical analysis of power distribution networks," *IEEE Trans. on CAD*, pp. 159–168, 2002.
[6] 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.
[7] E.J.Grimme, *Krylov projection methods for model reduction (Ph. D Thesis)*. Univ. of Illinois at Urbana-Champaign, 1997.
[8] G. H. Golub and C. F. V. Loan, *Matrix Computations*. Baltimore, MD: The Johns Hopkins University Press, 3 ed., 1989.
[9] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Fifth Berkeley Symposium on Math*, 1967.
[10] L. Vandenberghe, S. Boyd, and A. E. Gamal, "Optimizing dominant time constant in rc circuits," *IEEE Trans. on CAD*, pp. 110–125, 1998.
[11] R.A.Rohrer, "Circuit partitioning simplified," *IEEE Trans. on CAS*, pp. 2–5, 1988.
[12] J. Sherman and W.J.Morrison, "Adjustment of the an inverse matrix corresponding to changes in the elements of original matrix," in *Annu. Math. Statist.*, p. 621, 1949.
[13] G.Guardabassi and A. Sangiovanni-Vincentelli, "A two level algorithm for tearing," *IEEE Trans. on CAS*, pp. 783–791, 1976.