Decentralized and Passive Model Order Reduction of Linear Networks With Massive Ports

Boyuan Yan, Sheldon X.-D. Tan, *Senior Member, IEEE*, Lingfei Zhou, Jie Chen, *Fellow, IEEE*, and Ruijing Shen, *Student Member, IEEE*

***Abstract*—It is well known that model order reduction for circuits with many terminals remains a challenging problem. One reason is that existing approaches are based on a *centralized* framework, in which each input-output pair is implicitly assumed to be equally interacted and the matrix-valued transfer function is assumed to be fully populated. In this paper, we attempt to address this long-standing problem using a *decentralized* model order reduction scheme, in which a multi-input multi-output system is decoupled into a number of subsystems and each subsystem corresponds to one output and several dominant inputs. The decoupling process is based on the relative gain array, which measures the degree of interaction of each input-output pair. For each decoupled subsystem, passive reduction can be easily achieved using existing reduction techniques. The proposed method is suitable for resistance-dominant interconnects such as on-chip power grids, substrate planes where extremely compact models can be obtained. Simulation results demonstrate the advantage of the proposed method compared to the existing approaches.**

***Index Terms*—Decentralized, model order reduction, multi-port networks.**

I. Introduction

MODEL ORDER REDUCTION (MOR) [1] is an efficient technique to reduce the interconnect circuit complexity while producing a good approximation of the input-output behavior. Existing approaches may generally be divided into two broad categories: the moment-matching-based methods and the balanced truncation-based methods. In the former case, the system is projected onto the Krylov subspace to match dominant moments while in the latter case the system is projected onto a subspace, which is both easily controllable and easily observable.

Moment-matching-based methods have been a great success owing to their efficiency [8], [10], [11], [18], [24], [27], [28], [37]. Recent methods perform implicit moment-matching by projecting the original system onto a Krylov subspace, and in this process, stability, passivity, and structure information inherent to resistance-inductance-capacitance (*RLC*) circuits can be preserved by exploiting the internal structure of the *RLC* formulation. While suitable for reduction of large-scale circuits, moment-matching-based techniques do not necessarily generate models as compact as desired. Nevertheless, another approach, the truncated balanced realization (TBR), which has been well developed in the field of systems and control [12], [17], can be employed to advantage [14], [22], [23], [29], [31]–[33], [36].

However, the efficiency of existing model order reduction methods degrade as the number of ports increases. The reason for this degradation is fundamental and does not depend on any particular reduction algorithm. For Krylov subspace-based algorithms, the cost associated with model computation is directly proportional to the number of inputs, i.e., to the number of columns in the transfer function matrix. Similarly, in the TBR algorithm, for systems with many inputs, many states may be needed because of the high dimension of the controllable subspace.

There has been notable effort devoted to mitigating this long-standing problem recently. One strategy is to exploit input signals during the reduction. The examples are the extended Krylov subspace (EKS) method [30], the generalized second-order Arnoldi method [25], and the extended truncated balanced realization method [13], [26], which consider the dynamics of the circuit as well as the source excitations during the reduction. However, since the modeling process depends on the input signal, the model needs to be rebuilt once input signal is changed. The second strategy performs the terminal reduction before the model reduction [7], [9], [15], [16]. The early work like SVDMOR method [7], [9] utilizes the fact that the matrix transfer function may be numerically low rank at dc, or at some specific frequency. However, the transfer function matrix of a large-scale circuit is rarely low rank in practice.

In this paper, we propose a *decentralized* model order reduction scheme where a multi-input-multi-output (MIMO) network is decoupled into a number of multi-input-single-output (MISO) networks (subsystems) based on the input-output interaction strength. Each subsystem is then reduced individually, which can be done easily using the existing techniques. The decoupling process is based on the relative gain array (RGA) [2], which is a matrix of interaction measures for single-input

Manuscript received August 11, 2010; revised December 11, 2010; accepted February 26, 2011. Date of publication April 05, 2011; date of current version April 06, 2012. A portion of this paper appeared in the Proceedings of the 45th IEEE/ACM Design Automation Conference, Anaheim, CA, 2008. This work was supported in part by NSF under Grant CCF-0448534 and Grant CCF-1017090 and in part by the National Natural Science Foundation of China (NSFC) under Grant 60828008. The work of J. Chen was supported in part by NSF under Grant CCF-0541456 and Grant ECCS-0801874, by the Natural Science Foundation of China under Grant 60628301, by the Hong Kong RGC under Project 111810, and by the City University of Hong Kong under Project 9380054.

B. Yan is with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77840 USA (e-mail: byan@neo.tamu.edu).

S. X.-D. Tan and R. Shen are with the Department of Electrical Engineering, University of California, Riverside, CA 92521 USA (e-mail: stan@ee.ucr.edu; rshen@ee.ucr.edu).

L. Zhou is with the Servo Engineering Department, Western Digital, Irvine, CA 92612 USA (e-mail: lingfei.zhou@wdc.com).

J. Chen is with the Department of Electronic Engineering, City University of Hong Kong, Hong Kong, China (e-mail: jichen@cityu.edu.hk).

Digital Object Identifier 10.1109/TVLSI.2011.2126612

single-output (SISO) pairings in an MIMO linear time invariant (LTI) system.

We will show that, given a passive network, when the decoupled subsystems are reduced by passive reduction methods, then the overall reduced system will remain passive. We will also show that this method can be applied with both Krylov subspace methods and balanced truncation methods. Further, we will demonstrate experimentally that the new reduction algorithm, called *DeMOR*, can be efficient for interconnect circuits especially with dominant resistance couplings like on-chip power grids and substrate networks.

This paper is organized as follows. In Section II, we review relevant model order reduction methods. In Section III, we introduce the concept of relative gain array. In Section IV, we propose the decentralized framework for Krylov subspace methods and prove the preservation of passivity. In Section V, we show that the proposed method can be applied together with balanced truncation methods to generate more compact models for systems with a massive number of inputs. Simulation results are given in Section VI to demonstrate the effectiveness of our proposed method. Section VII concludes this paper.

II. Model Order Reduction

In this section, we review the model reduction framework and problems with the existing Krylov subspace methods.

A. Projection-Based Reduction Framework

An interconnect circuit can be formulated by the following state-space form using modified nodal analysis (MNA)

$$\mathbf{C}\dot{\mathbf{x}}(t) = -\mathbf{G}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \quad \mathbf{y}(t) = \mathbf{L}^T\mathbf{x}(t) \tag{1}$$

where $\mathbf{C}, \mathbf{G} \in \mathbb{R}^{n\times n}$, $\mathbf{B}, \mathbf{L} \in \mathbb{R}^{n\times p}$, $\mathbf{x}(t)$ is the state vector, and $\mathbf{u}(t)$ and $\mathbf{y}(t)$ represent the input and output, respectively. Typically, we have $p \ll n$. Model reduction algorithms seek to produce a smaller system

$$\tilde{\mathbf{C}}\dot{\tilde{\mathbf{x}}}(t) = -\tilde{\mathbf{G}}\tilde{\mathbf{x}}(t) + \tilde{\mathbf{B}}\mathbf{u}(t) \quad \tilde{\mathbf{y}}(t) = \tilde{\mathbf{L}}^T\tilde{\mathbf{x}}(t) \tag{2}$$

where $\tilde{\mathbf{C}}, \tilde{\mathbf{G}} \in \mathbb{R}^{r\times r}$, $\tilde{\mathbf{B}}, \tilde{\mathbf{L}}, \in \mathbb{R}^{r\times p}$. The reduced order r is much lower than the original order n, i.e., $r \ll n$, but the output $\mathbf{y}(t)$ and $\tilde{\mathbf{y}}(\mathbf{t})$ are approximately equal for inputs $\mathbf{u}(t)$ of interest. This can be achieved by constructing matrices $\mathbf{W}$ and $\mathbf{V}$, whose columns span a useful subspace, and by projecting the original equations in the column spaces of $\mathbf{W}$ and $\mathbf{V}$, i.e.,

$$\tilde{\mathbf{C}} = \mathbf{W}^T\mathbf{C}\mathbf{V}, \tilde{\mathbf{G}} = \mathbf{W}^T\mathbf{G}\mathbf{V}, \tilde{\mathbf{B}} = \mathbf{W}^T\mathbf{B}, \tilde{\mathbf{L}} = \mathbf{V}^T\mathbf{L}. \tag{3}$$

The accuracy of the reduced model can be measured by the approximation error between the two transfer functions matrices

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

For a prescribed threshold $\epsilon > 0$, if $\|\mathbf{H}(s) - \tilde{\mathbf{H}}(s)\| \le \epsilon$, for some norm in a region of the complex plane, then the reduced model is accepted to be accurate.

B. Problems With Existing Krylov Subspace Methods

The Krylov subspace $K_m(\mathcal{T}, \mathcal{R})$ generated by a matrix $\mathcal{A}$ and matrix $\mathcal{R}$, of order m, is the space spanned by the set of vectors $(\mathcal{R}, \mathcal{A}\mathcal{R}, \mathcal{A}^2\mathcal{R}, \ldots, \mathcal{A}^{m-1}\mathcal{R})$. Usually projection matrices $\mathbf{V}$ and $\mathbf{W}$ are constructed so that their columns span a Krylov subspace. For example, a typical implementation is to construct $\mathbf{V} = \mathbf{W}$ using the Arnoldi algorithm, thereby spanning a Krylov subspace with

$$\mathcal{A} = (\mathbf{G}+s_0\mathbf{C})^{-1}\mathbf{C} \quad \mathcal{R} = (\mathbf{G}+s_0\mathbf{C})^{-1}\mathbf{B}. \tag{5}$$

Because of the moment-matching properties of Krylov subspace, the reduced transfer function $\tilde{\mathbf{H}}(s)$ will agree with the original $\mathbf{H}(s)$ up to the first m derivatives on an expansion point in the complex plane (usually $s_0 = 0$), which results in

$$\mathbf{H}(s) = \tilde{\mathbf{H}}(s) + O((s-s_0)^m). \tag{6}$$

However, the cost associated with model computation and the size of the reduced model is directly proportional to the number of inputs in the Krylov subspace method, i.e., to the number of columns in the transfer function matrix. For example, in the PRIMA algorithm [18], if only two (block) moments are to be matched at each port and the network has 1000 ports, the resulting reduced model will have 2000 states.

