Parametric Structure-Preserving Model Order Reduction ⋆

Jorge Fernández Villena⋆⋆[1], Wil H.A. Schilders[23], and L. Miguel Silveira[14]

[1] INESC ID / IST - Tech. University of Lisbon, Rua Alves Redol 9 1000-029 Lisbon, Portugal
[2] NXP Semiconductors Research, High Tech Campus 37, 56656 AE Eindhoven
[3] Dept. of Mathematics and Computer Science, Tech. University of Eindhoven, 5600 MB, Eindhoven, The Netherlands
[4] Cadence Research Labs, Rua Alves Redol 9 1000-029 Lisbon, Portugal
{jorge.fernandez,lms}@inesc-id.pt wil.schilders@nxp.com

Abstract. Analysis and verification environments for next-generation nano-scale RFIC designs must be able to cope with increasing design complexity and to account for new effects, such as process variations and Electromagnetic (EM) couplings. Designed-in passives, substrate, interconnect and devices can no longer be treated in isolation as the interactions between them are becoming more relevant to the behavior of the complete system. At the same time variations in process parameters lead to small changes in the device characteristics that may directly affect system performance. These two effects, however, cannot be treated separately as the process variations that modify the physical parameters of the devices also affect those same EM couplings. Accurately capturing the effects of process variations as well as the relevant EM coupling effects requires detailed models that become very expensive to simulate. Reduction techniques able to handle parametric descriptions of linear systems are necessary in order to obtain better simulation performance. In this work we discuss parametric Model Order Reduction techniques based on Structure-Preserving formulations that are able to exploit the hierarchical system representation of designed-in blocks, substrate and interconnect, in order to obtain more efficient simulation models.

1 Introduction

New coupling and loss mechanisms, including EM field coupling and substrate noise as well as process-induced variability, are becoming too strong and too relevant to be neglected, whereas more traditional coupling and loss mechanisms are more difficult to describe given the wide frequency range involved and the greater variety of structures to be modeled. The performance of each device in the circuit is strongly affected by the environment surrounding it. In other words, the response of each circuit part depends not only on its own physical and electrical characteristics, but to a great extent also on its positioning in the IC, i.e. on the devices to which it is directly connected to or coupled with. The high level of integration available in current RFIC designs leads to proximity effects between the devices, which induce EM interactions, that can

⋆ Work partially supported by the EU/IST/FP6/027378 STREP CHAMELEON RF project.
⋆⋆ Also supported by EC/IST/FP6/019417 COMSON Network under a Marie Curie Fellowship.

lead to different behaviors of the affected parts. In any manufacturing process there is always a certain degree of uncertainty involved given our limited control over the environment. For the most part this uncertainty was previously ignored when analyzing or simulating complete systems, or assumed to be accounted for in the individual device models. However, as we step towards the nano-scale and higher frequency eras, such environmental, geometrical and electromagnetic fluctuations become more significant. Nowadays, parameter variability can no longer be disregarded, and its effect must be accounted for in early design stages so that unwanted consequences can be minimized. This leads to parametric descriptions of systems, including the effects of manufacturing variability, which further increases the complexity of such models. Reducing this complexity is paramount for efficient simulation and verification. However, the resulting reduced models must retain the ability to capture the effects of small fluctuations, in order to accurately predict behavior and optimize designs. This is the realm of *Parametric Model Order Reduction* (pMOR). Furthermore, these parametric fluctuations of the physical characteristics of the devices can affect not only the performance of such devices, but also the coupling between devices. For this reason the parametric models of the individual blocks of a system can no longer be simulated in isolation but must be treated as one entity and verified together. Such reduction must take advantage of the hierarchical description of those systems, namely to account for designed-in elements as well as interconnect effects. To this end, structure-preserving techniques must be used which not only retain structural properties of the individual systems but also its connections and couplings.

