Fast Analysis of a Large-Scale Inductive Interconnect by Block-Structure-Preserved Macromodeling

Hao Yu, *Member, IEEE*, Chunta Chu, Yiyu Shi, *Student Member, IEEE*, David Smart, *Member, IEEE*, Lei He, *Senior Member, IEEE*, and Sheldon X.-D. Tan, *Senior Member, IEEE*

***Abstract*—To efficiently analyze the large-scale interconnect dominant circuits with inductive couplings (mutual inductances), this paper introduces a new state matrix, called VNA, to stamp inverse-inductance elements by replacing inductive-branch current with flux. The state matrix under VNA is diagonal-dominant, sparse, and passive. To further explore the sparsity and hierarchy at the block level, a new matrix-stretching method is introduced to reorder coupled fluxes into a decoupled state matrix with a bordered block diagonal (BBD) structure. A corresponding block-structure-preserved model-order reduction, called BVOR, is developed to preserve the sparsity and hierarchy of the BBD matrix at the block level. This enables us to efficiently build and simulate the macromodel within a SPICE-like circuit simulator. Experiments show that our method achieves up to 7× faster modeling building time, up to 33× faster simulation time, and as much as 67× smaller waveform error compared to SAPOR [a second-order reduction based on nodal analysis (NA)] and PACT (a first-order 2 × 2 structured reduction based on modified NA).**

***Index Terms*—Circuit simulation, high-speed interconnect model, model-order reduction.**

I. Introduction

INDUCTANCE is important in the analysis of high-performance interconnects [1]–[6]. For the system in a package (SiP) design, the chip–package codesign, and the analog/RF design beyond 60 GHz, there exists a huge amount of interconnects from package tracings, C4 bumps, on-chip power grids, buses, high-speed clock nets, and parasitics with a strong electromagnetic coupling. To ensure signal and power integrity, a complete $\mathbf{RLC}$ model from all interconnects needs to be extracted from the layout, which leads to a network with the large-scale complexity. Model-order reduction is one technique to reduce the complexity. Most previous model-order reduction approaches [7]–[16], however, ignored the structure information of state matrices. Utilizing the tree structure, rapid interconnect circuit evaluator (RICE) [17] provided a fast macromodeling for $\mathbf{RC}$ trees by tree traversal. However, the show stopper for many fast simulation algorithms like RICE arises from mutual inductors. As magnetic coupling is a long-range effect, it makes each pair of branch currents coupled and destroys the structure benefit from trees. As such, there is a need to develop a fast model-order reduction algorithm that can explore the genetic matrix structure such as sparsity, hierarchy, and latency. The first step is the sparsification of the mutual inductance.

An inductor is defined with respect to a loop current. The switching from multiple adjacent currents causes difficulties when attempting to exactly specify a return path that composes the loop. The partial-element equivalent circuit (PEEC) model [18] describes one partial inductance with one piece of branch current through the metal and assumes its induced current at infinity. This results in a modified nodal analysis (MNA) [19], [20] with the state variable of a branch-inductive current, describing the partial inductance. In the PEEC model, each partial inductance has coupling (mutual inductance) with each other, and hence, the state matrix becomes dense with a large number of fill-ins.

An efficient model-order reduction thereby needs to first sparsify the dense inductance matrix. Unlike the capacitance matrix, since the magnetic coupling is determined by the long-range vector potential [21], [22], the extracted inductance matrix $\mathbf{L}$ is not diagonal-dominant. Sparsifying $\mathbf{L}$ directly thereby leads to an instable state matrix. Recently, people found that $\mathbf{L}^{-1}$ matrix is more diagonal-dominant [2], [6], [16] than $\mathbf{L}$, and hence, $\mathbf{L}^{-1}$ can be more effectively sparsified while preserving stability. Existing model-order reductions [7]–[10], [12], [14]–[16] directly worked on the dense inductance matrix $\mathbf{L}$. However, they become inefficient if $\mathbf{L}$ is not stably sparsified. On the other hand, there is no method similar to the MNA that can passively stamp $\mathbf{L}^{-1}$ into the state matrix. A nonpassive stamping results in a nonpassive reduction [10], [12], and the reduced model cannot reproduce a stable simulation result. To passively reduce the linear circuit containing $\mathbf{L}^{-1}$, ENOR [11] stamped the *nodal susceptance* ($\Gamma = \mathbf{E}_l\mathbf{L}^{-1}\mathbf{E}_l^T$, and $\mathbf{E}_l$ is the incident matrix for inductance). This results in a passive second-order system described by the nodal analysis (NA) matrix. SAPOR further improved the accuracy of orthonormalization by a second-order Arnoldi method [13]. Since most circuit simulators such as SPICE [20] assume a

Manuscript received November 23, 2008; revised March 25, 2009 and May 14, 2009. First published September 15, 2009; current version published September 24, 2010. This work was supported in part by the National Science Foundation under CCR-0401682, by SRC project 1100.001, and by a University of California-MICRO grant sponsored by Analog Devices, Intel, and Mindspeed. This paper was presented in part at the 2005 IEEE International Behavioral Modeling and Simulation Conference and the 2006 International Conference on Computer-Aided Design.

H. Yu was with Berkeley Design Automation, Santa Clara, CA 95054 USA. He is now with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore (e-mail: haoyu@ieee.org).

C. Chu is with Apache Design Solutions, San Jose, CA 95134 USA (e-mail: chunta@ucla.edu).

Y. Shi and L. He are with the Department of Electrical Engineering, University of California, Los Angeles, CA 90095 USA (e-mail: yshi@ee.ucla.edu; lhe@ee.ucla.edu).

D. Smart is with Analog Devices, Inc., Wilmington, MA 01887 USA (e-mail: david.smart@analog.com).

S. X.-D. Tan is with the Department of Electrical Engineering, University of California–Riverside, Riverside, CA 92521 USA (e-mail: stan@ee.ucr.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVLSI.2009.2024343

first-order MNA during the numerical transient integration, a reduced macromodel in the second-order form by ENOR or SAPOR cannot be reconnected with the remaining nonlinear devices into SPICE.

After sparsifying the long-range magnetic coupling, the resulting loosely coupled $\mathbf{RLC}$ network can be partitioned further to take the advantage of sparsity and hierarchy at the block level. With the use of a split congruence transformation, PACT first explored the 2×2 matrix structure of MNA [9]. The nodal voltage and branch current of the reduced system can be solved separately. SPRIM [14] further related the reciprocity of the second-order system to the first-order system. Using the split congruence transformation, the reduced system preserves the reciprocity, and hence, can match two times as many moments as PRIMA [10]. To further explore the hierarchy at the block level, a genetic block-structure-preserving macromodeling (BSMOR) is introduced in [23] to preserve the block-level sparsity and hierarchy during the reduction. However, the consideration of magnetic couplings between blocks is still unknown.

For large-scale inductive interconnects, this paper explores the block-level sparsity and hierarchy that can reduce computational cost during model-order reduction when generating and simulating a macromodel. Its contributions are twofold. First, using a new state variable of flux instead of the inductive-branch current to stamp $\mathbf{L}^{-1}$ elements, we introduce a new modified nodal analysis, called VNA, in this paper. Compared to the traditional MNA with dense $\mathbf{L}$, our VNA with $\mathbf{L}^{-1}$ has a better diagonal-dominance. As such, VNA is not only easier when exploring the sparsity, but is also formulated to preserve passivity. Compared to the traditional NA state matrix, our VNA state matrix is in the first-order form. As a result, the VNA state matrix is definite at dc and can be easily integrated with other devices into SPICE.

To utilize the sparsity and hierarchy of the state matrix at the block level, we further introduce a matrix-stretching method to build a state matrix with the bordered block diagonal (BBD) structure. Compared to the previous works [23]–[25], our BBD decomposition is the first work to consider magnetic couplings between blocks. The BBD-structure-based solver further reduces factorization cost during reduction. We also develop a generalized split congruence transformation to preserve the BBD structure, called BVOR, in this paper. Compared to the previous structured model-order reduction [9], [14], BVOR aims at preserving more fine-grained structures of the state matrix. The reduced system preserves not only the reciprocity and passivity similar to PACT and SPRIM, but also the sparsity and hierarchy at the block level. As a result, it can be rapidly factorized by a hybrid dense and sparse matrix solver. Experimental results show that compared to SAPOR, the second-order reduction method based on NA, our approach is up to $7\times$ faster to build, $33\times$ faster to analyze, and has up to $67\times$ smaller waveform error. In addition, compared to PACT, the first-order reduction method based on MNA with a 2×2 split congruence transformation, our approach is up to $5\times$ faster to build and $29\times$ faster to analyze.

The rest of this paper is organized as follows. In Section II, we present the background. In Section III, we derive a new stamping of $\mathbf{L}^{-1}$ elements, called VNA, and its corresponding reduction method, called VOR. In Section IV, we discuss a matrix-stretching method to build the BBD structure and present a BBD solver. In Section V, we introduce a BBD-structured model-order reduction, called BVOR. We present experimental results in Section VI and conclude in Section VII.

II. Background

A. Inductive Interconnect in State Space

An extracted $\mathbf{RLC}$ network for an interconnect can be described in the state space as follows. We assume that the interconnect is driven by the external voltage source $u(t)$ as the input and is probed by the voltage source $y(t)$ as the output. Next, we discuss how to compose the state variable $x(t)$. One obvious choice is to use the nodal voltage v_n at the terminals of each device. However, for an inductance or an external voltage source, the branch currents $\mathbf{i}_l$ and $\mathbf{i}_i$ have to be used as the state variables. For an inductor with two terminal nodal voltages $\mathbf{v}_1$ and $\mathbf{v}_2$, it is shorted at dc, i.e., $\mathbf{v}_1 - \mathbf{v}_2 = 0$. As such, the two state variables $\mathbf{v}_1$ and $\mathbf{v}_2$ are no longer independent, and a new state variable is needed by a branch current $\mathbf{i}_{1,2}$ through the inductor. Similarly, for an external voltage source u, its terminal voltages are not independent since $u = \mathbf{v}_1 - \mathbf{v}_2$, and hence, also needs to be described by a branch current $\mathbf{i}_{1,2}$.