C. SVDMOR

The SVDMOR/RecMOR methods [7], [9] were first proposed to explicitly reduce the terminals in the projection-based reduction framework. For many practical circuits, the input-output correspondence at various ports may be highly correlated. In this case, the input-output matrices can be approximated using low-rank matrices. Applying the standard SVD to the system transfer function matrix at dc gives

$$\mathbf{H}_{\mathrm{DC}} = \mathbf{L}^T\mathbf{G}^{-1}\mathbf{B} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{7}$$

where $\mathbf{U}$ and $\mathbf{V}$ are matrices that consist of the left and right singular vectors. If there exists a strong correlation between the responses at different input-output ports, the transfer function matrix can be well approximated based on $k(k < p)$ dominant left and right singular vectors in $\mathbf{U}_k$ and $\mathbf{V}_k$. These singular vectors are used to find a rank-k approximation for $\mathbf{L}$ and $\mathbf{B}$, i.e.,

$$\mathbf{B} \approx \mathbf{B}_k\mathbf{V}_k^T \quad \mathbf{L} \approx \mathbf{L}_k\mathbf{U}_k^T \tag{8}$$

where $\mathbf{B}_k \in \mathbb{R}^{n\times k}$ and $\mathbf{L}_k \in \mathbb{R}^{n\times k}$. The original transfer function matrix can then be approximated as

$$\mathbf{H}(s) \approx \mathbf{U}_k\mathbf{L}_k^T(s\mathbf{C}+\mathbf{G})^{-1}\mathbf{B}_k\mathbf{V}_k^T. \tag{9}$$

Since $\mathbf{H}_k(s) \approx \mathbf{L}_k^T(s\mathbf{C}+\mathbf{G})^{-1}\mathbf{B}_k$ represents a terminal-reduced MIMO network with $k(k < p)$ ports, it can be more readily reduced by existing reduction methods.

It should be pointed out that SVDMOR is a centralized reduction method. While it may perform well for low rank transfer

function matrices, a network with a large number of ports rarely has a low-rank transfer function matrix.

III. Measurement of Interaction

In this section, we introduce a tool to measure the degree of interaction of each input-output pair in an MIMO system.

A. Ideal Decentralized Systems

The input-output relationship of a $p \times p$ LTI system can be described by

$$\mathbf{y}(s) = \mathbf{H}(s)\mathbf{u}(s) \tag{10}$$

where $\mathbf{u}(s)$ and $\mathbf{y}(s)$ are p-dimensional vectors of inputs and outputs, respectively, and $\mathbf{H}(s)$ is the system's transfer function matrix. For a diagonal transfer function matrix

$$\mathbf{H}(s) = \begin{bmatrix} h_{11}(s) & 0 & \dots & 0 \\ 0 & h_{22}(s) & \dots & 0 \\ \dots & \dots & & \dots \\ 0 & 0 & \dots & h_{pp}(s) \end{bmatrix} \tag{11}$$

no interaction exists between different inputs and outputs and the system consists of p independent subsystems. In this case, if we are only interested in one of the response $y_i(s)$, where $i = 1, \dots, p$, it suffices to compute the subsystem $h_{ii}(s)$. Clearly, this computation can be performed in parallel for all the subsystems.

It is of course very often that output and input variables are coupled and one output may interact with many inputs. It is plausible, however, that in many cases of interest, one output may interact more strongly with a subset of inputs than with the others. An important observation in this paper is that this property can be exploited by quantifying the degree of interaction of each input-output pair in an MIMO system. This effort is facilitated by the use of RGA.

B. Relative Gain Array

RGA is a matrix of interaction measures for all possible single-input single-output (SISO) pairings in an MIMO LTI system [2]. This concept has found widespread utility in process control, and as a system robustness measure [4]. For a system $\mathbf{H}(s)$ with p inputs and p outputs, there are $p \times p$ relative gain elements λ_{ij}, which together form the matrix

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \dots & \lambda_{1p} \\ \lambda_{21} & \lambda_{22} & \dots & \lambda_{2p} \\ \dots & \dots & & \dots \\ \lambda_{p1} & \lambda_{p2} & \dots & \lambda_{pp} \end{bmatrix}. \tag{12}$$

The relative gain λ_{ij} between an output y_i and an input u_j is given by

$$\lambda_{ij} = \frac{g_{ij}^o}{g_{ij}^c} \tag{13}$$

where g_{ij}^o and g_{ij}^c are the open and closed loop gains of the transfer function $h_{ij}(s)$, respectively.

First, assume that except u_j, all other inputs $u_k (k = 1, \dots, p, k \neq j)$ are zeros, a step change in input u_j of magnitude $\triangle u_j$ will produce a change $\triangle y_i$ in output y_i. Thus, the gain between u_j and y_i when the other inputs are kept zeros is given by

$$g_{ij}^o = \frac{\triangle y_i}{\triangle u_j}\Big|_{u_k = 0 (k \neq j)} \tag{14}$$

which can be viewed as an open loop gain with respect to other inputs.

Second, except y_i, when keeping all the other outputs $y_l (l = 1, \dots, p, l \neq i)$ zeros, a step change in input u_j of magnitude $\triangle u_j$ will result in another change in y_i. In this process, other outputs will also be affected due to cross-coupling. In order to keep them zeros, we need to adjust other inputs correspondingly, which will also contribute to the change in y_i. The gain under the new set of conditions is denoted by

$$g_{ij}^c = \frac{\triangle y_i}{\triangle u_j}\Big|_{y_l = 0 (l \neq i)} \tag{15}$$

which can be viewed as a closed loop gain with respect to other inputs.

Although the above gains are between the same pair of variables, they may have different values because they have been obtained under different conditions. If interaction exists, the change in y_i due to a change in u_j for the two cases (when other inputs and when other outputs are kept zeros), will be different. The ratio

$$\lambda_{ij} = \frac{\frac{\triangle y_i}{\triangle u_j}\big|_{u_k = 0 (k \neq j)}}{\frac{\triangle y_i}{\triangle u_j}\big|_{y_l = 0 (l \neq i)}} \tag{16}$$

defines the relative gain between the output y_i and input u_j.

There are two extreme cases: first, if $\lambda_{ij} = 0$, y_i is not influenced by u_j at all; second, if $\lambda_{ij} = 1$, closed loop gain is equal to open loop gain, which means the interaction from other inputs is zero and y_i is influenced by u_j only. In fact, by taking the absolute value of each RGA element and taking the inverse for those larger than 1, the scaled elements $\tilde{\lambda}_{ij}$ will fall into the range of $[0, 1]$

$$\tilde{\lambda}_{ij} = |\lambda_{ij}| (|\lambda_{ij}| \leq 1) \quad \tilde{\lambda}_{ij} = \frac{1}{|\lambda_{ij}|} (|\lambda_{ij}| > 1). \tag{17}$$

Usually, the larger the scaled number is, the more important the corresponding input will be. For a given output i, the contribution of each input can be easily compared and those inputs can be arranged in a descending order in terms of their contribution. For a large number of systems in practice, most input-output pairs are magnitude-wise insignificant and their corresponding values are close to zero. In this case, an output is only predominantly influenced by a small number of dominant inputs and those inputs could be identified with the guidance of RGA.

C. Computation of RGA

Generally, the relative gain array of the system $\mathbf{H}(s)$ can be shown to be the frequency dependent function

$$\mathbf{\Lambda}(s) = \mathbf{H}(s) \circ \mathbf{H}(s)^{-T} \tag{18}$$

where $\circ$ denotes element-by-element multiplication (often called the Hadamard or Schur product) [38], and $\mathbf{H}^{-T}$ is the transpose of $\mathbf{H}^{-1}$.

For control system applications, as steady-state performance is usually more important, the relative gain array is typically evaluated at zero frequency (steady-state relative gain array)

$$\mathbf{\Lambda}(0) = \mathbf{H}(0) \circ \mathbf{H}(0)^{-T}. \tag{19}$$

For systems with non-square transfer matrix, we can use *pseudoinverse* instead

$$\mathbf{\Lambda}(0) = \mathbf{H}(0) \circ (\mathbf{H}(0)^{T})^{+}. \tag{20}$$

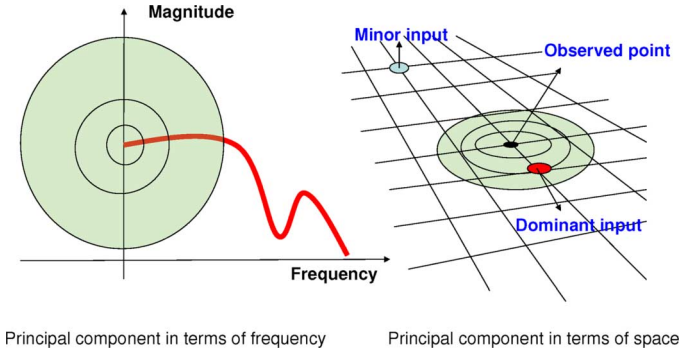


Fig. 1. Principal components in terms of frequency and electrical distance.

IV. Decentralized Model Order Reduction

In this section, we first propose a dominant Krylov subspace method to build compact models corresponding to individual outputs. Then we present the decentralized framework and prove the preservation of passivity.

A. Dominant Krylov Subspace Method

For the interconnect circuit in (1), the transfer function is

$$\mathbf{H}(s) = \mathbf{L}^{T}(\mathbf{C}s + \mathbf{G})^{-1}\mathbf{B} \tag{21}$$

and the steady-state gain $\mathbf{H}(0)$ is the dc gain $\mathbf{H}_{\mathrm{DC}}$

$$\mathbf{H}_{\mathrm{DC}} = \mathbf{L}^{T}\mathbf{G}^{-1}\mathbf{B}. \tag{22}$$

The RGA at dc can be computed as

$$\mathbf{\Lambda}(0) = \mathbf{H}_{\mathrm{DC}} \circ \mathbf{H}_{\mathrm{DC}}^{-T}. \tag{23}$$

Although RGA is frequency dependent in general, we only evaluate it at dc as the methods in this paper are mainly proposed for resistance-dominant interconnects like on-chip power grids, substrate planes, and clock meshes, which behave as low pass filters [3], [5], [19]–[21]. For a low-pass filter, the attenuation of high frequency components is much faster than the attenuation of low frequency components, which means high frequency components tend to be more localized and the evaluation of RGA at dc is actually valid for all the frequencies.