The goal of this paper is therefore to present and discuss techniques for model order reduction of interconnect, substrate or designed-in passives, taking into account their dependence on relevant process or fabrication parameters and their coupling and connections. The paper is structured as follows: Section 2 gives an introduction to Model Order Reduction. In Section 3 an overview of several existing pMOR techniques will be presented. In Section 4 an introduction to two-level hierarchical MOR will be done, and an extension to improve the reduction will be presented. In Section 5 the proposed methodology for combining the parametric techniques with the hierarchical reduction will be proposed. To illustrate the procedure, its pros and cons, in Section 6 some reduction results will be presented for several real-life structures. Finally conclusions will be drawn in Section 7.

2 Model Order Reduction

Model Order Reduction (MOR) is a framework whose aim is to efficiently find a behavioral equivalent yet reduced representation of a system. The system is usually represented in its state-space representation, which in descriptor form can be written as

$$\begin{aligned} C\dot{x}(t) + Gx(t) &= Bu(t) \\ y(t) &= Lx(t) \end{aligned} \tag{1}$$

where $C, G \in \mathbb{R}^{n\times n}$ are respectively the dynamic and static matrices describing circuit behavior, $B \in \mathbb{R}^{n\times m}$ is the matrix that relates the input vector $u \in \mathbb{R}^m$ to the inner states

$x \in \mathbb{R}^n$ and $L \in \mathbb{R}^{n \times p}$ is the matrix that links those inner states to the outputs $y \in \mathbb{R}^p$. This time-domain descriptor yields a frequency response modeled via the transfer function

$$H(s) = L(sC + G)^{-1}B \tag{2}$$

for which we seek to generate a reduced order approximation, able to accurately capture the input-output behavior of the system,

$$\hat{H}(s) = \hat{L}(s\hat{C} + \hat{G})^{-1}\hat{B}. \tag{3}$$

Existing methods for linear model reduction can be broadly characterized into two types: those based on balancing techniques (sometimes also referred to as SVD[5]-based [1]), and those that are based on projection methods.

The first set of techniques, those in the truncated balanced realization (TBR) family [2], perform reduction based on the concept of controllability and observability of the system states. They rely on balancing the system and then truncating the states with small controllability and observability, information given by the Hankel Singular Values of the product of the system Gramians, which are obtained by solving a pair of Lyapunov equations. These methods are purported to produce nearly optimal models and have easy to compute *a-posteriori* error bounds. There are also known techniques [3] that extent this framework in order to guarantee the passivity of the *Reduced Order Model* (ROM), independently of the structure of its representation. However, the TBR procedures are awkward to implement and expensive to apply, which limits their applicability to small and medium sized problems. Hybrid techniques that combine some of the features of each type of methods have also been presented [4–6].

Among the second set of techniques, Krylov subspace projection methods such as PVL [7] and PRIMA [8] have been the most widely studied over the past decade. They are very appealing because of their simplicity and performance in terms of efficiency and accuracy. They rely on the computation of a basis for the Krylov subspace, $colspan\{V\} = Kr\{A, R, q \times m\}$, which encloses information about the transfer function, with $A = G^{-1}C$, $R = G^{-1}B$, and q the number of block moments matched (each block with m columns). The projection of the high-dimensional original system in the lower-dimensional generated subspace guarantees such implicit moment matching and avoids numerical errors in the reduction process.

$$\hat{G} = V^T G V \qquad \hat{C} = V^T C V \qquad \hat{B} = V^T B \qquad \hat{L} = LV \tag{4}$$

Furthermore, this orthogonal projection (and congruence transformation), performed in PRIMA, guarantees the preservation of the passivity in the reduction process if C, G are positive definite and $B = L^T$ (see [8] for details). However the procedures in this framework exhibit several known shortcomings. The lack of a general strategy for error control and order selection, as well as a dependence on the original model's structure if passivity is to be guaranteed after the reduction are among the more obvious ones.