Therefore, in frequency domain (s), with the use of the state variable $x(s)$,

$$x(s) = \begin{bmatrix} v_n \\ \mathbf{i}_l \\ \mathbf{i}_i \end{bmatrix}$$

a first-order state equation can be derived by

$$(\mathcal{G} + s\mathcal{C})x(s) = \mathcal{B}u(s) \qquad y(s) = \mathcal{B}^T x(s) \tag{1}$$

based on the Kirchhoff's current law/Kirchhoff's voltage law (KCL/KVL) and branch equations, where

$$\mathcal{B} = \begin{bmatrix} 0 \\ 0 \\ -B \end{bmatrix}$$

and

$$\mathcal{G} = \begin{bmatrix} G & \mathbf{E}_l & \mathbf{E}_i \\ -\mathbf{E}_l^T & 0 & 0 \\ -\mathbf{E}_i^T & 0 & 0 \end{bmatrix} \qquad \mathcal{C} = \begin{bmatrix} C & 0 & 0 \\ 0 & \mathbf{L} & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{2}$$

Here, $\mathcal{G}$ and $\mathcal{C} \in R^{N\times N}$ $(N = n_v + n_i)$ and $\mathcal{B} \in R^{N\times n_p}$. n_v is the number of nodal voltage variables, n_i is the number of branch current variables, and n_p is the number of voltage sources at the ports. Note that $\mathbf{E}_l$ and $\mathbf{E}_i$ are incident matrices that describe the relation between nodal variables and branch variables. B is the adjacent matrix that describes how to select n_p inputs/outputs from N variables. This formulation is called an MNA [19], [20].

One important property of the aforementioned MNA formulation is that the transposed summations of state matrices are *symmetric* and *semipositive-definite*, i.e., $\mathcal{G} + \mathcal{G}^T \succ 0$ and $\mathcal{C} + \mathcal{C}^T \succ 0$. As discussed in [10], this is one of the sufficient

conditions for a system to be passive. However, if $\mathbf{L}^{-1}$ is directly stamped by the inductive-branch current, $\mathcal{G}$ and $\mathcal{C}$ become

$$\mathcal{G} = \begin{bmatrix} G & \mathbf{E}_l & \mathbf{E}_i \\ -\mathbf{L}^{-1}\mathbf{E}_l^T & 0 & 0 \\ -\mathbf{E}_i^T & 0 & 0 \end{bmatrix} \qquad \mathcal{C} = \begin{bmatrix} C & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix}. \quad (3)$$

The difference between (2) and (3) is that the state matrix $\mathcal{G}$ in (3) does not satisfy the requirement for passivity.

To symmetrically stamp $\mathbf{L}^{-1}$ into a passive state matrix, a second-order system [11], [13] by NA can be used

$$(sC + G + \Gamma/s)\,x(s) = \mathbf{I}(s) \qquad y(s) = \mathbf{E}_i^T x(s). \quad (4)$$

The state variable x only has the nodal voltage $x(s) = v_n$, where the inductive-branch current $\mathbf{i}_l$ is eliminated as an intermediate variable. Moreover, the external source $\mathbf{I}$ relates the current flowing at terminals by $\mathbf{I}(s) = \mathbf{E}_i(-\mathbf{i}_i)$. In addition, the *nodal susceptance* Γ is $\Gamma = \mathbf{E}_l\mathbf{L}^{-1}\mathbf{E}_l^T$ for inductance. It is observed in [11] and [13] that Γ is symmetric and positive-definite, and that the intrinsic form in NA [see (4)] has both passivity and reciprocity. Therefore, with the use of a second-order Arnoldi orthonormalization, (4) can be reduced by a flat projection matrix with the preserved passivity and reciprocity.

However, compared to the MNA, inductances, independent voltage sources, and some of the controlled sources cannot be properly described by NA in the second-order form. For example, since an inductor is shorted at dc, the dc solution is indefinite as the NA state matrix is indefinite. Next, as shown by (4), the state matrix in the NA is denser than the one in the MNA, and hence, its factorization cost is increased. More importantly, all existing circuit simulators such as SPICE [20] assume a first-order differential algebraic equation (DAE) with a corresponding local truncation error (LTE) control since the numerical transient integration for a second-order DAE is much more complicated than the first-order DAE. As such, in order to enjoy the sparsity of $\mathbf{L}^{-1}$ in SPICE, we need to find a first-order stamping that is also passive.

B. Model-Order Reduction

To take into account the skin effect and for the sake of quasi-static extraction, the long and thick metal needs to be first discretized into filaments. This may lead to a large-dimensioned state space by (1). Model-order reduction is one approach to reduce the dimension while preserving accuracy in terms of moments.

Solving (1) results in a transfer function

$$\frac{y(s)}{u(s)} = H(s) = \mathcal{B}^T(\mathcal{G} + s\mathcal{C})^{-1}\mathcal{B}. \quad (5)$$

By expanding $H(s)$ at some frequency point s_0 ($s = s_0 + \sigma$), it can be easily verified that the state variable is contained in a block Krylov subspace

$$\text{span}\{\mathcal{R}, \mathcal{A}\mathcal{R}, \ldots, \mathcal{A}^{q-1}\mathcal{R}, \ldots\}$$

with two moment generation matrices

$$\mathcal{A} = (\mathcal{G} + s_0\mathcal{C})^{-1}\mathcal{C} \qquad \mathcal{R} = (\mathcal{G} + s_0\mathcal{C})^{-1}\mathcal{B}.$$

The work in [26] relates to the Krylov subspace and the model-order reduction by the projection as follows.

Lemma 1: If a small-dimensioned matrix Q ($\in R^{N\times n} n \ll N$) that spans the qth-order ($q = \lfloor n/n_p \rfloor$) block Krylov subspace is

$$\mathcal{K}(\mathcal{A}, \mathcal{R}, q) = \text{span}\{\mathcal{R}, \mathcal{A}\mathcal{R}, \ldots \mathcal{A}^{q-1}\mathcal{R}\} \subseteq \text{span}\{Q\}$$

then applying Q to project the original system

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

the first q block moments of the reduced system $\hat{H}(s)$

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

expanded at s_0 are identical to the original one $H(s)$.

The projection-based model-order reduction is essentially to construct an invariant subspace that can approximate the dominant system response in terms of the first few moments expanded at s_0. The order q determines the accuracy of the reduced model and depends on both N and n_p. A detailed analysis on how to select q can be found in [15] and [27]. Note that PRIMA [10] applies a block Arnoldi method to construct an orthonormalized projection matrix Q.

In addition, as shown by PACT [9] and PRIMA [10], we state the following lemma.

Lemma 2: When the input and the output are symmetric, the reduced model $\hat{H}(s)$ is passive when projected by an orthonormalized Q in a fashion similar to that of the congruence transformation (6).

For the preservation of passivity, this paper makes the same assumptions for the symmetric inputs and outputs.

Moreover, instead of projecting a flat matrix, PACT [9] and SPRIM [14] further split the projection matrix Q into a 2×2 block form

$$Q = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} \rightarrow \mathcal{Q} = \begin{bmatrix} Q_1 & 0 \\ 0 & Q_2 \end{bmatrix}$$

according to the size of n_v and n_i. Such a structured matrix $\mathcal{Q}$ leads to a projection that preserves the 2×2 MNA structure and the symmetry of input and output (or reciprocity). Therefore, the nodal voltage and branch current can be solved separately, and the number of matched moments is improved in PACT and SPRIM by two times (input moments + output moments) when compared to PRIMA.

III. VNA Stamping of Inverse Inductance

Though it is well known that a second-order system can be transformed into a first-order system, finding one that preserves passivity and reciprocity is not easy. In this section, we show that by using a new state variable, flux Φ, to replace an inductive-branch current i_l, we can form a first-order stamping of $\mathbf{L}^{-1}$

that preserves passivity and reciprocity and leads to a passive reduction similar to PRIMA.

A. Derivation of VNA

The fundamental electromagnetics of inductance [21], [22] shows the following relation between a magnetic flux Φ, a branch inductance $\mathbf{L}$, and an inductive-branch current i_l:

$$\Phi = \mathbf{L}\mathbf{i}_l. \tag{7}$$

Recall that in a PEEC model, the inductive-branch current i_l composes of a loop with the returned current at infinity. Hence, both Φ and $\mathbf{L}$ are defined based on such a loop.

Note that the branch voltage drop v_l at an inductor can be calculated as $\mathbf{v}_l = s\mathbf{L}\mathbf{i}_l$, or can be selected from the nodal voltage v_n by $\mathbf{v}_l = \mathbf{E}_l^T v_n$. As such, the following relation between the magnetic flux Φ (branch variable) and the nodal voltage v_n can be derived by

$$s\Phi = \mathbf{E}_l^T v_n. \tag{8}$$

Furthermore, the MNA (1) leads to the following two independent equations:

$$s\mathbf{i}_l = \mathbf{L}^{-1}\mathbf{E}_l^T v_n \tag{9}$$

and

$$(G + sC)v_n + \mathbf{E}_l\mathbf{i}_l + \mathbf{E}_i\mathbf{i}_i = 0. \tag{10}$$

According to (8), (9) becomes

$$\mathbf{i}_l = \mathbf{L}^{-1}\Phi \tag{11}$$

and (10) becomes

$$(G + sC)v_n + \mathbf{E}_l(\mathbf{L}^{-1}\Phi) + \mathbf{E}_i\mathbf{i}_i = 0. \tag{12}$$

As a result, a new MNA equation with a first-order admittance can be obtained

$$\mathcal{G}x(s) + s\mathcal{C}x(s) = \mathcal{B}u(s) \qquad y(s) = \mathcal{B}^T x(s) \tag{13}$$

based on (8) and (12), where

$$x = \begin{bmatrix} v_n \\ \Phi \\ \mathbf{i}_i \end{bmatrix}$$

is a new vector of state variables composed by the nodal voltage, flux, and branch current for the external voltage source. The new state matrices $\mathcal{G}$ and $\mathcal{C}$ become

$$\mathcal{G} = \begin{bmatrix} G & (\mathbf{E}_l\mathbf{L}^{-1}) & \mathbf{E}_i \\ -\left(\mathbf{L}^{-1}\mathbf{E}_l^T\right) & 0 & 0 \\ -\mathbf{E}_i^T & 0 & 0 \end{bmatrix}$$
$$\mathcal{C} = \begin{bmatrix} C & 0 & 0 \\ 0 & \mathbf{L}^{-1} & 0 \\ 0 & 0 & 0 \end{bmatrix}$$
$$\mathcal{B} = \begin{bmatrix} 0 \\ 0 \\ -B \end{bmatrix}. \tag{14}$$

We call such a new "MNA" the VNA in this paper.