If we are interested in an individual output, the ith output, a projection matrix $\mathbf{V}_i$ is constructed so that the columns span a *dominant Krylov subspace* $K_m(\mathcal{A}, \mathcal{R}_i)$, where

$$\mathcal{A} = (\mathbf{G} + s_0\mathbf{C})^{-1}\mathbf{C} \quad \mathcal{R}_i = (\mathbf{G} + s_0\mathbf{C})^{-1}\mathbf{B}_i. \tag{24}$$

In this approach, instead of all the inputs, $\mathbf{B}_i$ is composed of the dominant inputs of the ith output, which are identified based on RGA. Then the reduced model for the ith output $\tilde{\mathbf{H}}_i(s)$ is obtained by

$$\tilde{\mathbf{C}}_i = \mathbf{V}_i^{T}\mathbf{C}\mathbf{V}_i, \tilde{\mathbf{G}}_i = \mathbf{V}_i^{T}\mathbf{G}\mathbf{V}_i, \tilde{\mathbf{B}}_i = \mathbf{V}_i^{T}\mathbf{B}, \tilde{\mathbf{L}}_i = \mathbf{V}_i^{T}\mathbf{L}. \tag{25}$$

In the reduction process, although only dominant inputs are used to construct the projection matrix, we do not eliminate any column corresponding to insignificant input from the input matrix. By doing this, the signal transfers from those insignificant inputs are still coarsely preserved because those inputs, although very little individually, could not be ignored when adding up together. Therefore, if the original model is a $p \times p$ system, the reduced model $\tilde{\mathbf{H}}_i(s)$ is still a $p \times p$ system. Note that, only the ith output is to be used in future simulation.

If the original model (1) has the following structure information:

$$\mathbf{C} \geq 0 \quad \mathbf{G} + \mathbf{G}^{T} \geq 0 \quad \mathbf{B} = \mathbf{L} \tag{26}$$

where ≥ 0 means positive semi-definite, after the projection (25), the reduced model will inherit the structure information such that

$$\tilde{\mathbf{C}}_i \geq 0 \quad \tilde{\mathbf{G}}_i + \tilde{\mathbf{G}}_i^{T} \geq 0 \quad \tilde{\mathbf{B}}_i = \tilde{\mathbf{L}}_i \tag{27}$$

which means the reduced model $\tilde{\mathbf{H}}_i(s)$ is passive [18].

Different from existing Krylov subspace methods, where only principle components in terms of frequency is considered, the dominant Krylov subspace method takes into consideration principal components in terms of both frequency (temporal) and electrical distance to reduce the system complexity, as illustrated in Fig. 1.

B. Decentralized Framework

If more outputs are to be observed, reduced models can be built for each output, which results in a decentralized framework. In such a framework, the computation can be performed individually and in parallel as the reduced models corresponding to different outputs are independent.

The DeMOR algorithm based on the dominant Krylov subspace is shown in Algorithm. 1. Note that, as the subsystems corresponding to different outputs share the same system matrix $\mathbf{G}$, only one sparse matrix factorization is required in the whole algorithm.

Decentralized Model Order Reduction

Input: $\mathbf{H} : (\mathbf{G}, \mathbf{C}, \mathbf{B}, \mathbf{L}), m$

Output: $\tilde{\mathbf{H}}_i : (\tilde{\mathbf{G}}_i, \tilde{\mathbf{C}}_i, \tilde{\mathbf{B}}_i, \tilde{\mathbf{L}}_i)(i = 1, \ldots, p)$

1. Solve $\mathbf{GM} = \mathbf{B}$ for $\mathbf{M}_0$
2. Compute $\mathbf{H}_{\mathrm{DC}} = \mathbf{L}^{T}\mathbf{M}_0$

3. Compute relative gain array $\mathbf{\Lambda}(0) = \mathbf{H}_{\mathrm{DC}} \circ \mathbf{H}_{\mathrm{DC}}^{-T}$
4. Scale the RGA values to the range of [0,1]
5. For output i $(i = 1, \ldots, p)$
   Determine the dominant input matrix $\mathbf{B}_i$ corresponding to $\tilde{p}_i$ dominant inputs to construct a dominant Krylov subspace.
   Model order reduction using PRIMA to obtain $\tilde{\mathbf{H}}_i$

$$\tilde{\mathbf{C}}_i = \mathbf{V}_i^T \mathbf{C} \mathbf{V}_i, \tilde{\mathbf{G}}_i = \mathbf{V}_i^T \mathbf{G} \mathbf{V}_i,$$
$$\tilde{\mathbf{B}}_i = \mathbf{V}_i^T \mathbf{B}, \tilde{\mathbf{L}}_i = \mathbf{V}_i^T \mathbf{L}$$

where $\mathrm{colspan}(\mathbf{V}_i) = K_m(\mathbf{G}^{-1}\mathbf{C}, \mathbf{G}^{-1}\mathbf{B}_i)$

The reduced order $r = m\tilde{p}$ is determined by both the number of moments matched m and the number of dominant inputs $\tilde{p}$.

As other moment-matching methods, the number of moments matched m corresponds to the frequency range of accuracy. Typically, the reduced model is required to be accurate in a certain frequency range. A larger accurate frequency range can be obtained by matching more moments. However, it is hard to know exactly how much the frequency range will increase with one more moment matched.

Similarly, for the number of dominant inputs $\tilde{p}$, although RGA can provide some guidance, it is hard to know exactly how many inputs should be treated as dominant inputs to achieve the best tradeoff between accuracy and reduced size.

Therefore, given an example, a few tests are often needed to determine the best number of m and $\tilde{p}$.

C. Preservation of Passivity

After reduction (25), we get p reduced subsystems $\tilde{\mathbf{H}}_i(s) (i = 1, \ldots, p)$ and each subsystem is a p-port network, which has p inputs and p outputs. The reduced subsystems together can be viewed as a p^2-port network $\tilde{\mathcal{H}}(s)$, i.e.,

$$\tilde{\mathcal{H}}(s) = \begin{bmatrix} \tilde{\mathbf{H}}_1(s) & \mathbf{O} & \ldots & \mathbf{O} \\ \mathbf{O} & \tilde{\mathbf{H}}_2(s) & \ldots & \mathbf{O} \\ \ldots & \ldots & & \ldots \\ \mathbf{O} & \mathbf{O} & \ldots & \tilde{\mathbf{H}}_p(s) \end{bmatrix} = \tilde{\mathcal{L}}^T (s\tilde{\mathcal{C}} + \tilde{\mathcal{G}})^{-1} \tilde{\mathcal{B}} \tag{28}$$

where $\tilde{\mathcal{C}}$, $\tilde{\mathcal{G}}$, $\tilde{\mathcal{B}}$, and $\tilde{\mathcal{L}}$ are block diagonal matrices

$$\tilde{\mathcal{C}} = \begin{bmatrix} \tilde{\mathbf{C}}_1 & \mathbf{O} & \ldots & \mathbf{O} \\ \mathbf{O} & \tilde{\mathbf{C}}_2 & \ldots & \mathbf{O} \\ \ldots & \ldots & & \ldots \\ \mathbf{O} & \mathbf{O} & \ldots & \tilde{\mathbf{C}}_p \end{bmatrix}$$
$$\tilde{\mathcal{G}} = \begin{bmatrix} \tilde{\mathbf{G}}_1 & \mathbf{O} & \ldots & \mathbf{O} \\ \mathbf{O} & \tilde{\mathbf{G}}_2 & \ldots & \mathbf{O} \\ \ldots & \ldots & & \ldots \\ \mathbf{O} & \mathbf{O} & \ldots & \tilde{\mathbf{G}}_p \end{bmatrix}$$
$$\tilde{\mathcal{B}} = \begin{bmatrix} \tilde{\mathbf{B}}_1 & \mathbf{O} & \ldots & \mathbf{O} \\ \mathbf{O} & \tilde{\mathbf{B}}_2 & \ldots & \mathbf{O} \\ \ldots & \ldots & & \ldots \\ \mathbf{O} & \mathbf{O} & \ldots & \tilde{\mathbf{B}}_p \end{bmatrix}$$
$$\tilde{\mathcal{L}} = \begin{bmatrix} \tilde{\mathbf{L}}_1 & \mathbf{O} & \ldots & \mathbf{O} \\ \mathbf{O} & \tilde{\mathbf{L}}_2 & \ldots & \mathbf{O} \\ \ldots & \ldots & & \ldots \\ \mathbf{O} & \mathbf{O} & \ldots & \tilde{\mathbf{L}}_p \end{bmatrix}. \tag{29}$$

As the reduced subsystems $\tilde{\mathbf{H}}_i(s)$ have the structure information (27), we have

$$\tilde{\mathcal{C}} \geq 0 \quad \tilde{\mathcal{G}} + \tilde{\mathcal{G}}^T \geq 0 \quad \tilde{\mathcal{B}} = \tilde{\mathcal{L}} \tag{30}$$

because a block diagonal matrix is positive semi-definite if and only if each diagonal block is positive semi-definite. This means the network $\tilde{\mathcal{H}}(s)$ is passive [18].

Let u_{ij} denote the jth input at the ith subsystem and y_{ij} denote the jth output at the ith subsystem , where $i = 1, \ldots, p$ and $j = 1, \ldots, p$. The input-output relationship of the passive network $\tilde{\mathcal{H}}(s)$ is given as follows:

$$\begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1p} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2p} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ y_{p1} \\ y_{p2} \\ \vdots \\ y_{pp} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{H}}_1(s) & \mathbf{O} & \ldots & \mathbf{O} \\ \mathbf{O} & \tilde{\mathbf{H}}_2(s) & \ldots & \mathbf{O} \\ \ldots & \ldots & & \ldots \\ \mathbf{O} & \mathbf{O} & \ldots & \tilde{\mathbf{H}}_p(s) \end{bmatrix} \begin{bmatrix} u_{11} \\ u_{12} \\ \vdots \\ u_{1p} \\ u_{21} \\ u_{22} \\ \vdots \\ u_{2p} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ u_{p1} \\ u_{p2} \\ \vdots \\ u_{pp} \end{bmatrix}. \tag{31}$$

For this p^2-port network, we use Port_{ij} to denote the jth port of the ith reduced subsystem and (u_{ij}, y_{ij}) are the corresponding input-output variables.