A different technique that attempts to establish a bridge between the two families of methods was also proposed. The Poor Man's TBR [9] is based on a projection scheme

[5] SVD – Singular value decomposition.

where the projection matrix approximately spans the dominant eigenspaces of the controllability and observability matrices and provides an interesting platform for bridging between the two types of techniques. The controllability Gramian is estimated via a frequency-based quadrature rule of its integral form

$$X = \int_{-\infty}^{\infty} (j\omega C + G)^{-1} B B^T (j\omega C + G)^{-H} d\omega \tag{5}$$

where X is the controllability Gramian, and ω is the frequency. The Gramian can be estimated by

$$\bar{X} = \sum_k z_k z_k^H = Z Z^H \tag{6}$$

where $Z = [z_1 \; z_2 \; \ldots]$ and $z_k = \left(j\omega^{(k)} C + G\right)^{-1} B$, with $\omega^{(k)}$ the k^{th} frequency sample. In [9] it was shown that if the quadrature scheme is accurate enough, then the estimated Gramian $\bar{X}$ in (6) converges to the original one X, which implies that the dominant eigenspace of $\bar{X}$ converges to the dominant eigenspace of X (and in fact it converges faster than the Gramian).

Still the technique is not without drawbacks, as it relies on proper choice of sampling points, a non-trivial task in general.

3 Parametric Model Order Reduction

Actual fabrication of physical devices is susceptible to the variation of technological and geometrical parameters due to deliberate adjustment of the process or from random deviations inherent to the manufacturing procedures. This variability leads to a dependence of the extracted circuit elements on several parameters, of electrical or geometrical origin. This dependence results in a parametric state-space system representation, which in descriptor form can be written as

$$\begin{aligned} C(\lambda)\dot{x}(t,\lambda)(\lambda) + G(\lambda)x(t,\lambda) &= Bu(t) \\ y(t,\lambda) &= Lx(t,\lambda) \end{aligned} \tag{7}$$

where the various elements have the meaning described for (1). The elements of the matrices C and G, as well as the states of the system x, depend on a set of Q parameters $\lambda = [\lambda_1, \lambda_2, \ldots, \lambda_Q]$ which model the effects of the mentioned uncertainty. Usually the system is formulated so that the matrices related to the inputs and outputs (B and L) do not depend on the parameters. This time-domain descriptor yields a parametric dependent frequency response modeled via the transfer function

$$H(s,\lambda) = L(sC(\lambda) + G(\lambda))^{-1} B \tag{8}$$

for which we again seek to generate a reduced order approximation, able to accurately capture the input-output behavior of the system for any point in the multidimensional frequency-parameter space.

$$\hat{H}(s,\lambda) = \hat{L}(s\hat{C}(\lambda) + \hat{G}(\lambda))^{-1} \hat{B} \tag{9}$$

In general, one attempts to generate a ROM whose structure is as much similar to the original as possible, i.e. exhibiting a similar parametric dependence and retaining as much of the original structure as possible. In this situation, the generated models can be efficiently combined and used inside simulation environments.

3.1 Representation of the Parametric System

The treatment of the system matrices as appear in (7) is quite inappropriate, as the parameter dependence can vary from element to element inside the matrices, and the reduction methodology will likely not maintain this dependence. For this reason, an approximate representation is generally used as the original system. An affine model based on a Taylor Series expansion can be used for accurately approximating the behavior of the static and dynamic matrices, $G(\lambda)$ and $C(\lambda)$, expressed as a function of the parameters.

$$\begin{aligned} G(\lambda_1,\ldots,\lambda_Q) &= \sum_{k_1=0}^{\infty}\ldots\sum_{k_Q=0}^{\infty} G_{k_1,\ldots,k_Q}\lambda_1^{k_1}\ldots\lambda_Q^{k_Q} \\ C(\lambda_1,\ldots,\lambda_Q) &= \sum_{k_1=0}^{\infty}\ldots\sum_{k_Q=0}^{\infty} C_{k_1,\ldots,k_Q}\lambda_1^{k_1}\ldots\lambda_Q^{k_Q} \end{aligned} \tag{10}$$

where $G_{k_1,\ldots,k_Q}$ and $C_{k_1,\ldots,k_Q}$ are, respectively, the sensitivities of the static and dynamic system matrices. The Taylor series can be extended up to the desired (or required) order, including cross derivatives, for the sake of accuracy.