Since the aforementioned derivation is just to replace the state variable of $\mathbf{i}_l$ by Φ, the resulting MNA has the same structure as the original MNA, both preserving passivity and reciprocity when compared to the NA stamping in (4). The new VNA stamping results in a passive formulation of $\mathbf{L}^{-1}$ because both

$$\mathcal{G} + \mathcal{G}^T = \begin{bmatrix} 2G & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad \mathcal{C} + \mathcal{C}^T = \begin{bmatrix} 2C & 0 & 0 \\ 0 & 2\mathbf{L}^{-1} & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{15}$$

are symmetric and semipositive-definite. Therefore, a passive model-order reduction can be performed similarly to PRIMA by finding a projection matrix Q from the block Arnoldi method. In addition, according to Lemma 2, the projection by Q leads to a reduced model with preserved passivity when the input and the output are symmetric. By further creating a 2×2 structured projection matrix $\mathcal{Q}$ similarly to PACT and SPRIM, the reduced model further preserves the reciprocity and matches twice as many moments as does PRIMA. We call the reduction based on the VNA the VOR in this paper.

B. Advantages of VNA

The advantages of VNA are further illustrated. First, let us compare the VNA with the MNA. Since the VNA with $\mathbf{L}^{-1}$ has a better diagonal-dominance than the MNA with $\mathbf{L}$, it results in a better pivoting to reduce fill-ins than stamping $\mathbf{L}$ directly. More importantly, $\mathbf{L}^{-1}$ can also be stably sparsified by applying truncation or windowing approaches [2], [6], [12]. When stamping a sparsified $\mathbf{L}^{-1}$ and a sparse $\mathbf{E}_l$, their product $\mathbf{E}_l\mathbf{L}^{-1}$ is sparse as well. As such, though two off-diagonal blocks $(\mathbf{E}_l\mathbf{L}^{-1})$ introduce new fill-ins, the additional fill-ins are still much smaller than the fill-ins of a dense $\mathbf{L}$.

Next, different from the second-order NA, since VNA is still in the first-order form, both the original VNA and the reduced VNA can be reconnected with the nonlinear devices in SPICE for the numerical transient simulation. Therefore, the VNA stamping within a SPICE-like simulator can dramatically reduce the factorization cost for inductive interconnect dominant circuits. In addition, as the VNA is definite at dc, the expansion point s_0 can be selected at dc $(s_0 = 0)$. As such, tree-traversal algorithms such as RICE [17] can be extended to handle the inductance-sparsified $\mathbf{RLC}$ trees.

However, to handle most nontree interconnects such as buses and meshes, we need to explore the sparsity and hierarchy of state matrices at the block level. As such, we further introduce a BBD-structured state matrix and its block-structure-preserved model-order reduction.

IV. Bordered Block Diagonal Matrix

One efficient solution for a large-scale $\mathbf{RLC}$ network is to apply the network decomposition [23]–[25], which usually results in a state matrix with a BBD structure. As shown in Fig. 1, a network is partitioned into a few blocks, and each block may have couplings to a top-level global block. This is different from the preordering algorithm such as the nested dissection (NS) [25]. The permutation in NS may destroy the passivity of the state matrix.

However, due to the magnetic coupling, it couples each pair of branch currents, and hence, limits the use of BBD to only the $\mathbf{RC}$ network. In this section, after stably sparsifying the

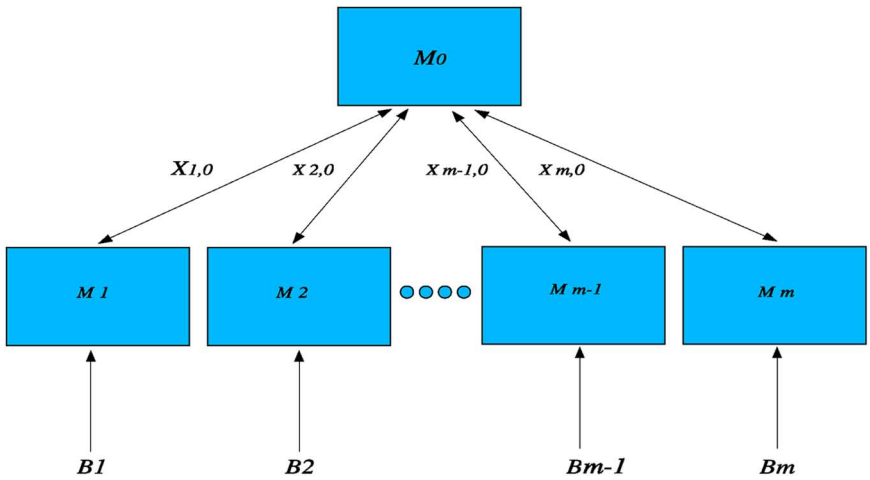

Fig. 1. Two-level BBD rpresentation of a flat VNA circuit. There is no coupling in bottom blocks, but each bottom block is connected to a centric interconnection block.

mutual inductance, we introduce a matrix-stretching method to construct a passive BBD-structured matrix.

A. BBD Matrix Stretching

A network decomposition is usually achieved by *node splitting* and *branch tearing* for nodal voltage and branch current variable, respectively. Though the concepts of node splitting and branch tearing are not new [23]–[25], the previous methods cannot deal with magnetic couplings. The BBD matrix stretching developed in this paper basically introduces new columns/rows for decoupled nodes or branches. It provides a way to tear the state matrix with sparse inverse inductance in the frame work of VNA, which is not discussed before.

1) Stretching of Nodal Voltages: A matrix stretching of nodal voltages can be described by the following rule.

> *Rule 1*: Assume that two resistive (or capacitive) branches b_i and b_j are coupled at a common node v_2 with a conductor g_x (or a capacitor c_x). Branch b_i has nodal voltages (v_1, v_2) and b_j has nodal voltages (v_2, v_3). They can be decoupled by introducing: 1) a duplicated state variable v_2' with $v_2 = v_2'$ and 2) an auxiliary state variable $i_{2,2'}$ for a new branch current between nodes v_2 and v_2'.

The corresponding transformations from the old $\mathcal{G}$ and $\mathcal{C}$ to the new G and C are

$$\mathcal{G}: \begin{array}{c|ccc} & v_1 & v_2 & v_3 \\ \hline v_1 & g_0+g_x & -g_x & \\ v_2 & -g_x & g_0+2g_x & -g_x \\ v_3 & & -g_x & g_0+g_x \end{array} \rightarrow$$

$$G: \begin{array}{c|ccccc} & v_1 & v_2 & v_2' & v_3 & i_{2,2'} \\ \hline v_1 & g_0+g_x & -g_x & & & \\ v_2 & -g_x & 2g_0+g_x & & & +1 \\ v_2' & & & 2g_0+g_x & -g_x & -1 \\ v_3 & & & -g_x & g_0+g_x & \\ i_{2,2'} & & -1 & +1 & & \end{array} \tag{16}$$

and

$$\mathcal{C}: \begin{array}{c|ccc} & v_1 & v_2 & v_3 \\ \hline v_1 & c_0+c_x & -c_x & \\ v_2 & -c_x & c_0+2c_x & -c_x \\ v_3 & & -c_x & c_0+c_x \end{array} \rightarrow$$

$$C: \begin{array}{c|ccccc} & v_1 & v_2 & v_2' & v_3 & i_{2,2'} \\ \hline v_1 & c_0+c_x & -c_x & & & \\ v_2 & -c_x & 2c_0+c_x & & & +1 \\ v_2' & & & 2c_0+c_x & -c_x & -1 \\ v_3 & & & -c_x & c_0+c_x & \\ i_{2,2'} & & -1 & +1 & & \end{array} \tag{17}$$

respectively, where $g_{0\mathrm{S}}/c_{0\mathrm{S}}$ are the self-conductance/capacitance at each node.

Since the resistor and the capacitor are represented by nodal voltage, such a node splitting can be efficiently applied to decouple the $\mathbf{RC}$ network. The node splitting, however, cannot handle inductance or its inverse element because inductance is described by branch current. It is quite possible that two branch currents at two partitioned blocks are still coupled by mutual inductance.

2) Stretching of Branch Currents: We call the entries of $\mathbf{L}^{-1}$ the *susceptor*, which includes the self-susceptor k_0 and the mutual susceptor k_x. To cleanly decouple the inductive couplings between two partitioned blocks, a matrix stretching of inductive-branch currents can be described by the following rule.

> *Rule 2*: Assume that two inductive branches b_i and b_j are coupled by a mutual susceptor k_x. Branch b_i has nodal voltages (v_1, v_2) and b_j has nodal voltages (v_3, v_4). Moreover, b_i has a branch flux ϕ_i and b_j has a branch flux ϕ_j. These can be decoupled by introducing an auxiliary state variable ϕ_{ij} that describes the flux difference by $\phi_{ij} = \phi_i - \phi_j$.