Among the p^2 ports, there are p *external* ports $\mathrm{Port}_{ii} (i = 1, \ldots, p)$, which corresponds to the p ports of the original system $\mathbf{H}(s)$. For those p ports, the input variables $u_{ii} (i = 1, \ldots, p)$ correspond to the input variables of the original model $\mathbf{H}(s)$ $u_i (i = 1, \ldots, p)$

$$u_{ii} = u_i (i = 1, \ldots, p) \tag{32}$$

and the output variables $y_{ii} (i = 1, \ldots, p)$ are approximant to the output variables of the original model $\mathbf{H}(s)$ $y_i (i = 1, \ldots, p)$ due to reduction

$$y_{ii} \approx y_i (i = 1, \ldots, p). \tag{33}$$

The other $p^2 - p$ ports are *internal* ports, where the input variables $u_{ki} (k \neq i)$ are controlled by the input variables of the p external ports u_{ii} as follows:

$$u_{ki} = u_{ii} (i = 1, \ldots, p, k \neq i). \tag{34}$$

This means the $p^2 - p$ internal ports are interconnected with $p^2 - p$ controlled sources, respectively.

Therefore, in terms of the p external ports, the decentralized reduced system described by (31) (32) (33) (34) is a p-port network composed of a passive model $\tilde{\mathcal{H}}(s)$ (31) interconnected

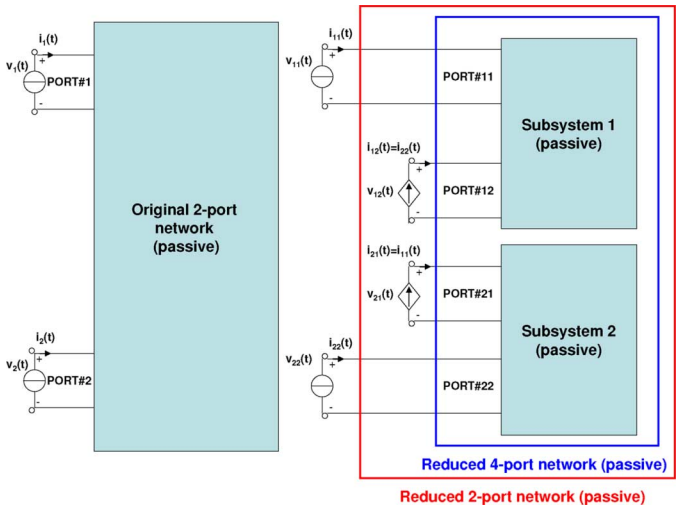


Fig. 2. Passivity of decentralized reduced model.

with p^2-p controlled sources (34). Note that, controlled sources are passive elements. As interconnection of passive systems are still passive, the decentralized reduced system is passive [18].

To better illustrate this point, we use a 2-port passive network as an example. As shown in Fig. 2 (left), the original model is a 2-port passive network and each port is connected with a current source as external stimuli. After reduction via DeMOR, two subsystems are obtained and each subsystem itself is a passive 2-port network (27). In subsystem 1, the voltage response at the first port $v_{11}(t)$ is the approximant to the voltage response at the first port of the original model $v_1(t)$. In subsystem 2, the voltage response at the second port $v_{22}(t)$ is the approximant to the voltage response at the second port of the original model $v_2(t)$. The two subsystems together (in the blue box) can be viewed as a 4-port network, which is also passive (30). Among the four ports, the first (Port_{11}) and the last port (Port_{22}) are external ports. The second (Port_{12}) and the third port (Port_{21}) are internal ports, where the input currents $i_{12}(t), i_{21}(t)$ are controlled by the currents at the last port $i_{22}(t)$ and the first port $i_{11}(t)$, respectively. The decentralized reduced model corresponds to the parts inside the red box, which is composed of the interconnection of a passive 4-port model (in the blue box) with two current controlled current sources, which are also passive elements. Therefore, the overall decentralized reduced system is passive.

D. Computational Cost Analysis

First, we discuss the reduction cost of DeMOR compared with PRIMA.

Given an interconnect model of order n and with p ports, where $p \ll n$, to match m moments of the original model, Krylov subspace method PRIMA first takes a matrix factorization of the sparse matrix $\mathbf{G}$ at the cost of $O(n^\beta)$. In the Arnoldi procedure, it takes mp solves for the generation of vectors at the cost of $O(mpn^\alpha)$, and a QR factorization of the mp vectors at the cost of $O((mp)^2 n)$ (typically, $1 \le \alpha \le 1.2$ and $1.1 \le \beta \le 1.5$ for sparse circuits). So the total cost of reduction in PRIMA is

$$O(n^\beta + mpn^\alpha + (mp)^2 n). \tag{35}$$

For DeMOR method, we first need to compute the dc response $\mathbf{H}_{\mathrm{DC}}$, which mainly takes one matrix factorization of the sparse matrix $\mathbf{G}$ at the cost of $O(n^\beta)$. The computational cost of RGA from $\mathbf{H}_{\mathrm{DC}}$ is $O(p^3)$. Given an output of interest, if a number of $\tilde{p}$ dominant inputs are identified based on RGA, where $\tilde{p} \ll p$, in order to match m moments corresponding to $\tilde{p}$ inputs, the Arnoldi procedure requires $m\tilde{p}$ solves for the generation of vectors at the cost of $O(m\tilde{p}n^\alpha)$, and a QR factorization of the $m\tilde{p}$ vectors at the cost of $O((m\tilde{p})^2 n)$ after the factorization of matrix $\mathbf{G}$. So the total reduction cost for one subsystem is

$$O(n^\beta + p^3 + m\tilde{p}n^\alpha + (m\tilde{p})^2 n). \tag{36}$$

Note that, if all the p outputs are to be observed, we sill only need to perform one matrix factorization of $\mathbf{G}$ and the total cost is

$$O(n^\beta + p^3 + pm\tilde{p}n^\alpha + p(m\tilde{p})^2 n). \tag{37}$$

For a large number of problems of interests, where $\tilde{p} \ll p$ and $p \ll n$, the cost of both methods is usually dominated by the sparse matrix factorization $O(n^\beta)$. Even if DeMOR is more expensive than PRIMA, the overhead will not have much effect on the overall efficiency of Krylov subspace methods for large-scale sparse systems.

Second, we discuss the simulation cost of the reduced models. If the original system has p ports, then a reduced system of order mp is needed for PRIMA to match m block moments. The cost of solving the reduced system is $O(m^3p^3)$ as the matrices in the reduced system are full. On the other hand, if we decentralize the original system into p subsystems and assume each subsystem has $\tilde{p}$ dominant inputs ($\tilde{p} \ll p$), to match the same number of moments m, the reduced order of each subsystem is $m\tilde{p}$. If only one output is to be observed, only one reduced subsystem needs to be solved at the cost of $O(m^3\tilde{p}^3)$. If all the outputs are to be observed, the cost to solve a number of p reduced subsystems is $O(pm^3\tilde{p}^3)$. In the second case, the cost ratio for solving the two reduced models is

$$\frac{O(m^3p^3)}{O(pm^3\tilde{p}^3)} \gg 1 \ (\text{for } p \gg \tilde{p}) \tag{38}$$

which means that solving DeMOR reduced model is much less expensive.

Moreover, since the reduced subsystems are independent, the runtime can be reduced significantly if the parallel computing can be applied in the simulation process. The proposed method is most efficient for large systems with a massive number of inputs, where only a small number of outputs are to be observed.

V. Extension to Balanced Truncation

In this section, we extend the proposed method to another family of model reduction methods: balanced truncation.

A. Balanced Truncation

Given a stable LTI system with minimal realization $(\mathbf{A}, \mathbf{B}, \mathbf{C})$

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \quad \mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) \tag{39}$$

where $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times p}$, $\mathbf{C} \in \mathbb{R}^{q\times n}$, $\mathbf{y}(t) \in \mathbb{R}^q$, $\mathbf{u}(t) \in \mathbb{R}^p$, the controllability and observability gramians are the unique positive definite solutions to the Lyapunov equations

$$\mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T = 0 \tag{40}$$

$$\mathbf{A}^T\mathbf{Y} + \mathbf{Y}\mathbf{A} + \mathbf{C}^T\mathbf{C} = 0. \tag{41}$$

Since the eigenvalues of the product $\mathbf{XY}$ are input-output invariant [17], we can perform a similarity transformation

$$\mathbf{A}_b = \mathbf{T}^{-1}\mathbf{A}\mathbf{T},\ \mathbf{B}_b = \mathbf{T}^{-1}\mathbf{B},\ \mathbf{C}_b = \mathbf{C}\mathbf{T} \tag{42}$$

to diagonalize the product $\mathbf{XY}$

$$\mathbf{T}^{-1}\mathbf{X}\mathbf{Y}\mathbf{T} = \mathbf{\Sigma} = \operatorname{diag}\left(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2\right) \tag{43}$$

where the Hankel singular values σ_k of the system are arranged in a descending order.

The resulted coordinate system$(\mathbf{A}_b, \mathbf{B}_b, \mathbf{C}_b)$ is called a balanced realization and the states of the balanced system corresponding to the small Hankel singular values are difficult to reach and to observe at the same time. Such states are less involved in the energy transfer from inputs to outputs.

If we partition $\mathbf{\Sigma}$ into

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_1 & 0 \\ 0 & \mathbf{\Sigma}_2 \end{bmatrix} \tag{44}$$

and partition the transformed matrices as

$$\mathbf{A}_b = \begin{bmatrix} \mathbf{A}_{b11} & \mathbf{A}_{b12} \\ \mathbf{A}_{b21} & \mathbf{A}_{b22} \end{bmatrix} \quad \mathbf{B}_b = \begin{bmatrix} \mathbf{B}_{b1} \\ \mathbf{B}_{b2} \end{bmatrix} \quad \mathbf{C}_b = [\,\mathbf{C}_{b1} \quad \mathbf{C}_{b2}\,] \tag{45}$$

a reduced model $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}})$ of order r

$$\dot{\tilde{\mathbf{x}}}(t) = \tilde{\mathbf{A}}\tilde{\mathbf{x}}(t) + \tilde{\mathbf{B}}\mathbf{u}(t) \quad \tilde{\mathbf{y}}(t) = \tilde{\mathbf{C}}\tilde{\mathbf{x}}(t) \tag{46}$$

is obtained by taking the $r \times r$, $r \times p$, $q \times r$ leading blocks of $\mathbf{A}_b$, $\mathbf{B}_b$, $\mathbf{C}_b$, respectively

$$\tilde{\mathbf{A}} = \mathbf{A}_{b11} \quad \tilde{\mathbf{B}} = \mathbf{B}_{b1} \quad \tilde{\mathbf{C}} = \mathbf{C}_{b1}. \tag{47}$$

B. Modified Balanced Truncation for Systems With Massive Inputs

Given the system$(\mathbf{A}, \mathbf{B}, \mathbf{C})$ (39), if the number of inputs of the system p is much larger than the number of outputs q

$$q \ll p \tag{48}$$

the dimension of controllable subspace is hight and the decay of eigenvalues of the gramian product $\mathbf{XY}$ will be slow even if the number of outputs is small. This means a higher order reduced model is needed based on the criteria of balanced truncation.