The techniques here presented can be combined with any order of the Taylor Series in (10). However, for simplicity, in the following a first order approximation, with first order sensitivities and no cross terms, will be used to illustrate the procedure.

$$\begin{aligned} G(\lambda_1,\ldots,\lambda_Q) &= G_0 + G_{\lambda_1}\lambda_1 + G_{\lambda_2}\lambda_2 + \ldots + G_{\lambda_Q}\lambda_Q \\ C(\lambda_1,\ldots,\lambda_Q) &= C_0 + C_{\lambda_1}\lambda_1 + C_{\lambda_2}\lambda_2 + \ldots + C_{\lambda_Q}\lambda_Q \end{aligned} \tag{11}$$

where G_0 and C_0 are the nominal matrices, whereas G_{λ_i} and C_{λ_i} represent the 1^{st} order derivatives of the original matrices with respect to the i^{th} parameter. Under this representation of the parametric system, the structure for parameter dependence may be maintained under a projection scheme, as long as the projection is not only applied to the nominal matrices, but to the sensitivities as well.

$$\begin{aligned} \hat{G}(\lambda_1,\ldots,\lambda_Q) &= \hat{G}_0 + \hat{G}_{\lambda_1}\lambda_1 + \hat{G}_{\lambda_2}\lambda_2 + \ldots + \hat{G}_{\lambda_Q}\lambda_Q \\ \hat{C}(\lambda_1,\ldots,\lambda_Q) &= \hat{C}_0 + \hat{C}_{\lambda_1}\lambda_1 + \hat{C}_{\lambda_2}\lambda_2 + \ldots + \hat{C}_{\lambda_Q}\lambda_Q \end{aligned} \tag{12}$$

where $\hat{C}_0 = V^T C_0 V$, $\hat{G}_0 = V^T G_0 V$, $\hat{C}_{\lambda_i} = V^T C_{\lambda_i} V$, and $\hat{G}_{\lambda_i} = V^T G_{\lambda_i} V$, with V the projector whose columns span the desired subspace basis. This is one of the main features that make the projection-based reduction the procedure followed by most of the parametric model order reduction techniques.

Another important issue is the passivity of the system. Taylor Series is not globally accurate, and, under large parameter variations, can lead to loss of accuracy, and more important, passivity. To avoid such pitfalls, the building of the Taylor Series formulation must be done such that under any expected parameter setting, the system matrices retain

their positive definiteness. A simple scheme for ensuring this is to compute the derivatives element-wise, i.e. for each resistor, capacitor, etc..., consistently with the nominal, and thus under any possible parameter setting the stamping of this value still yields a positive definite matrix. Projection techniques are able to guarantee the passivity of the reduced models under certain circumstances (as pointed in Section 2), usually fulfilled in the case of electric models (for details see [8]).

3.2 Reduction via Multi-Dimensional Moment Matching

Some of the most appealing techniques for the reduction of the parametric systems are multi-dimensional moment matching algorithms. These techniques appear as extensions to nominal moment-matching ones [8, 7, 10]. Moment matching algorithms have gained a well deserved reputation in nominal MOR due to their simplicity and efficiency. The extensions of these techniques to the parametric case feature a similar simplicity. They are usually based in the implicit or explicit matching of the moments of the parametric transfer function (8). These moments depend not only on the frequency, but on the set of parameters affecting the system, and thus are denoted as multi-dimensional moments.