The transformations from the old $\mathcal{G}$ and $\mathcal{C}$ to the new G and C are

$$\mathcal{G}: \begin{bmatrix} & v_1 & v_2 & \phi_i & v_3 & v_4 & \phi_j \\ v_1 & \star & \star & k_0 & \star & \star & k_x \\ v_2 & \star & \star & -k_0 & \star & \star & -k_x \\ \phi_i & -k_0 & k_0 & 0 & -k_x & k_x & 0 \\ v_3 & \star & \star & k_x & \star & \star & k_0 \\ v_4 & \star & \star & -k_x & \star & \star & -k_0 \\ \phi_j & -k_x & k_x & 0 & -k_0 & k_0 & 0 \end{bmatrix} \rightarrow$$

$$G: \begin{bmatrix} & v_1 & v_2 & \phi_i & v_3 & v_4 & \phi_j & \phi_{ij} \\ v_1 & \star & \star & k_0' & & & & -k_x \\ v_2 & \star & \star & -k_0' & & & & k_x \\ \phi_i & -k_0' & k_0' & 0 & & & & 0 \\ v_3 & & & & \star & \star & k_0' & k_x \\ v_4 & & & & \star & \star & -k_0' & -k_x \\ \phi_j & & & & -k_0' & k_0' & 0 & 0 \\ \phi_{ij} & k_x & -k_x & 0 & -k_x & k_x & 0 & 0 \end{bmatrix} \tag{18}$$

and

$$\mathcal{C}: \begin{bmatrix} & v_1 & v_2 & \phi_i & v_3 & v_4 & \phi_j \\ v_1 & \star & \star & 0 & \star & \star & 0 \\ v_2 & \star & \star & 0 & \star & \star & 0 \\ \phi_i & 0 & 0 & k_0 & 0 & 0 & k_x \\ v_3 & \star & \star & 0 & \star & \star & 0 \\ v_4 & \star & \star & 0 & \star & \star & 0 \\ \phi_j & 0 & 0 & k_x & 0 & 0 & k_0 \end{bmatrix} \rightarrow$$

$$C: \begin{bmatrix} & v_1 & v_2^i & \phi_i & v_3 & v_4 & \phi_j & \phi_{ij} \\ v_1 & \star & \star & 0 & & & & 0 \\ v_2 & \star & \star & 0 & & & & 0 \\ \phi_i & 0 & 0 & k_0' & & & & 0 \\ v_3 & & & & \star & \star & 0 & 0 \\ v_4 & & & & \star & \star & 0 & 0 \\ \phi_j & & & & 0 & 0 & k_0' & 0 \\ \phi_{ij} & 0 & 0 & 0 & 0 & 0 & 0 & -k_x \end{bmatrix} \tag{19}$$

respectively, where $k_0' = k_0 + k_x$. Note that Rule 2 obtains an equivalent solution by finding a summed equivalent state matrix

$$\mathcal{G} + s\mathcal{C} \rightarrow G + sC.$$

This is due to the following node–branch relations in (8):

$$v_1^i - v_2^i = s \cdot \phi_i \qquad v_1^j - v_2^j = s \cdot \phi_j.$$

3) BBD Partition: Under these two BBD-transformation rules, it is easy to verify that the resulting BBD state matrices are equivalent to the original and are passive. In the following, we further illustrate the steps to build the BBD-structured state matrix with partition.

The first step is to partition the network at dc. In this step, the flat VNA circuit at dc is first mapped into a hypergraph. In one hypergraph, the nodes represent the nodal voltage and the branches represent the resistive path for the current; the capacitive path is open and the (inverse) inductive path is short at dc. In this paper, we assume that each branch has uniform weight, and apply a *min-cut* multilevel algorithm hmetis [28] to partition the hypergraph into user-specified m blocks. Here, the partition needs to be cautious about the port nodes with external voltage sources. To guarantee that each block has a dc path to ground, we prespecified a uniform distribution of ports for m blocks, where each block has at least one port with an external voltage source to the ground. A more advanced method based on spectral clustering can be applied to obtain the distribution of ports by studying their correlation [29]. After the partition, each block has a size of n_i ($\sum_{i=1}^{m} n_i = N$) and has n_{p_i} ($\sum_{i=1}^{m} n_{p_i} = n_p$) external ports.

Next, the node splitting is applied according to (16) and (17) to split the connected resistive or capacitive branches between two coupled blocks, while the branch tearing is applied according to (18) and (19) to tear the coupled inductive branches between two coupled blocks. The resulting m blocks with no coupling in between are at the bottom level of the BBD, and are represented by $\mathbf{M}_i$ ($i = 1, \ldots, m$). The top level is the global interconnection block represented by $\mathbf{M}_0$, which is connected with one diagonal block $\mathbf{M}_i$ ($i \neq 0$) by the corresponding connection matrix $X_{i,0}$. The interconnection block has size n_0 and contains all coupling branches between any pair of blocks at the bottom level.

Precisely, the resulting system equation is

$$(G + sC)\mathrm{x}(s) = Bu(s) \tag{20}$$

where

$$G = \begin{bmatrix} \mathcal{G}_{1,1} & 0 & \ldots & 0 & X_{1,0}^g \\ 0 & \mathcal{G}_{2,2} & \ldots & 0 & X_{2,0}^g \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & \mathcal{G}_{m,m} & X_{m,0}^g \\ -\left(X_{1,0}^g\right)^T & -\left(X_{2,0}^g\right)^T & \ldots & -\left(X_{m,0}^g\right)^T & \mathcal{G}_{0,0} \end{bmatrix}$$

$$C = \begin{bmatrix} \mathcal{C}_{1,1} & 0 & \ldots & 0 & X_{1,0}^c \\ 0 & \mathcal{C}_{2,2} & \ldots & 0 & X_{2,0}^c \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & \mathcal{C}_{m,m} & X_{m,0}^c \\ -\left(X_{1,0}^c\right)^T & -\left(X_{2,0}^c\right)^T & \ldots & -\left(X_{m,0}^c\right)^T & \mathcal{C}_{0,0} \end{bmatrix}$$

$$B = \begin{bmatrix} \mathcal{B}_1 & 0 & \ldots & 0 & 0 \\ 0 & \mathcal{B}_2 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & \mathcal{B}_m & 0 \\ 0 & 0 & \ldots & 0 & \mathbf{0} \end{bmatrix}$$

and

$$\mathrm{x} = [x_1, x_2, \ldots, x_m, x_0]^T \qquad u = [u_1, u_2, \ldots, u_m, 0]^T$$

where $G, C \in R^{\mathrm{N}\times\mathrm{N}}$, and $B \in R^{\mathrm{N}\times n_p}$.

For each block $\mathbf{M}_i$, the state variable x_i includes the nodal voltage v_n for the block conductance and capacitance, and the branch vector potential $\mathbf{A}_l$ for the block inverse inductance. Its first-order VNA admittance is $\mathcal{G}_{ii} + s\mathcal{C}_{ii}$ ($\in R^{n_i \times n_i}$), and $\mathcal{B}_i$ is the excitation-port incidence matrix ($\in R^{n_i \times n_{p_i}}$) for external current sources.

The diagonal blocks $\mathbf{M}_i$ are interconnected with $\mathbf{M}_0$ by the torn branches $X_{i0}^{g,c}$ ($\in R^{n_i \times n_0}$) in the border. For the global

interconnection block $\mathbf{M}_0$ at the bottom, the state variable x_0 includes the interfacing current variables for resistive or capacitive node splitting and the new state variable describing the flux difference for inductive branch tearing. Its first-order VNA admittance is $(G)_{0,0} + s(C)_{0,0}$ $(\in R^{n_0 \times n_0})$. In addition, all external sources are counted by $\mathcal{B}_i$ for block $\mathbf{M}_i$ $(i = 1, \ldots, m)$ and there are no external sources in $\mathbf{M}_0$.

B. BBD Solver

The BBD structure enables a hierarchical solution by a two-level domain decomposition [25] to solve each partitioned block independently.

Decompose the admittance $G + sC$ into the diagonal $\mathbf{M}$ and off-diagonal $\mathbf{X}$ parts, where the diagonal part $\mathbf{M}$ is

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_1 & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \mathbf{M}_m & 0 \\ 0 & \ldots & 0 & \mathbf{M}_0 \end{bmatrix}$$
$$\mathbf{M}_i = \mathcal{G}_{i,i} + s\mathcal{C}_{i,i} \qquad (i = 1, \ldots, m) \tag{21}$$

and the off-diagonal part $\mathbf{X}$ is

$$\mathbf{X} = \begin{bmatrix} 0 & 0 & \ldots & \mathbf{X}_1 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & 0 & \mathbf{X}_m \\ -\mathbf{X}_1 & \ldots & -\mathbf{X}_m & 0 \end{bmatrix}$$
$$\mathbf{X}_i = X^g_{i,0} + sX^c_{i,0} \qquad (i = 1, \ldots, m). \tag{22}$$

Algorithm 1 Solve BBD matrix

1. Block LU-factor of x_k

for every k *in* m $(k++)$ **do**

(1.1) *input*: M_k, X_k, $\mathcal{B}_k$;

(1.2) *factor*: $M_k = L_k U_k$;

(1.3) *solve*: $L_k \Phi_k = X_k$ for Φ_k, $\Psi_k U_k = (X_k)^T$ for Ψ_k, and $L_k \xi_k = \mathcal{B}_k$ for ξ_k;

(1.4) *form*: $F_k = \Psi_k \Phi_k$, and $G_k = \Psi_k \xi_k$

(1.5) *output*: F_k, G_k.

end for

2. Update Interconnection x_0

(2.1) *input*: M_0, F_k, G_k;

(2.2) *form*: $F = M_0 + \sum_{k=1}^{m} F_k$, $G = \sum_{k=1}^{m} G_k$;

(2.3) *solve*: $Fx_0 = G$ for x_0;

(2.4) *output*: x_0.