The system$(\mathbf{A}, \mathbf{B}, \mathbf{C})$(39) can be decomposed as p single input systems $(\mathbf{A}, \mathbf{b}_j, \mathbf{C})$ [6], where $\mathbf{b}_j (j = 1, 2, \ldots, p)$ is the jth column of $\mathbf{B}$. Let $\mathbf{X}_j$ be the controllability gramians for the system $(\mathbf{A}, \mathbf{b}_j, \mathbf{C})$ satisfying

$$\mathbf{A}\mathbf{X}_j + \mathbf{X}_j\mathbf{A}^T + \mathbf{b}_j\mathbf{b}_j^T = 0. \tag{49}$$

Given the fact that

$$\mathbf{B}\mathbf{B}^T = \sum_{j=1}^{p} \mathbf{b}_j\mathbf{b}_j^T \tag{50}$$

it is easy to show that the controllability gramian of the system $\mathbf{X}$ can be decomposed as the sum of the controllability gramians associated to the each single input system [6]

$$\mathbf{X} = \sum_{j=1}^{p} \mathbf{X}_j. \tag{51}$$

In this summation, each input is implicitly assumed to have an equal weight. As a result, if the number of inputs is large, the decay of the eigenvalues of the gramian $\mathbf{X}$ will be slow.

However, as shown in the previous section, those inputs are not equally important. In fact, in most cases, only a small number of inputs are dominant. To take advantage of this property, we present a modified balanced truncation method for systems with a large number of inputs (48). Similar to the dominant Krylov subspace method, we first construct an input matrix $\mathbf{B}_d$ to include those dominant inputs, which are determined by the relative gain array. After that, a *dominant controllability gramian* $\mathbf{X}_d$ can be obtained by solving the following Lyapunov equation:

$$\mathbf{A}\mathbf{X}_d + \mathbf{X}_d\mathbf{A}^T + \mathbf{B}_d\mathbf{B}_d^T = 0. \tag{52}$$

The system can be balanced in terms of the gramian product $\mathbf{X}_d\mathbf{Y}$

$$\mathbf{T}^{-1}\mathbf{X}_\mathbf{d}\mathbf{Y}\mathbf{T} = \mathbf{\Sigma} = \operatorname{diag}\left(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2\right) \tag{53}$$

where $\mathbf{Y}$ is the observability gramian obtained from (41).

Given the matrix $\mathbf{T}$, similar to the classical balanced truncation, a reduced order model $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}})$ (46) can be obtained by balancing (42) and then truncating (47) the system. Note that, the balanced truncation is performed on $\mathbf{B}$ instead of $\mathbf{B}_d$, which means the energy transfers from weak inputs are also coarsely preserved. Given a small number of outputs, the proposed method often leads to very compact models for systems with a large number of inputs.

The cost of balanced truncation is dominated by the cost of solving Lyapunov equations. If the matrix $\mathbf{A}$ is sparse, the cost of solving Lyapunov equation (40) is $O(Jpn)$ with iterative methods [14], where p is the rank of input matrix $\mathbf{B}$ (the number of inputs p). Therefore, as the number of dominant inputs $\tilde{p}$ is much smaller $(\tilde{p} \ll p)$, the cost will be reduced from $O(Jpn)$ to $O(J\tilde{p}n)$ by solving (52) in the proposed method.

Note that, the proposed method cannot be relied to preserve passivity with classical passivity preserving balanced truncation method like PRTBR [22]. However, just like with PRIMA, passivity can be preserved with congruency transformation based balanced truncation methods like PMTBR [23] for interconnect modeling.

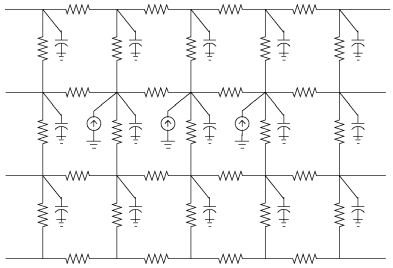


Fig. 3. Mesh structure network.

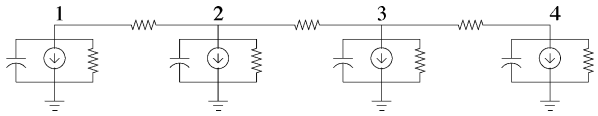


Fig. 4. Illustration of RGA.

VI. Experimental Results

The proposed methods have been implemented in MATLAB 7.0 and tested on a computer with an Intel core 2 duo 3.17 GHz processor and 4 GB memory. The performance of the proposed methods is demonstrated with examples with mesh structures (see Fig. 3), which are widely used for distributing critical global signals on a chip such as clock and power/ground [3], [5], [7], [9], [19], [21], [30], [35].

A. Example 1

First, we consider an illustrative example. As shown in Fig. 4, given a simple circuit with unit elements $(R = 1)$, the scaled relative gain array $|\Lambda(0)|$ is obtained as

$$|\Lambda(0)| = \begin{bmatrix} 0.808 & 0.238 & 0.000 & 0.000 \\ 0.238 & 0.700 & 0.191 & 0.000 \\ 0.000 & 0.191 & 0.700 & 0.238 \\ 0.000 & 0.000 & 0.238 & 0.808 \end{bmatrix} \tag{54}$$

where each row corresponds to one output (node voltage) and each column to one input (current source). For instance, if the voltage response at node 2 is of interests, the metric indicates that, in terms of importance, the inputs should be arranged in a descending order as $2 > 1 > 3 > 4$ and this can be verified by applying a constant current input at different nodes respectively to observe the corresponding voltage responses at node 2.

B. Example 2

The second example is a simple RC mesh ($R = 1\ \Omega$ and $C = 1$ pf) with 1600 nodes and 33 ports. In this example, we demonstrate how to divide and conquer an MIMO coupled network by decoupling it into a number of MISO subsystems and reduce each subsystem individually. As the reduced subsystems are independent, the simulation can be performed in parallel on the reduced subsystems.

We verify the reduction accuracy in the time and frequency domain. Current sources are connected to the ports to stimulate the circuit with a series of pulses of unit magnitude. The voltage responses at the ports are observed. The RGA value is shown in Fig. 5. We can see that the most input-output pairs are magnitude-wise insignificant and their corresponding values are close to zero.

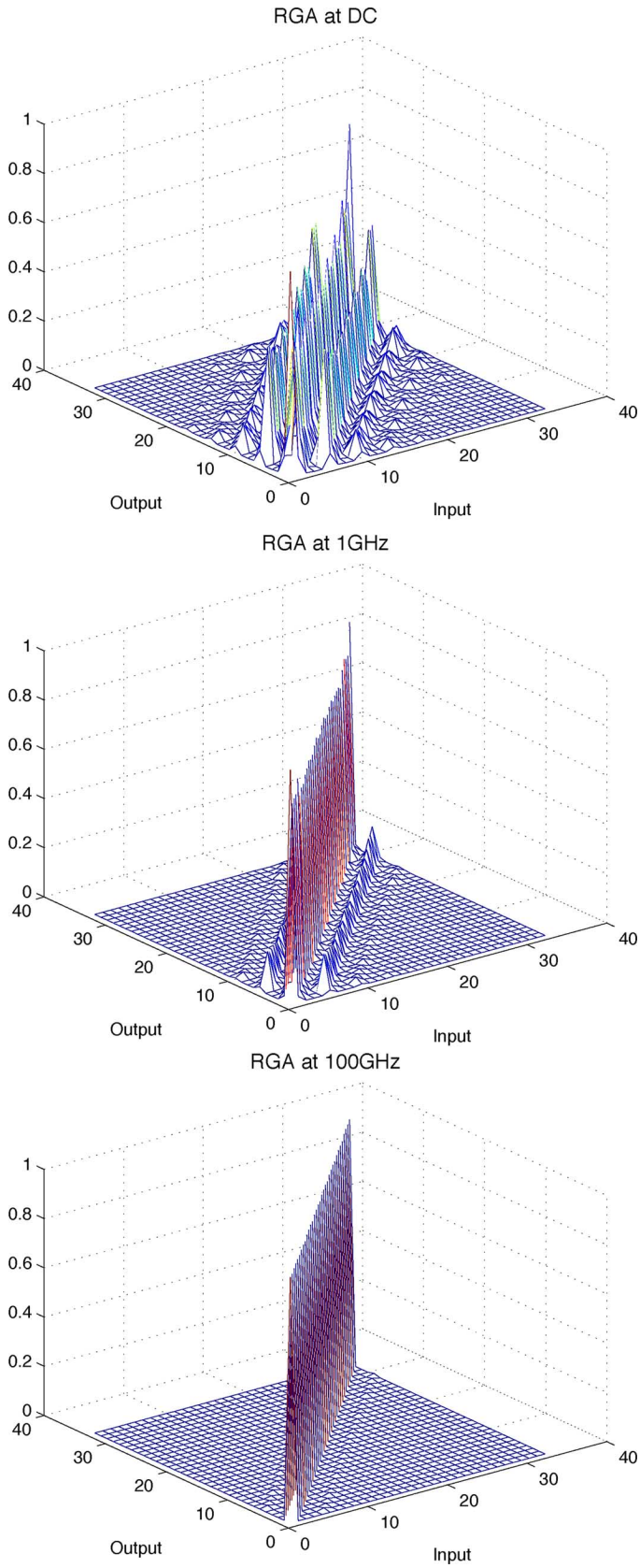


Fig. 5. RGA computed at different frequencies.

In addition, as the frequency increases, RGA becomes more and more locally concentrated. This is because the RC networks behave as low pass filters. For a low-pass filter, the attenuation of high frequency components is much faster than the attenuation of low frequency components, which means that high frequency components tend to be more localized and the evaluation of RGA at dc is actually valid for all the frequencies.