Some schemes, denoted as Multi-Parameter Moment Matching [11], rely on matching, via different approaches, the multi-parameter moments of (8).

$$x(s,\lambda_1,\ldots,\lambda_Q)=\sum_{k=0}^{\infty}\sum_{k_s=0}^{k}\sum_{k_1=0}^{k-k_s}\cdots\sum_{k_Q=0}^{k-k_s-k_1\ldots-k_{Q-1}}M_{k,k_s,k_1,\ldots,k_Q}s^{k_s}\lambda_1^{k_1}\ldots\lambda_Q^{k_Q} \tag{13}$$

where $M_{k,k_s,k_1,\ldots,k_Q}$ is a k-th ($k=k_s+k_1+\ldots+k_Q$) order multi-parameter moment corresponding to the coefficient term $s^{k_s}\lambda_1^{k_1}\ldots\lambda_Q^{k_Q}$. Following the same idea used in the nominal moment matching techniques, a basis for the subspace formed from these moments can be built

$$colspan[V]=colspan\{M_{0,0,0,\ldots,0},M_{1,1,0,\ldots,0},\ldots,M_{k,k_s,k_1,\ldots,k_Q}\} \tag{14}$$

and the resulting matrix V can be used as a projection matrix for reducing the original system. The generated parameterized ROM matches up to the k-th order multi-parameter moment of the original system. The main inefficiencies of these techniques arise from the fact that the same number of moments is matched for all the parameters (including the frequency), and the expansion is performed around a single point. However, it should be noticed that the parameters usually fluctuate in small ranges around their nominal values, whereas the frequency has a much wider range of variation. To match the number of moments required to maintain the accuracy for large frequency ranges may lead to large basis, and thus oversized reduced models. Some schemes have been proposed to cope with this issue [12], but still suffer from some drawbacks.

A slightly different approach, that provides more compact ROMs, is presented in [13], which relies on the computation of several subspaces, built separately for each dimension, i.e. the frequency s and the parameter set λ. Given a parametric system (7), the first step of the algorithm is to obtain the k_s block moments of the transfer function with respect to the frequency when the parameters take their nominal value (for

example, via [8]). These block moments will be denoted as V_s. The next step is to obtain the subspace which matches the k_{λ_i} block moments of x with respect to each of the parameters λ_i (with the values for the rest of the parameters $j \neq i$ fixed to their nominal values), and will be denoted by V_{λ_i}. Once all the subspaces have been computed, an orthonormal basis can be obtained so that its columns span the union of all previously computed subspaces.

$$colspan[V] = colspan\{V_s, V_{\lambda_1}, \ldots, V_{\lambda_Q}\} \tag{15}$$

Applying the resulting matrix in a projection scheme ensures that the parametric ROM matches k_s moments of the original system with respect to the frequency, and k_{λ_i} moments with respect to each parameter λ_i. If the cross-term moments are needed for accuracy reasons, the subspace that spans these moments can be also included by following the same procedure.

Still a different alternative was proposed in [14], where the number of multi-parameter moments matched is increased by applying a two step moment matching scheme. The first step matches the parameter moments explicitly, and a second projection step is applied to capture the frequency moments (for details see [14]). Unfortunately, the parameter dependence is lost and the passivity is not preserved.

Recent approaches [15, 16] were presented to overcome these shortcomings. They rely on a recursive procedure to compute the same moments spanned by the approach in [14]. Basically, the frequency moments of the nominal transfer function are generated, and from these moments, in a recursive fashion, the frequency moments for each parameter moment are also generated. As an example, for first order moment with respect to the parameters

$$\begin{aligned} colspan[V_0] &= colspan\{V_0^0, V_0^1, \ldots, V_0^{\alpha_0-1}\} \\ colspan[V_{\lambda_i}] &= colspan\{V_{\lambda_i}^0, V_{\lambda_i}^1, \ldots, V_{\lambda_i}^{\alpha_i-1}\} \\ V_{\lambda_i}^j &= -(G_0 + s_e C_0)^{-1}(G_{\lambda_i} V_0^j + s_e C_{\lambda_i} V_0^{j-1} + s_e C_0 V_{\lambda_i}^{j-1}) \end{aligned} \tag{16}$$