3. Block-backward Substitution of x_k

for every k *in* m $(k--)$ **do**

(3.1) *input*: x_0, Φ_k, ξ_k, U_k;

(3.2) *form*: $\xi_k = \xi_k - \Phi_k x_0$;

(3.3) *solve*: $U_k x_k = \xi_k$ for x_k;

(3.4) *output*: x_k.

end for

The overall procedure is outlined in Algorithm 1. Each block matrix $\mathbf{M}_i$ $(i = 1, \ldots, m)$ is first solved individually with LU factorization and substitution (1.1–1.5). Since each $\mathbf{M}_i$ is nonsingular, the aforementioned factorization can be performed at dc. The results x_i $(i = 1, \ldots, m)$ from each reduced block are then further used to solve the coupling block $\mathbf{M}_0$ for x_0 (2.1–2.4). The final x_i of each reduced block is updated (3.1–3.4) with the result from the coupling current x_0. Its computational cost is analyzed as follows. Typically, LU factorization requires $O(n_i^3)$ flops for a dense matrix and $O(n_i^\alpha)$ $(1 < \alpha < 2)$ for a sparse matrix. The computational cost of Algorithm 1 is therefore $O(\sum_{i=0}^{m} n_i^3)$ for the dense matrix and $O(\sum_{i=0}^{m} n_i^\alpha)$ for the sparse matrix. When multithread or message-passing interface (MPI) implementation is applied to factor each block in parallel, the *summation* becomes the *maximum*. Moreover, some block symbolic or numeric factorizations can be skipped if two blocks have identical fill-in patterns or even the identical stamped values.

Note that in (2.2), the factorization cost of F, *Schur's compliment* grows when the size of the coupling block $\mathbf{M}_0$ is much larger than the other blocks $\mathbf{M}_i$ $(i = 1, \ldots, m)$ in the diagonal. This happens when a large number of auxiliary variables are added after the BBD transformation. For linear circuits with the structure such as buses, clock-trees, and power/ground (P/G) meshes, $\mathbf{L}^{-1}$ can be effectively sparsified [2], [6] with the use of structured pruning such as windowing methods. In general, a sparse and direct extraction of the inverse inductance can be obtained through the box integral within the vector-potential-equivalent-circuit (VPEC) model [4]. The number of the new state variable for the flux difference can therefore be well controlled. Note that since the *min-cut* algorithm can help find the minimum nodal splittings, the number of new state variable of the interfacing current is also well controlled. However, for circuits with parasitics randomly distributed in the layout, it is difficult to design a general method determining the physical screening range for the pruning. Therefore, the size (n_0) of $\mathbf{M}_0$ may not be under control. As such, an iterative solver, such as generalized minimal residual (GMRES), needs to be utilized to effectively solve Schur's compliment F [25].

A detailed discussion of GMRES method can be found in [25]. The critical step for a successful GMRES iteration is to find a proper preconditioner. Since F usually contains zero diagonal entries in circuit simulation, in this paper, we apply a banded block diagonal preconditioner P

$$p_{ij} = \begin{cases} f_{ij}, & \text{if } |f_{ij}| > \epsilon \cdot \max(|f_{ii}|, |f_{jj}|) \\ 0, & \text{otherwise} \end{cases} \tag{23}$$

for the GMRES iterations. If the number of GMRES iterations is n_G, the computational cost of GMRES is about $O(n_G \cdot n_0^2)$, mainly from the matrix-vector multiplication.

V. Block-Structure-Preserved Reduction

Although stretching introduces a sparse yet hierarchical representation, the block size and the overall system size may still be large. The size of the state matrix can be reduced by

the model-order reduction by finding a set of orthonormalized projection vectors Q, which spans the following block Krylov subspace:

$$\mathcal{K}(A,R,q)=\text{span}\{R,AR,\ldots,A^{q-1}R\}\subseteq\text{span}\{Q\}$$

with

$$A=(G+s_0C)^{-1}C \qquad R=(G+s_0C)^{-1}B.$$

Algorithm 2 BBD-based block Arnoldi method

(1.1) *input*: G, C, B;

(1.2) *BBD-solve*: $(G+s_0C)Q^{(0)}=B$ for $Q^{(0)}$;

(1.3) *orthonormalize*: each column in $Q^{(0)}$;

for every $i\ in\ q-1$ **do**

(1.4) *BBD-solve*: $(G+s_0C)Q^{(i)}=CQ^{(i-1)}$ for $Q^{(i)}$;

(1.5) *orthogonalize*: $Q^{(i)}$ to all $Q^{(j)}$ $(j=0,\ldots,i-1)$;

(1.6) *orthonormalize*: each column in $Q^{(i)}$;

end for

(1.7) *compose*: $Q=[Q^{(0)},Q^{(1)},\ldots,Q^{(q-1)}]$;

(2.1) *partition*: $Q:\rightarrow Q_{0_{n_0\times n_{p_0}}},\ldots,Q_{m_{n_m\times n_{p_m}}}$;

for every $j\ in\ m$ **do**

(2.2) *merge*: Q_i and Q_{i+1} till a new Q_i' is nonsingular;

(2.2) *orthonormalize*: each column in Q_i;

end for

(2.3) *compose*: $Q=diag[Q_1,\ldots,Q_m,Q_0]$;

The orthonormalization is usually performed by the block Arnoldi procedure [10]. It needs a heavy computation to factorize the state matrix. As we have formulated a BBD-structured state matrix, the BBD solver in Algorithm 1 can be applied to reduce the computational cost during the block Arnoldi procedure. A new block-Arnoldi orthonormalization enhanced by the BBD solver is presented in Algorithm 2. Clearly, due to the BBD formulation, the factorization cost in (1.2) and (1.4) can be dramatically reduced.

However, by directly applying a flat Q to project the original state matrices, the BBD structure is destroyed and the reduced model is still inefficient to be reused in SPICE. As shown in Algorithm 2, this can be solved by constructing a structured projection matrix. The flat Q is split into a structured $\mathcal{Q}$

$$Q=\begin{bmatrix}Q_1\\Q_2\\\vdots\\Q_m\\Q_0\end{bmatrix}\rightarrow\mathcal{Q}=\begin{bmatrix}Q_1&&&&\\&Q_2&&&\\&&\ddots&&\\&&&Q_m&\\&&&&Q_0\end{bmatrix} \tag{24}$$

according to the size of each block, where $Q_i\in R^{n_i\times n}$ $(1\leq i\leq m,\ q=\lfloor n/n_p\rfloor)$.

As each Q_i is further orthonormalized, it leads to

$$\mathcal{Q}^T\mathcal{Q}=I.$$

As such, each column of $\mathcal{Q}$ is still linearly independent and the total column rank of $\mathcal{Q}$ $(\text{rank}=m\times n)$ is increased by a factor of m compared to the column rank of Q $(\text{rank}=n)$. The m-times increased column rank means m-times more poles can be approximated after projection. Note that sometimes a truncated Q_i may become singular and impossible to be orthonormalized with previous Q_{i-1}'s. As shown in Algorithm 2, we resolve this by merging Q_i with Q_{i+1} until the new Q_i' becomes nonsingular. This is similar to the deflation/look-ahead procedure in Pade Via Lanczos (PVL) for the same reason [8].

The order-reduced state matrices by $\mathcal{Q}$ become

$$\widetilde{G}=\mathcal{Q}^TG\mathcal{Q}, \qquad \widetilde{C}=\mathcal{Q}^TC\mathcal{Q}, \qquad \widetilde{B}=\mathcal{Q}^TB \tag{25}$$

where

$$\widetilde{G}_{i,j}=\mathcal{Q}_i^TG_{i,j}\mathcal{Q}_j, \qquad \widetilde{C}_{i,j}=\mathcal{Q}_i^TC_{i,j}\mathcal{Q}_j, \qquad \widetilde{B}_i=\mathcal{Q}_i^TB_i.$$

As a result, the transfer function is

$$\widetilde{H}(s)=\widetilde{B}^T[\widetilde{G}+s\widetilde{C}]^{-1}\widetilde{B}. \tag{26}$$

We call this reduction the BVOR (BBD-based VOR) method. The reduced model $\widetilde{H}(s)$ has the following properties.

Theorem 1: The first q block moments expanded at s_0 are identical for $\widetilde{H}(s)$ and $H(s)$.

Since $\mathcal{K}(\mathcal{A},\mathcal{R},q)\subseteq\text{span}\{Q\}\subseteq\text{span}\{\mathcal{Q}\}$, there are q block moments matched according to Grimme's projection theorem [26].

In addition, the reduced system is passive.

Theorem 2: When the input and the output are symmetric, $\widetilde{H}(s)$ projected by $\mathcal{Q}$ is passive.

The proof of this theorem is similar to [10], where a symmetry between input and output ports is assumed. Note that the BBD form of (20) becomes

$$\begin{bmatrix}\mathbf{M}&\mathbf{X}\\-\mathbf{X}^T&\mathbf{M}_0\end{bmatrix}\begin{bmatrix}\mathbf{x}\\x_0\end{bmatrix}=\begin{bmatrix}\mathbf{B}\\\mathbf{0}\end{bmatrix}. \tag{27}$$

This leads to the following equivalent passivity check:

$$\begin{aligned}H(s)+H(s)^*&=[\mathbf{B}^T\quad 0]\begin{bmatrix}\mathbf{M}+\mathbf{M}^*&\mathbf{0}\\\mathbf{0}&\mathbf{M}_0+\mathbf{M}_0^*\end{bmatrix}\begin{bmatrix}\mathbf{B}\\\mathbf{0}\end{bmatrix}\\&=\mathbf{B}^T(\mathbf{M}+\mathbf{M}^*)\mathbf{B}.\end{aligned}$$

Clearly, because each block in $\mathbf{M}$ is symmetric and semipositive-definite, their block diagonal matrix $\mathbf{M}+\mathbf{M}^T$ is also symmetric and semipositive-definite. With the use of the congruence transformation by $\mathcal{Q}$, the reduced $\widetilde{H}(s)$ thereby preserves the passivity.