First, we build the reduced model for the voltage response at port 1. From Fig. 6, we see that port 1 is only dominantly interacted with itself. In this case, a reduced model of order 7 can match the original output well. The results of PRIMA, SVDMOR, and DeMOR are shown in Fig. 6 with the same order. We notice that SVDMOR does not work well. The reason is that dc matrix has full rank, which is usually the case for a complete matrix-valued transfer function. The frequency responses at port 1 are also shown in Fig. 6. The simulation times of the reduced models of order 7 is 0.006 s.

Second, we build the reduced model for the voltage response at port 12, which is located in the center of the circuit. From the RGA values, there are three dominant inputs: input 8, input 12, and input 16. A reduced model of order 12 is needed for a good match, where four moments of the corresponding inputs are matched. The reduction results of PRIMA, SVDMOR, and DeMOR are compared in Fig. 7. The simulation time of reduced models of order 12 is 0.008 s.

In the previous case, the same reduced order $r = 12$ is used for different methods. Now, instead of the same reduced order, the same number of moments $m = 4$ is used for different methods. In this case, different methods will lead to reduced models of different orders. To match four moments, the reduced orders of PRIMA and DeMOR are 132 and 12. The maximum errors for PRIMA and DeMOR reduced models in stimulation are 0.002 and 0.013 V. Although the results for PRIMA are better, such small difference can be actually ignored. For SVDMOR, in order to get better results than DeMOR, a reduced model of order 80 is needed. In this case, the reduction times for PRIMA, SVDMOR, and DeMOR are 0.316, 0.223, and 0.086 s. The simulation times of reduced models of PRIMA, SVDMOR, and DeMOR are 1.765, 0.233, and 0.008 s.

In this example, as the original model has the special circuit formulation structure (26) and the reduced subsystems inherit such structure, the reduced subsystems are guaranteed to be passive and the overall reduced system remains passive as proved in Section IV-C.

C. Example 3

DeMOR is quite suitable for analyzing a number of nodes in a local region. We can perform the RGA analysis for all those nodes and find their dominant inputs. Typically, those nodes to be observed may share a very small number of dominant inputs, which is the case for power grid networks where input sources are not attached to every node to be observed.

To demonstrate this point, we use a power grid network with 10 000 nodes and 1000 distributed current sources. The current sources are used to model the typical switching activities of nonlinear devices, which are obtained in the standalone simulations.

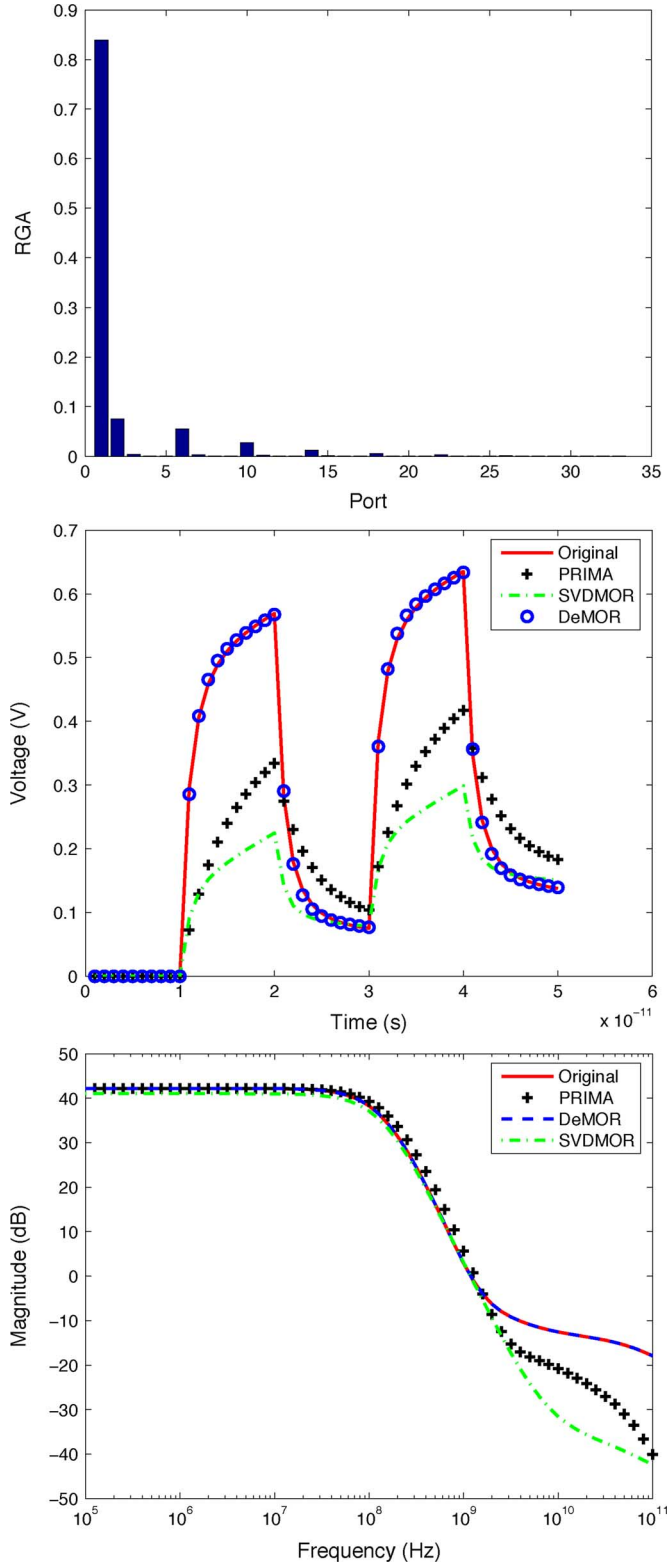


Fig. 6. (top) Relative gains, (middle) time-domain, and (bottom) frequency-domain responses at port 1 for an *RC* circuit.

The resistance and capacitance on the grid are on the order of Ω and pf, respectively.

Now we are interested in the transient responses for 500 nodes in a local region. As shown in Fig. 8, the distribution of those nodes in terms of dominant inputs are very concentrated, which

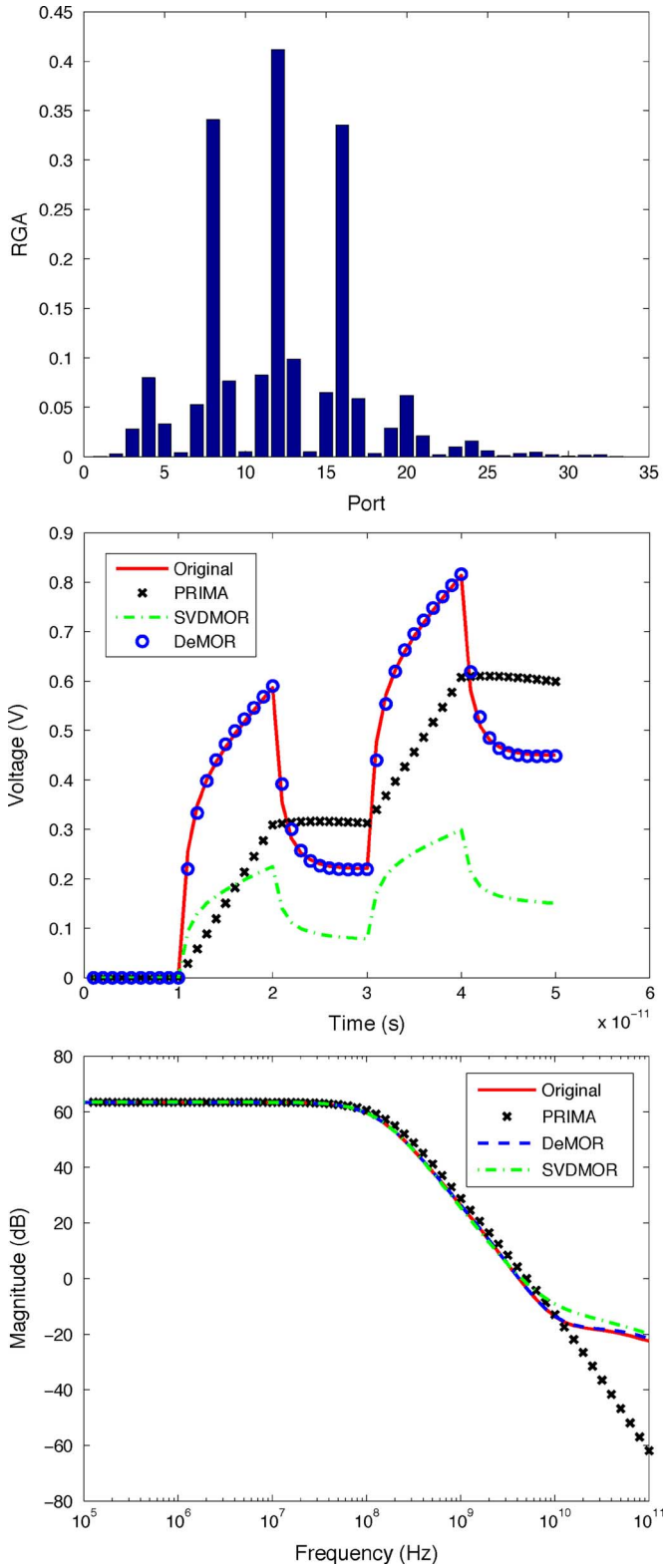


Fig. 7. (top) Relative gains, (middle) time-domain, and (bottom) frequency-domain responses at port 12 for an *RC* circuit.

means a small number of dominant inputs are shared by a large number of nodes. For each node, we choose the most dominant input. Since many inputs are shared, the redundant ones are eliminated. In this example, 25 representative inputs are identified and 2 moments are matched for each representative input,

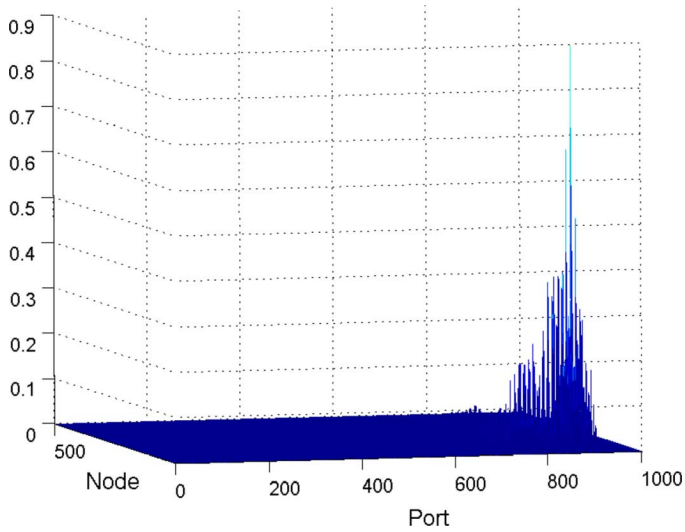


Fig. 8. Distribution of nodes in terms of dominant inputs in a local region.

which results in a reduced model of order 50. The transient responses of the 500 nodes, however, can be well approximated by the localized reduced model. Fig. 9 shows the transient responses at one of the nodes. Given the same reduced order 50, while DeMOR can match the original response well, there is still noticeable error for SVDMOR. In this example, the reduction times for DeMOR, SVDMOR, and PRIMA are 13.199, 12.856, and 68.241 s, respectively. The simulation time of reduced models of order 50 is 0.326 s.