where V_0 is the basis that allows matching α_0 frequency moments for the nominal system (V_0^j is related to the j^{th} moment of the nominal system with respect to frequency), and V_{λ_i} is the basis that allows matching α_i frequency moments for the first moment of the i^{th} parameter, that is $M_{1,0,\cdots,1,\cdots}$ up to $M_{\alpha_i,\alpha_i-1,\cdots,1,\cdots}$ in Eqn.(13 ($V_{\lambda_i}^j$ is related to the j^{th} frequency moment for the first moment of the i^{th} parameter; see [15, 16] for further details). This adds flexibility as the number of moments to match with respect to each parameter and the frequency can be different. Furthermore, the number of frequency moments generated for each parameter moment can be also different. Both techniques differ in the methodology for selecting the most relevant moments, either by applying sampling on a tree scheme [15], or by generating the moments exhaustively until no rank is added [16]. The moments generated are orthonormalized and applied as an overall basis, V, in a projection scheme on the Taylor Series matrices.

$$colspan[V] = colspan\{V_0, V_{\lambda_1}, \ldots, V_{\lambda_Q}\} \tag{17}$$

These schemes avoid or minimize the growth of the ROM with the number of parameters and moments to match.

3.3 Reduction via Variational PMTBR

A different approach was also proposed that extends the PMTBR [9] algorithm to include variability [17]. This approach is based on the statistical interpretation of the algorithm (see [9] for details) and enhances its applicability. In this interpretation, the approximated Gramian is seen as a covariance matrix for a Gaussian variable, $x(0)$, obtained by exciting the underlying system description with white noise. Rewriting the Gramian from (5) as

$$X_\lambda = \int_{S_\lambda} \int_{-\infty}^{\infty} (j\omega C(\lambda) + G(\lambda))^{-1} BB^T (j\omega C(\lambda) + G(\lambda))^{-H} p(\lambda) d\omega d\lambda \qquad (18)$$

where $p(\lambda)$ is the joint *Probability Density Function* (PDF) of λ in the parameter space, S_λ. Just as in the original PMTBR algorithm, a quadrature rule can be applied in the parameter plus frequency space to approximate the Gramian via numerical computation

$$z_k = z(s = s^{(k)}, \lambda = \lambda^{(k)}) = (s^{(k)} C(\lambda^{(k)}) + G(\lambda^{(k)}))^{-1} B \qquad (19)$$

where z_k is the k^{th} sample, obtained for a frequency value of $s^{(k)}$ and a parameter set $\lambda^{(k)}$ (i.e. $\lambda^{(k)} = [\lambda_1^{(k)} \dots \lambda_Q^{(k)}]$). The sampling scheme can be combined with any representation, i.e. does not require the computation of the sensitivities of the Taylor Series representation as in the case of multi-dimensional moment matching techniques. On the other hand, the Taylor Series representation exhibits further advantages in terms of maintenance of the parametric dependence and reuse, which are useful for efficient use of the reduced models in simulation environments. Note that the accuracy of the resulting ROM does not depend on the accuracy of the approximation of the integral, but on the projection subspace. After the quadrature is performed in the overall variational subspace, the deterministic procedure is followed and the most relevant vectors are selected via *Singular Value Decomposition* (SVD) in order to build a projection matrix

$$Z = [z_1, \dots, z_k, \dots] \longrightarrow VSU = SVD(Z) \qquad (20)$$

where S is the diagonal matrix with the singular values σ_i in its entries, and V and U are the unitary matrices that span the vectors associated with such singular values. The vectors of V whose associated singular values do not fall below a desired tolerance, are used in a congruence transformation on the parametric system matrices (7) (and thus retain the projection-based reduction advantages when applied to a Taylor Series representation). As in the deterministic case, an error analysis and control can be included, via the singular values, but in this variational case, only a bound on the expected error can be given (as we are working with statistical analysis)

$$E\{\|\hat{x}_0 - x_0\|_2^2\} \leq \sum_{i=r+1}^{n} \sigma_i^2 \qquad (21)$$

where r is the reduced order, n the original number of states, and σ_i are the singular values obtained from (20). The complexity and computational cost is generally the same as that of the deterministic PMTBR plus the previous quadrature operations, and, it has been shown that the size of the reduced model is less sensitive to the number of