Since the structured projection preserves the BBD structure, the BBD solver in Algorithm 1 can be employed further to efficiently solve the reduced system. As sparsity is preserved at the block level, a sparse solver can still be applied on the top, and a

dense-matrix solver is then applied to solve each reduced block with a much smaller dimension than the original.

VI. Numerical Experiments

Numerical experiments are presented to demonstrate the accuracy and efficiency of the proposed VNA and BVOR methods. They are compared to the exact MNA without reduction, the second-order reduction SAPOR [13] in the NA, and the first-order reduction PACT [9] in the MNA with the use of 2×2 split congruence transformation. We have implemented a SPICE transient simulator with the VNA and BVOR in MATLAB. A modernized SPICE simulator ngspice (http://ngspice.sourceforge.net/) is used to parse the netlist and modified to dump the VNA state matrix into a sparse matrix for MATLAB. The algorithm hmetis [28] partitions the mapped hypergraph, and BBD stretching builds a transformed BBD state matrix. As the dense-matrix solver basic linear algebra subprograms/linear algebra package (BLAS/LAPACK) and the sparse-matrix solver such as unsymmetric multifrontal sparse LU factorization package (UMFPACK) are all available in MATLAB, the model-order reduction and the numerical transient integration by Backward–Euler with the LTE are implemented in MATLAB for the prototyping in this paper. The model-order reduction is first performed to obtain the reduced state matrices, which are further stamped back for the frequency and time-domain simulations. A corresponding hybrid dense and sparse solver is developed to solve the reduced model with the BBD structure.

Moreover, the column-approximated minimum degree (colamd) is utilized from MATLAB as the permutation method to reduce fill-ins. It is applied at the block level in BVOR and applied to the whole matrix for SAPOR, PACT, and VOR during each LU. The scaling is applied when the state matrix is ill-conditioned, where each column entry is divided by the maximum entry at that row. In addition, the GMRES is implemented in MATLAB with a pruning factor $\epsilon = 1e^{-3}$ and iteration tolerance $1e^{-6}$, when the size of the interconnecting block is as large as 30% of the total size. Experiments are run on a Linux workstation with Intel Pentium IV 2.66G CPU and 2 GB RAM. A number of structured $\mathbf{RLC}$ circuits such as buses, clock-trees, and P/G meshes are extracted as provided inputs. The sparsification method in [6] is applied to obtain $\mathbf{L}^{-1}$ elements.

A. Validation of VNA

To validate our VNA formulation, i.e., the VNA stamping of inverse inductance, we use a quarter-wavelength-long transmission line (Tline) at 1 GHz, i.e., $\text{length} = 7.5\,\text{cm}$. We describe the Tline by a distributed $\mathbf{RLC}$ circuit in the PEEC model, and stamp it in the MNA and VNA according to the circuit topology of the PEEC model. We then short one end of the transmission line and compute the impedance at the other end using VNA and MNA, respectively. Fig. 2 plots the magnitude of impedance at frequencies between 0.1 and 10 GHz. Clearly, the VNA stamping can correctly capture the resonant (electromagnetic interaction) as accurately as does the MNA stamping in the whole frequency range.

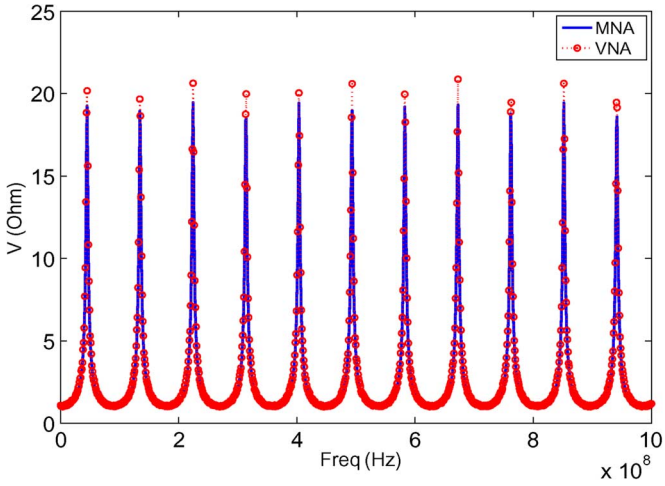

Fig. 2. Frequency-domain waveform comparison between VNA and MNA for a one-end shorted Tline at far-ends of buses.

B. Accuracy of VOR and BVOR

We first compare the waveform accuracy between VOR and SAPOR. The example is an 8-bit bus with 20 segments per bit, which is similar to the example used by SAPOR [13]. In addition, the driver is modeled by a 100 Ω resistor and the receiver is modeled by a 1 pF capacitor.

We first expand first-order admittance in VNA and the second-order admittance in NA around dc ($s_0 = 10$ Hz). The order of q of all reductions is selected as 80 when the maximum error of the frequency-domain waveform by VOR is below to 1%. The near-end of the first bit is excited with a unit voltage impulse source for the MNA/VNA, and an equivalent impulse current source with source resistor is used for NA. The response at far-ends are observed. Fig. 3 shows the frequency- and time-domain responses of the exact MNA, reduced models by SAPOR and by VOR. As state matrices in NA are indefinite around dc, SAPOR shows large numerical errors beyond 1 GHz. In contrast, VOR has an accuracy similar to the exact MNA for up to 50 GHz. The reduced model by VOR is also as accurate as the MNA in a time-domain transient simulation. However, the one by SAPOR is not included as the state matrices are indefinite for the initial time step.

The second example is a 32-bit bus with 20 segments per bit, where shielding is inserted for every 8 bits. The partitioning results in four blocks, each with 180 resistors, 250 capacitors, and around 25 600 nodal susceptors. The near-end of the first bit in each block is excited with a unit current impulse. BVOR is applied to reduce each block independently. Fig. 4(a) and (b) shows the nonzero BBD pattern of the G matrix in VNA before and after the reduction, and Fig. 4(c) and (d) shows those for the C matrix in VNA. Clearly, the reduced-state matrices preserve the BBD structure and are sparse. $\widetilde{G}$ and $\widetilde{C}$ have the sparsification ratio of 16.1% and 14.6%, respectively.

We study the scalability of accuracy versus circuit size in Fig. 5 between VOR and SAPOR. Since there are many differently sized circuits with different numbers of ports, all reductions are performed close to dc ($s_0 = 10\,\text{Hz}$) using order $q = 50$ for the convenience of comparison. The standard deviation of

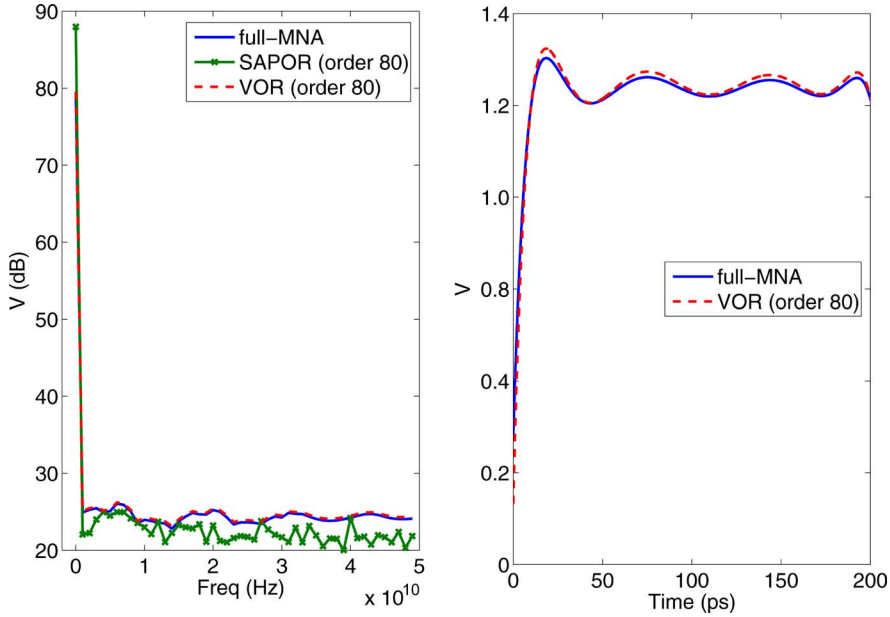

Fig. 3. Frequency- and time-domain waveform comparison of full-MNA, SAPOR ($\mathrm{order} = 80$) and VOR ($\mathrm{order} = 80$) at far-ends of buses. The reduced models are expanded close to dc ($s_0 = 10$ Hz). VOR and original are visually identical. The reduced model by SAPOR cannot converge in time-domain simulation.

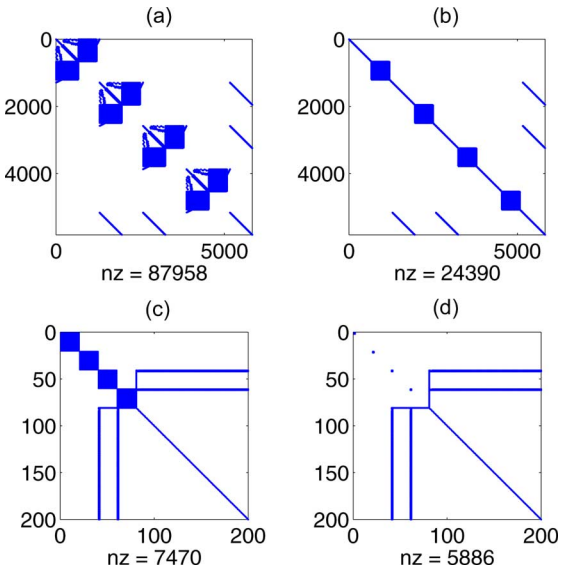

Fig. 4. BBD structure preserving of state matrices $\mathcal{G}$ and $\mathcal{C}$ before and after reduction, where NZ is the number of nonzero entries. (a) *G* before reduction. (b) *C* before reduction. (c) *G* after reduction. (d) *C* after reduction.