In this example, as the voltages responses to be observed are not limited to the nodes with input sources, $\mathbf{L} \neq \mathbf{B}$ and thus the original model does not have the special circuit formulation structure (26), which means the reduced model cannot be guaranteed to be passive. However, as stability is still preserved in the reduction process [27], the simulation will be stable if the reduced network is driven by independent sources and not connected with any load, which is the typical setting for power grid simulation [30].

D. Example 4

Finally, we show the effectiveness of the modified balanced truncation method (mBT) for systems with a large number of inputs. As balanced truncation is very expensive for large systems, we use a smaller RC mesh with 400 nodes and 10 input sources to demonstrate the accuracy. We use the voltage response at one of the nodes as the output. In mBT, the principle components are obtained from two steps: first, we choose the dominant inputs based on RGA; then, we choose the dominant states in terms of the energy transfers from the dominant inputs based on the Hankel singular values.

As shown in Fig. 11, given one dominant input, the Hankel singular values in mBT are decaying much faster than those in BT: only 5 states are dominant in mBT while more than 20 states seem important in BT. Given the same reduced order 5, the performances of two methods are shown in Fig. 10: while mBT can match the original frequency response well, there is still noticeable error for BT. This verifies that we do not need to pay attention to each input equally. In this example, the reduction

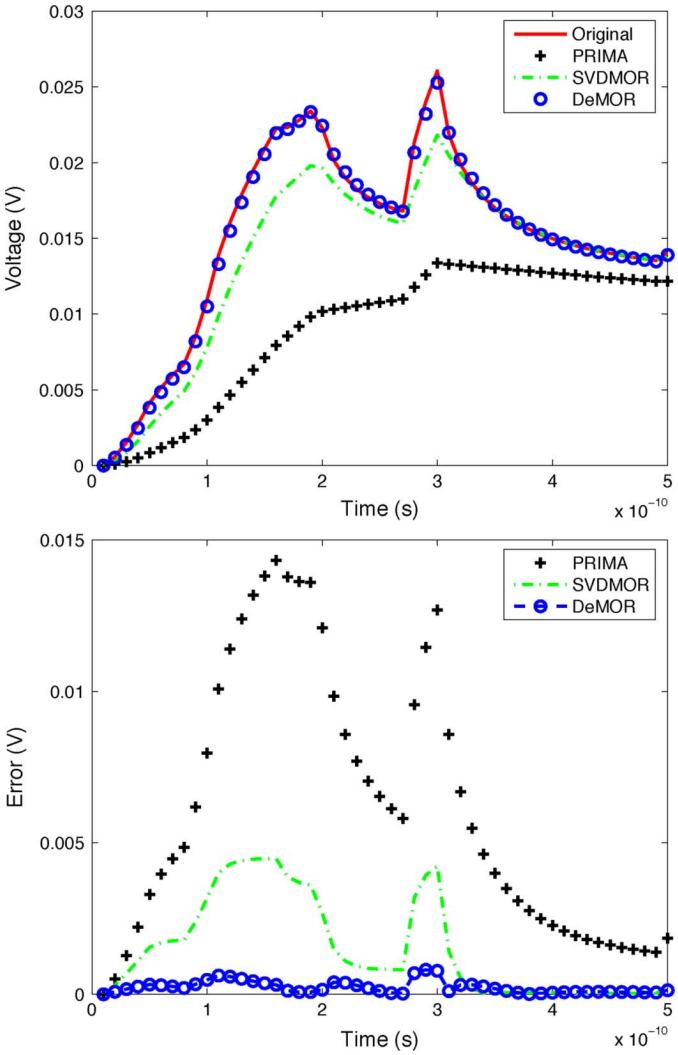


Fig. 9. Simulation results of the part of the grid.

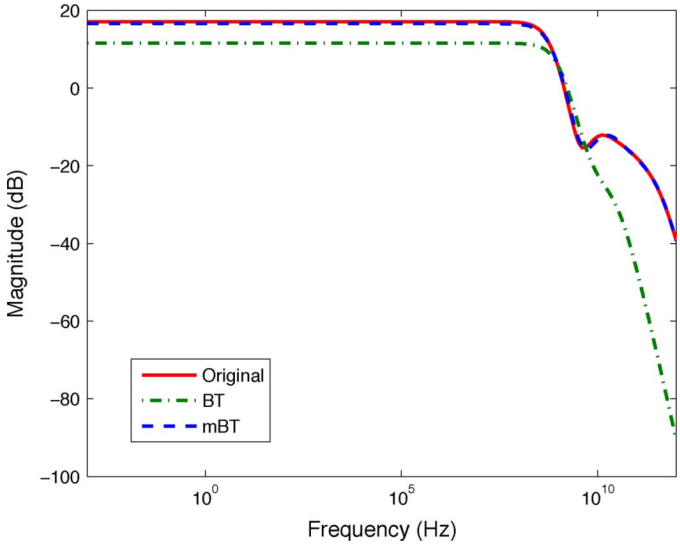


Fig. 10. Frequency-domain responses.

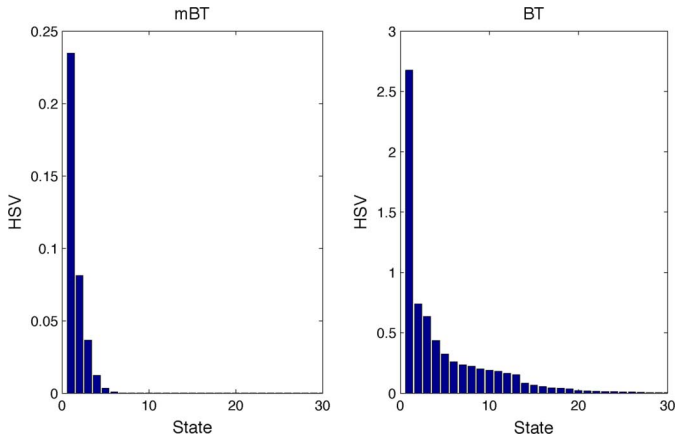


Fig. 11. Hankel singular values.

times of BT and mBT are 4.436 and 1.125 s, respectively. As discussed in Section V, the reduced model is not guaranteed to be passive.

VII. Conclusion

In this paper, we have proposed a novel method for the model order reduction of linear networks with many ports. The new method *DeMOR* adopts a *decentralized* reduction scheme, in which an MIMO system is decoupled into a number of MISO subsystems and each subsystem is reduced and simulated individually. The decouple process is carried out with the aid of the RGA, which measures the degree of interaction of each input-output pair. As a result, efficient passive reduction of each subsystem becomes possible and the whole reduced system can be proved to be passive. The DeMOR method is suitable for resistance-dominant interconnects like on-chip power grids and substrate planes, and can lead to extremely compact models for systems with massive ports compared to the traditional MOR methods. The simulation results show indeed that DeMOR is more advantageous than the existing approaches.

Acknowledgment

The authors would like to thank the reviewers for their helpful comments on this paper.

References

[1] A. C. Antoulas, "Approximation of large-scale dynamical systems," SIAM, Warrendale, PA, 2005.
[2] E. H. Bristol, "On a new measure of interaction for multivariable process control," *IEEE Trans. Autom. Control*, vol. AC-11, no. 1, pp. 133–134, Jan. 1966.
[3] H. Chen, C. Yeh, G. Wilke, S. Reddy, H. Nguyen, W. Walker, and R. Murgai, "A sliding window scheme for accurate clock mesh analysis," in *Proc. Int. Conf. Comput.-Aided Des. (ICCAD)*, 2005, pp. 939–946.
[4] J. Chen, J. S. Freudenberg, and C. N. Nett, "The role of the condition number and the relative gain array in robustness analysis," *Automatica*, vol. 30, pp. 1029–1035, 1994.
[5] E. Chiprout, "Fast flip-chip power grid analysis via locality and grid shells," in *Proc. Int. Conf. Comput.-Aided Des. (ICCAD)*, 2004, pp. 485–488.
[6] A. Conley and M. E. Salgado, "Gramian based interaction measure," in *Proc. IEEE Conf. Decision Control*, 2000, pp. 5020–5022.
[7] P. Feldmann, "Model order reduction techniques for linear systems with large numbers of terminals," in *Proc. Des., Autom. Test Eur. (DATE)*, 2004, pp. 944–947.
[8] P. Feldmann and R. W. Freund, "Efficient linear circuit analysis by pade approximation via the lanczos process," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 14, no. 5, pp. 639–649, May 1995.