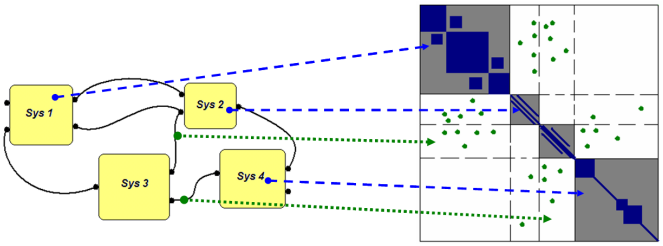

Fig. 1. Illustration of two-level block hierarchy in the system matrix.

parameters in the description, and to how this parameter dependence is modeled. On the other hand, the issue of sample selection, already an important one in the non-parametric context, becomes even more relevant, since the sampling must now be done in a potentially much higher-dimensional space.

4 Block Hierarchical Model Order Reduction

4.1 Structure Preservation

As pointed out, individual blocks inside an RFIC can no longer be treated in isolation, and for this reason the complete system must be treated as an entity. Considering the linear system composed of all these interconnected component blocks including designed-in passives, interconnect, etc, the joint description has an interesting structure, where the diagonal blocks correspond to the individual block matrices, whereas the off-diagonal blocks correspond to the static interconnections (in the G matrix) and dynamic couplings (C matrix), as shown in Figure 1. Standard model order reductions techniques can be applied to this joint, global system and while the resulting reduced model will usually be able to accurately capture the input-output behavior of the complete set of blocks, this approach leads to full reduced matrices. Furthermore, the original two-level hierarchy with interconnections and couplings, in where the individual sub-systems can be recognized, can no longer be recovered (as seen in top of Figure 2).

An alternative approach is to perform the reduction of the individual models in a hierarchical fashion, i.e to reduce each model independently without taking into account the rest of the models or the environment (as seen in bottom of Figure 2). Hence every model is reduced separately and thus the hierarchy and structure of the global system is maintained. However, to apply MOR to each model implies capturing its individual behavior, not the global one. This can be inefficient as too much effort may be spent capturing some local behavior that is not relevant for the global response (maybe filtered by another model). Furthermore certain aspects of the global response might be missed as it is not clear at the component level how relevant they are.

To avoid these problems, one can reduce each component block separately but oriented to capture the global input-output response. This approach will provide us with more control in the reduction stage while also preserving the structure of the interconnections. The transfer function to match is the global one, so the most relevant behavior