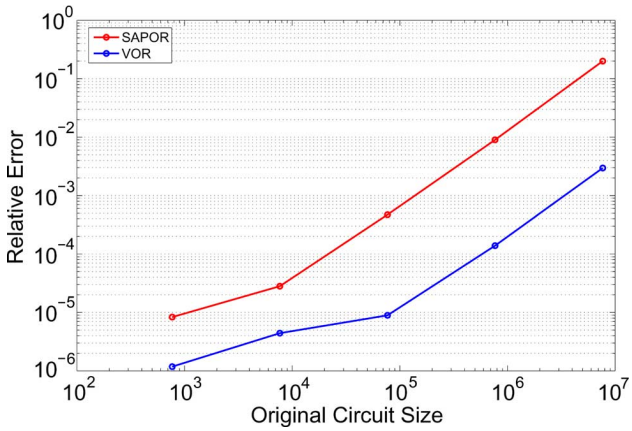

Fig. 5. Frequency-domain waveform error scalability between SAPOR and VOR.

waveform differences at frequency domain is used as the measurement of error. As state matrices in NA are indefinite around dc, the moment matching in SAPOR breaks down easily and results in a reduced model with low accuracy. In contrast, state matrices in VNA form are not singular, and the reduced model by BVOR has up to 67× smaller error than that by SAPOR.

C. Capacity and Runtime

We further compare the capacity and runtime in Table I. Column 1 shows three types of **RLC** circuits, including parallel buses, clock trees, and P/G meshes. Column 2 shows the sizes of state variables before and after model-order reductions by BVOR. All circuits are reduced to macromodels with a similar accuracy. Ten percent of nodes in each circuit are assigned as unit impulse voltage sources. Column 3 shows the fill-ins before and after $\mathbf{L}^{-1}$ sparsification by BVOR. Experiment results show that the total fill-ins of $G + s_0C$ by the sparsified VNA are up to 10× smaller than the one by the full MNA. Column 4 shows the number of blocks used in BVOR.

Columns 5–11 compare the runtime of the original model in the MNA and the reduced models by SAPOR, PACT, and BVOR, respectively. The runtime here includes both the macromodel building and simulation time. The reduction and simulation are performed in the frequency domain. For a circuit with 4500^2 elements (clktree3), BVOR reduces the orthonormalization cost, and hence, is about 5× (13.2 s versus 62.6 s) faster than PACT when building macromodels. BVOR is also 7× (13.2 s versus 101.9 s) faster to build than SAPOR for the same circuit. The additional speedup is due to the fact that the first-order MNA matrix used in PACT is sparser than the NA matrix used in

TABLE I
RUNTIME COMPARISON OF SAPOR, PACT, AND BVOR WITH SIMILAR ACCURACY

ckt	states (before:after)	fill-ins (before:after)	block #	Original(MNA) sim(s)	SAPOR(NA) build(s)	SAPOR(NA) sim(s)	PACT(MNA) build(s)	PACT(MNA) sim(s)	BVOR(BBD-VNA) build(s)	BVOR(BBD-VNA) sim(s)
bus8x20	325:40	6725:1882	4	99.3	2.02	11.9	1.45	10.2	0.55	4.32
bus32x20	1200:200	84K:26K	16	1.96e3	16.9	102.3	14.4	90.0	3.42	12.5
bus128x20	6000:500	2.2M:0.3M	64	8.73e4	92.7	1.05e3	61.3	906.8	15.6	31.6
bus256x20	12000:800	10M:0.9M	64	NA	NA	NA	NA	NA	142.5	158.8
clktree1	1215:200	91K:18K	4	2.09e3	18.7	104.1	15.9	91.3	2.82	11.2
clktree2	2100:300	242K:41K	8	6.02e3	23.4	553.3	24.3	431.8	4.89	22.7
clktree3	4500:400	562K:67K	16	1.83e4	101.9	640.0	62.6	572.4	13.2	25.2
clktree4	36000:800	12M:1.3M	16	NA	NA	NA	NA	NA	1.59e3	2.38e3
mesh1	800:80	41K:9K	8	524.1	10.0	31.2	5.12	17.1	1.64	7.48
mesh2	2000:200	112K:21K	16	5.42e3	19.1	105.5	16.9	58.2	2.81	12.2
mesh3	6000:400	4M:0.7M	32	1.08e5	96.3	649.9	53.6	581.4	18.9	25.2
mesh4-1	48000:1200	24M:2.5M	32	NA	NA	NA	NA	NA	7.27e3	1.69e4
mesh4-2	48000:1200	24M:2.5M	64	NA	NA	NA	NA	NA	9.05e3	1.42e4
mesh4-3	48000:1200	24M:2.5M	128	NA	NA	NA	NA	NA	1.05e4	1.82e4

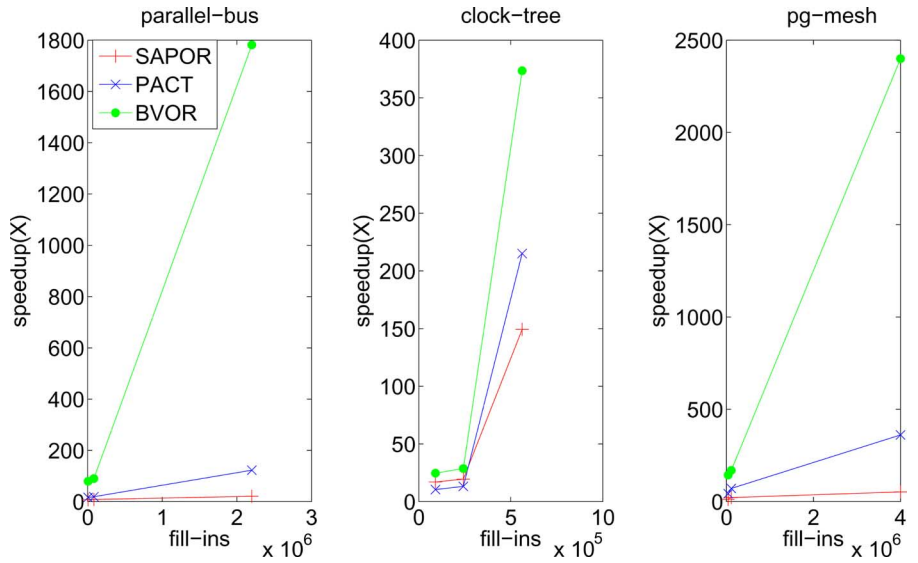

Fig. 6. Runtime scalability of SAPOR, PACT, and BVOR for parallel bus, clock trees, and P/G mesh, all with mutual inductances. Results of macromodels are compared (divided) with the result of the full MNA.

SAPOR. Moreover, for a circuit with 6000^2 elements (bus 128×20), because the BBD structure is preserved after the reduction, the two-level analysis in BVOR further reduces the simulation by $33\times$ and $29\times$ when compared to SAPOR and PACT, respectively. Fig. 6 further shows the scalability of SAPOR, PACT, and BVOR for three types of inductive interconnects. The y-axis is the speedup ($\times$) compared to the baseline of the original MNA model with dense mutual inductances and the x-axis is the fill-ins of the original MNA model. Clearly, BVOR shows a dramatic speedup when compared to other MORs for all three types.

Note that the size of the interconnecting block is large when more partitions are used for mesh4. The iterative solver is then utilized to factor Schur's compliment. We find that the overall runtime to build and simulate the macromodel is still reduced for mesh4 with 64 blocks compared to mesh4 with 32 blocks. The reason is that the saving of factorization cost from the small sized block is larger than the increase of factorization cost from Schur's compliment. However, we notice that when more partitions are used, the increased factorization cost for Schur's compliment could slow down the overall runtime.

VII. CONCLUSION

To handle large-scale inductive interconnects, this paper presents a block-structure-preserved macromodeling by exploring the efficiency of the matrix structure such as sparsity and hierarchy. First step is taken to obtain a loosely coupled inductive network. To stably stamp the sparsified inductive couplings into passive and definite state matrices, a VNA stamping is introduced to replace the state variable of the inductive-branch current by magnetic flux. After the stable sparsification, the sparsity and hierarchy of state matrices are further utilized by a matrix-stretching method to a BBD structure. Such a matrix stretching is the first work in literature that considers the coupling from inductive-branch currents. In addition, the BBD structure is preserved in the reduced macromodel by the block-structure-preserved model-order reduction. This leads to a hybrid matrix solver to efficiently factor the resulting system. The dense-matrix solver is applied to solve each reduced block, and then the sparse solver is applied on the top. In experiments performed on several test cases, our method achieved up to $7\times$ faster modeling building time, up to $33\times$

faster simulation time, and as much as 67× smaller waveform error compared to SAPOR (a second-order reduction based on NA) and PACT (a first-order 2×2 structured reduction based on MNA). Note that most circuits also show a block-level latency, i.e., a distribution of time constants. Some blocks have a fast changing rate and some others have a slow changing rate. A structure-preserved macromodeling algorithm to consider the block-level latency is developed in [30]. Future studies will also compare the performance of all these structure-preserved macromodeling methods.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their insightful comments to make this paper better. This work was performed at the University of California, Los Angeles.

REFERENCES

[1] Y. I. Ismail, E. G. Friedman, and J. L. Neves, "Exploiting on-chip inductance in high speed clock distribution networks," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 9, no. 6, pp. 963–973, Dec. 2001.

[2] A. Devgan, H. Ji, and W. Dai, "How to efficiently capture on-chip inductance effects: Introducing a new circuit element K," in *Proc. Int. Conf. Comput. Aided Des. (ICCAD)*, 2000, pp. 150–155.