[9] P. Feldmann and F. Liu, "Sparse and efficient reduced order modeling of linear subcircuits with large number of terminals," in *Proc. Int. Conf. Comput.-Aided Des. (ICCAD)*, 2004, pp. 88–92.
[10] R. W. Freund, "SPRIM: Structure-preserving reduced-order interconnect macromodeling," in *Proc. Int. Conf. Comput.-Aided Des. (ICCAD)*, 2004, pp. 80–87.
[11] E. J. Grimme, "Krylov projection methods for model reduction," Ph.D. dissertation, Elect. Comput. Eng. Dept., Univ. Illinois at Urbana-Champaign, Urbana-Champaign, 1997.
[12] A. J. Laub, M. T. Heath, C. C. Paige, and R. C. Ward, "Computation of system balancing transformations and other applications of simultaneous diagonalization algorithms," *IEEE Trans. Autom. Control*, vol. 32, no. 2, pp. 115–122, Feb. 1987.
[13] D. Li, S. X.-D. Tan, and B. McGaughy, "ETBR: Extended truncated balanced realization method for on-chip power grid network analysis," in *Proc. Des., Autom. Test Eur. (DATE)*, 2008, pp. 432–437.
[14] J. R. Li, "Model reduction of large linear systems via low rank system gramians," Ph.D. dissertation, Dept. Math., Massachusetts Institute of Technology (MIT), Cambridge, 2002.
[15] P. Li and W. Shi, "Model order reduction of linear networks with massive ports via frequency-dependent port packing," in *Proc. Des. Autom. Conf. (DAC)*, 2006, pp. 267–272.
[16] P. Liu, S. X.-D. Tan, H. Li, Z. Qi, J. Kong, B. McGaughy, and L. He, "An efficient method for terminal reduction of interconnect circuits considering delay variations," in *Proc. Int. Conf. Comput.-Aided Des. (ICCAD)*, 2005, pp. 821–826.
[17] B. Moore, "Principal component analysis in linear systems: Controllability, and observability, and model reduction," *IEEE Trans. Autom. Control*, vol. 26, no. 1, pp. 17–32, Jan. 1981.
[18] A. Odabasioglu, M. Celik, and L. Pileggi, "PRIMA: Passive reduced-order interconnect macromodeling algorithm," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 17, no. 8, pp. 645–654, Aug. 1998.
[19] H. Qian, S. R. Nassif, and S. S. Sapatnekar, "Random walks in a supply network," in *Proc. Des. Autom. Conf. (DAC)*, 2003, pp. 93–98.
[20] S. Pant and E. Chiprout, "Power grid physics and implications for CAD," in *Proc. Des. Autom. Conf. (DAC)*, 2006, pp. 199–204.
[21] X. Ye, P. Li, M. Zhao, R. Panda, and J. Hu, "Analysis of large clock meshes via harmonic-weighted model order reduction and port sliding," in *Proc. Int. Conf. Comput.-Aided Des. (ICCAD)*, 2007, pp. 627–631.
[22] J. R. Phillips, L. Daniel, and L. M. Silveira, "Guaranteed passive balancing transformation for model order reduction," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 22, no. 8, pp. 1027–1041, Aug. 2003.
[23] J. R. Phillips and L. M. Silveira, "Poor man's TBR: A simple model reduction scheme," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 24, no. 1, pp. 43–55, Jan. 2005.
[24] L. T. Pillage and R. A. Rohrer, "Asymptotic waveform evaluation for timing analysis," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 9, no. 4, pp. 352–366, Apr. 1990.
[25] Y. Shi, H. Yu, and L. He, "Samson: A generalized second-order arnoldi method for reducing multiple source linear network with susceptance," in *Proc. Int. Symp. Phys. Des. (ISPD)*, 2006, pp. 25–32.
[26] L. M. Silveira and J. R. Phillips, "Exploiting input information in a model reduction algorithm for massively coupled parasitic networks," in *Proc. Des. Autom. Conf. (DAC)*, 2004, pp. 385–388.
[27] M. Silveira, M. Kamon, I. Elfadel, and J. White, "A coordinate-transformed Arnoldi algorithm for generating guaranteed stable reduced-order models of RLC circuits," in *Proc. Int. Conf. Comput.-Aided Des. (ICCAD)*, 1996, pp. 288–294.
[28] Y. Su, J. Wang, X. Zeng, Z. Bai, C. Chiang, and D. Zhou, "SAPOR: Second-order Arnoldi method for passive order reduction of RCS circuits," in *Proc. Int. Conf. Comput.-Aided Des. (ICCAD)*, 2004, pp. 74–79.
[29] D. Vasilyev and J. White, "A more reliable reduction algorithm for behavioral model extraction," in *Proc. Int. Conf. Comput.-Aided Des. (ICCAD)*, 2005, pp. 813–820.
[30] J. M. Wang and T. V. Nguyen, "Extended Krylov subspace method for reduced order analysis of linear circuit with multiple sources," in *Proc. Des. Autom. Conf. (DAC)*, 2000, pp. 247–252.
[31] N. Wang and V. Balakrishnan, "Fast balanced stochastic truncation via a quadratic extension of the alternating direction implicit iteration," in *Proc. Int. Conf. Comput.-Aided Des. (ICCAD)*, 2005, pp. 801–805.
[32] B. Yan, S. X.-D. Tan, P. Liu, and B. McGaughy, "SBPOR: Second-order balanced truncation for passive model order reduction of RLC circuits," in *Proc. Des. Autom. Conf. (DAC)*, 2007, pp. 158–161.
[33] B. Yan, S. X.-D. Tan, and B. McGaughy, "Second-order balanced truncation for passive-model order reduction of RLCK circuits," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 55, no. 9, pp. 942–946, Sep. 2008.
[34] B. Yan, L. Zhou, S. X.-D. Tan, J. Chen, and B. McGaughy, "DeMOR: Decentralized model order reduction of linear networks with massive ports," in *Proc. Des. Autom. Conf. (DAC)*, 2008, pp. 409–414.
[35] B. Yan, S. X.-D. Tan, G. Chen, and L. W. , "Modeling and simulation for on-chip power grid networks by locally dominant Krylov subspace method," in *Proc. Int. Conf. Comput.-Aided Des. (ICCAD)*, 2008, pp. 744–749.
[36] B. Yan, S. X.-D. Tan, G. Chen, and Y. Cai, "Efficient model reduction of interconnects via double gramians approximation," in *Proc. Asia South Pacific Des. Autom. Conf. (ASP-DAC)*, 2010, pp. 25–30.
[37] B. Yan, S. X.-D. Tan, and J. Fan, "Passive rational interpolation based reduction via caratheodory extension for general systems," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 57, no. 9, pp. 750–755, Sep. 2010.
[38] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1994.

Boyuan Yan received the B.S. degree in electrical engineering from Dalian University of Technology, Dalian, China, in 2004, and the M.S. degree in electrical engineering, the M.S. degree in mathematics, and the Ph.D. degree in electrical engineering, from The University of California, Riverside, in 2007, 2008, and 2009, respectively.

He is currently a Postdoctoral Research Associate with the Department of Electrical and Computer Engineering, Texas A&M University, College Station. His research interests include modeling and simulation of large-scale circuits and dynamical systems, computer-aided design of VLSI circuits, computational neuroscience, and biomedical engineering. His current research involves biologically realistic modeling of human brain and computer-aided diagnosis and therapy of brain disorders.

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

He is a Professor with the Department of Electrical Engineering, University of California, Riverside (UCR). He is the Associate Director of Compute Engineering Program (CEN), Bourn College of Engineering, UCR. He also is a cooperative faculty member with the Department of Computer Science and Engineering, UCR. His research interests include statistical modeling, simulation, and optimization of mixed-signal/RF/analog circuits, fast thermal analysis and modeling for microprocessors and platform systems, parallel circuit simulation techniques based on GPU and multicore systems, and embedded system designs based on FPGA platforms. He also coauthored the books *Symbolic Analysis and Reduction of VLSI Circuits* (Springer/Kluwer, 2005) and *Advanced Model Order Reduction Techniques for VLSI Designs* (Cambridge University Press, 2007).

Dr. Tan is currently serving as an Associate Editor for three journals: *ACM Transaction on Design Automation of Electronic Systems (TODAE), Integration, The VLSI Journal, and Journal of VLSI Design*. He was a recipient of the Outstanding Oversea Investigator Award from the National Natural Science Foundation of China (NSFC) in 2008, the NSF CAREER Award in 2004, the Best Paper Award from 2007 IEEE International Conference on Computer Design (ICCD'07), two Best Paper Award Nominations from 2005 and 2009 IEEE/ACM Design Automation Conference, and the Best Paper Award from 1999 IEEE/ACM Design Automation Conference. He served as a technical program committee member for DAC, ICCAD, ASPDAC, ICCD, ISQED, BMAS, and ASICON.

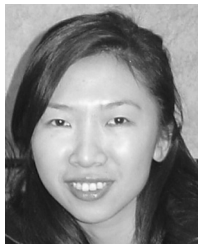
Lingfei Zhou received the B.S. degree in mechanical and electrical engineering from Xiamen University, China, in 2002, and the M.S. degree from the Department of Electrical Engineering, University of California, Riverside, in 2006.

He is with the Servo Engineering Department, Western Digital, Irvine, CA. His research interests include model order reduction and application on circuit simulation, position precision control, embedded real-time control system, control system design, optimization, and implementation on hard disk drive.

Jie Chen (F'07) received the B.S. degree in aerospace engineering from Northwestern Polytechnic University, Xian, China, in 1982, the M.S.E. degree in electrical engineering, the M.A. degree in mathematics, and the Ph.D. degree in electrical engineering from The University of Michigan, Ann Arbor, in 1985, 1987, and 1990, respectively.

He teaches in the field of systems and control, and signal processing. From 1990 to 1993, he was with the School of Aerospace Engineering and the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta. He joined the University of California, Riverside, in 1994, where he has been a Professor since 1999, and served as Chair for the Department of Electrical Engineering from 2001 to 2006. While on leave from the University of California, he currently holds the appointment of Chair Professor of Electronic Engineering at City University of Hong Kong, Hong Kong, China. He has also held a number of guest positions and visiting appointments with institutions in Australia, China, and Japan. His main research interests include the areas of linear multivariable systems theory, system identification, robust control, optimization, and networked control. He is the author of two books (with G. Gu) *Control-Oriented System Identification: An H-infinity Approach* (Wiley-Interscience, 2000), and (with K. Gu and V. L. Kharitonov) *Stability of Time-Delay Systems* (Birkhauser, 2003).

Dr. Chen is a Fellow of AAAS and a Yangtze Scholar/Chair Professor of China. He was a recipient of 1996 U.S. National Science Foundation CAREER Award, the 2004 SICE International Award, and the 2006 Natural Science Foundation of China Outstanding Overseas Young Scholar Award. He served on a number of journal editorial boards, as an Associate Editor and a Guest Editor for the IEEE TRANSACTIONS ON AUTOMATIC CONTROL, a Guest Editor for *IEEE Control Systems Magazine*, an Associate Editor for *Automatica*, *Journal of Control Theory and Applications*, and the founding Editor-in-Chief for the *Journal of Control Science and Engineering*.



Ruijing Shen (S'08) received the B.S. degree in electronic engineering and the M.S. degree from the Institute of Microelectronics, Tsinghua University, Beijing, China, in 2005 and 2007, respectively. She is currently pursuing the Ph.D. degree from the Department of Electrical Engineering, University of California, Riverside.

Her current research interests include voltage binning techniques for yield optimization, software thermal modeling for mobile computing system, statistical leakage analysis of VLSI, and variational capacitance extraction and modeling.