[3] Y. Massoud and J. White, "Simulation and modeling of the effect of substrate conductivity on coupling inductance and circuit crosstalk," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 10, no. 3, pp. 286–291, Jun. 2002.

[4] A. Pacelli, "A local circuit topology for inductive parasitics," in *Proc. Int. Conf. Comput. Aided Des. (ICCAD)*, 2002, pp. 208–214.

[5] R. Escovar, S. Ortiz, and R. Suaya, "An improved long distance treatment for mutual inductance," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 24, no. 5, pp. 783–793, May 2005.

[6] H. Yu and L. He, "A provably passive and cost efficient model for inductive interconnects," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 24, no. 8, pp. 1283–1294, Aug. 2005.

[7] T. L. Pillage, R. A. Roher, and C. Visweswariah, *Electronic Circuit and System Simulation Methods*. New York: McGraw-Hill, 1994.

[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] K. J. Kerns and A. T. Yang, "Preservation of passivity during RLC network reduction via split congruence transformations," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 17, no. 7, pp. 582–591, Jul. 1998.

[10] A. Odabasioglu, M. Celik, and L. Pileggi, "PRIMA: Passive reduced-order interconnect macro-modeling algorithm," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 17, no. 8, pp. 645–654, Aug. 1998.

[11] B. N. Sheehan, "ENOR: Model order reduction of RLC circuits using nodal equations for efficient factorization," in *Proc. Des. Autom. Conf. (DAC)*, 1999, pp. 17–21.

[12] H. Zheng and L. T. Pileggi, "Robust and passive model-order reduction for circuits containing susceptance elements," in *Proc. Int. Conf. Comput. Aided Des. (ICCAD)*, 2002, pp. 761–766.

[13] Y. Su, J. Wang, X. Zeng, Z. Bai, C. Chang, 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.

[14] R. W. Freund, "SPRIM: Structure-preserving reduced-order interconnect macro-modeling," in *Proc. Int. Conf. Comput. Aided Des. (ICCAD)*, 2004, pp. 80–87.

[15] G. Shi and C.-J. R. Shi, "Model order reduction by dominant subspace projection: Error bounds, subspace approximation and circuit applications," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 52, no. 5, pp. 975–993, May 2005.

[16] P. Heydari and M. Pedram, "Model-order reduction using variational balanced truncation with spectral shaping," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 53, no. 4, pp. 879–891, Apr. 2006.

[17] C. Ratzlaff and L. Pillage, "RICE: Rapid interconnect circuit evaluation using AWE," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 13, no. 6, pp. 763–776, Jun. 1994.

[18] A. E. Ruehli, "Equivalent circuits models for three dimensional multiconductor systems," *IEEE Trans. Microw. Theory Tech.*, vol. MTT-22, no. 3, pp. 216–221, Mar. 1974.

[19] C. W. Ho, A. E. Ruehli, and P. A. Brennan, "The modified nodal approach to network analysis," *IEEE Trans. Circuits Syst.*, vol. 22, no. CAS-6, pp. 504–509, Jun. 1975.

[20] L. Nagel, "Spice2: A computer program to simulate semiconductor circuits," Electron. Res. Lab., Univ. California, Berkeley, Rep. ERL-M520, 1975.

[21] J. D. Jackson, *Classical Electrodynamics*. Hoboken, NJ: Wiley, 1975.

[22] C. A. Balanis, *Advanced Engineering Electromagnetics*. Hoboken, NJ: Wiley, 1989.

[23] H. Yu, L. He, and S. Tan, "Block structure preserving model reduction," in *Proc. IEEE Int. Behav. Model. Simul. Conf. (BMAS)*, 2005, pp. 1–6.

[24] F. Wu, "Solution of large-scale networks by tearing," *IEEE Trans. Circuits Syst.*, vol. CAS-23, no. 12, pp. 706–713, Dec. 1976.

[25] Y. Saad, *Iterative Methods for Linear Systems*. Boston, MA: PWS-Kent, 2000.

[26] E. J. Grimme, "Krylov projection methods for model reduction," Ph.D. dissertation, Univ. Illinois at Urbana-Champaign, Urbana-Champaign, IL, 1997.

[27] M. Celik, L. Pileggi, and A. Odabasioglu, *IC Interconnect Analysis*. New York: Springer-Verlag, 2002.

[28] G. Karypis, R. Aggarwal, and V. K. S. Shekhar, "Multilevel hypergraph partitioning: Application in VLSI domain," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 7, no. 1, pp. 69–79, Mar. 1999.

[29] H. Yu, C. Chu, and L. He, "Off-chip decoupling capacitor allocation for chip package co-design," in *Proc. Des. Autom. Conf. (DAC)*, 2007, pp. 618–621.

[30] H. Yu, Y. Shi, and L. He, "Fast analysis of structured power grid by triangularization based structure preserving model-order reduction," in *Proc. Des. Autom. Conf. (DAC)*, 2006, pp. 205–210.

Hao Yu (S'02–M'06) received the Ph.D. degree in electrical engineering from the University of California, Los Angeles (UCLA), in 2006.

Between 2002 and 2006, he was a Graduate Student Researcher of electrical engineering at UCLA. Since 2006, he was a Senior Member of the Technical Staff, Berkeley Design Automation, Santa Clara, CA, where he was engaged in developing the fastest analog/RF simulation. He is currently an Assistant Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His current research interests include numerical prototyping of high-performance mixed-mode circuits: structured and parameterized model-order reduction and fast differential/integral equation solvers for SPICE and TCAD, and macromodel-based synthesis platform: integrity and robustness driven hybrid-system-integration from device to system level.

Dr. Yu was a recipient of the Best Paper Award at the Design Automation Conference in 2006, the International Conference of Computer-Aided-Design in 2006, and the Global Research Collaboration Inventor Award in 2008.

Chunta Chu received the B.S. degree in electrical engineering from Tsinghua University, Beijing, China, in 2003, and the M.S. degree in electrical engineering from the University of California, Los Angeles, in 2007.

In 2007, he joined Apache Design Solutions, San Jose, CA, where he is currently a Senior Software Engineer and is engaged in the development of analog power noise analysis tool. His research interests include design automation of VLSI circuits, modeling and simulation of power/ground network, model-order reduction technique, and power noise integrity.

Yiyu Shi (S'07) received the B.E. degree (with honors) in electronic engineering from Tsinghua University, Beijing, China, in 2005, and the M.S. degree (with honors) in electrical engineering from the University of California, Los Angeles, in 2007, where he is currently working toward the Ph.D. degree.

His current research interests include design automation for VLSI circuits and systems and large-scale optimization.

David Smart (M'76) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Illinois at Urbana-Champaign, Urbana-Champaign, in 1976, 1977, and 1988, respectively.

He worked at GTE Communication Systems for seven years, and at the IBM Watson Research Center for one summer. Since 1988, he has been with Analog Devices Inc., where he is currently an ADI Fellow, leading the development of simulation-based tools and methods for the design of analog and mixed-signal ICs. His current research interests include simulation algorithms, interconnect modeling and reduction, optimizing model compilers, and tools for robust ESD design.

Lei He (S'94–M'99–SM'08) received the Ph.D. degree in computer science from the University of California, Los Angeles (UCLA), in 1999.

Between 1999 and 2001, he was a faculty member at the University of Wisconsin, Madison. He also held visiting or consulting positions at Intel, Hewlett-Packard, Cadence, Synopsys, Rio Design Automation, and Apache Design Solutions. He is currently an Associate Professor with the Electrical Engineering Department, UCLA. His research interests include VLSI circuits and systems, and electronic design automation. He has authored or coauthored more than 160 technical papers.

Dr. He was a recipient of the National Science Foundation CAREER Award in 2000, the UCLA Chancellor's Faculty Career Development Award (highest class) in 2003, the IBM Faculty Award in 2003, the Northrop Grumman Excellence in Teaching Award in 2005, the Best Paper Award at the 2006 International Symposium on Physical Design, and Multiple Best Paper Nominations at the Design Automation Conference and the International Conference on Computer-Aided Design. He has been a member of technical program committees of a number of conferences, including the Design Automation Conference, the International Conference on Computer-Aided Design, the International Symposium on Low Power Electronics and Design, and the International Symposium on Field Programmable Gate Array.

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.

From 1995 to 1996, he was a faculty member in the Electrical Engineering Department, Fudan University. He is currently an Associate Professor in the Department of Electrical Engineering, University of California–Riverside, Riverside. His research interests include modeling and simulation of analog/RF/mixed-signal and interconnect circuits, analysis and optimization of high-performance power and clock distribution networks, architecture-level thermal, power, modeling, and simulation for multicore microprocessors and embedded system designs based on field-programmable gate array platforms. He is a coauthor of the book *Symbolic Analysis and Reduction of VLSI Circuits* (Springer-Verlag/Kluwer, 2005) and *Advanced Model Order Reduction Techniques for VLSI Designs* (Cambridge University Press, 2007).

Dr. Tan was a recipient of the Outstanding Oversea Investigator Collaboration Award from the National Natural Science Foundation (NSF) of China in 2008, the NSF CAREER Award in 2004, the Best Paper Award at the 2007 IEEE International Conference on Computer Design, a Best Paper Award Nomination at the 2005 and the 2009 IEEE/ACM Design Automation Conference, and the Best Paper Award at the 1999 IEEE/ACM Design Automation Conference. He was a member of the technical program committees of the Asia and South Pacific Design Automation Conference, the IEEE Behavioral Modeling and Simulation Conference, the IEEE International Symposium on Quality Electronic Design, and the International Conference on Computer Aided Design. He is an Associate Editor for three journals: *Association for Computing Machinery (ACM) Transaction on Design Automation of Electronic Systems*, *Integration, The VLSI Journal*, and the *Journal of VLSI Design*.