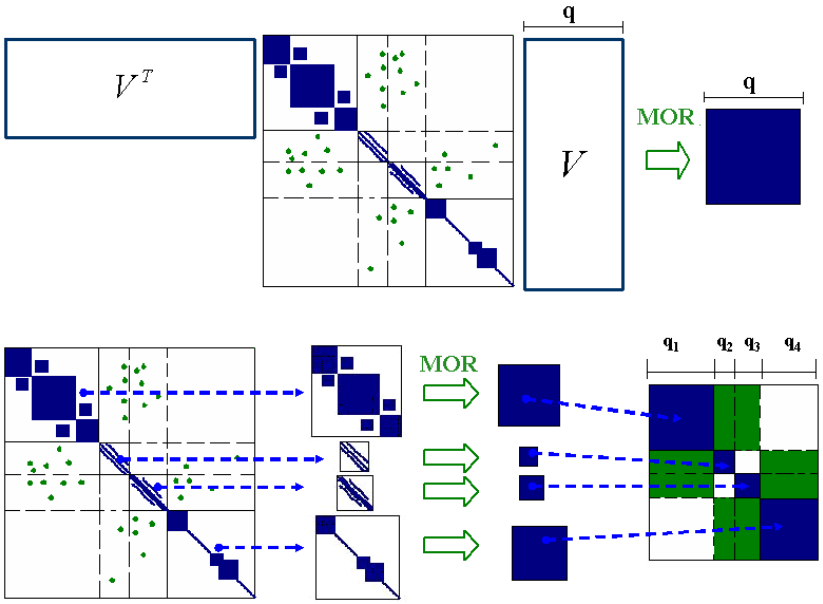

Fig. 2. Illustration of flat reduction in a Block Structured System (Up), and illustration of independent reduction of each system (Down).

for the complete RF system is captured. Hence the global matrices are used in the process of generating the basis for the projector, and thus only the global inputs and outputs of the complete interconnected system are relevant. Therefore, the inefficiencies caused by the large number of ports of the individual component blocks can be avoided. The basis is later used for the reduction of the individual blocks, so the hierarchy can be maintained.

Some recent methods have advocated this approach. In [18] a control theoretic viewpoint of reduction of interconnected systems was presented, but it has the disadvantage that it is unable to treat capacitive couplings, and it is cumbersome to define the interconnections in complex settings. A generalization that overcomes such problems is the *Block Structure Preserving* (BSP) framework, first presented in [19], in which it was applied to separate variables of different nature, and later generalized in [20, 21].

Considering a system composed of N_b sub-systems, the resulting description matrices can be written as

$$G = \begin{bmatrix} G_{11} & \dots & G_{1N_b} \\ \vdots & \ddots & \vdots \\ G_{N_b1} & \dots & G_{N_bN_b} \end{bmatrix} \quad C = \begin{bmatrix} C_{11} & \dots & C_{1N_b} \\ \vdots & \ddots & \vdots \\ C_{N_b1} & \dots & C_{N_bN_b} \end{bmatrix}$$

$$B = \begin{bmatrix} {B_1}^T & \dots & {B_{N_b}}^T \end{bmatrix}^T \quad L = \begin{bmatrix} L_1 & \dots & L_{N_b} \end{bmatrix}. \tag{22}$$

The main idea is to retain the system block structure, i.e. the two-level hierarchy and thus some degree of sparsity, after reduction via projection, allowing for a more efficient

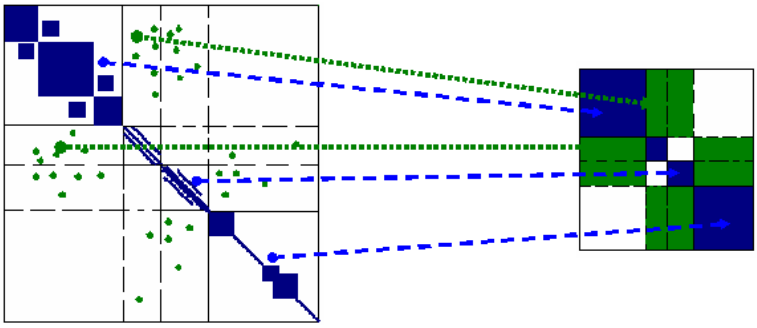

Fig. 3. Illustration of the effect of the Block Structure Preserving reduction.

reduction and use of the reduced model. The procedure relies on expanding the projector of the global system (obtained via any classical MOR projection technique) into a block diagonal matrix, with block sizes equal to the sizes of its N_b individual component blocks (22). A basis that spans a suitable subspace for reduction via projection is then computed (for example a Krylov subspace). The projector built from that basis can be split and restructured into a block diagonal one so that the 2-level structure is preserved under congruence transformation.

$$V = \begin{bmatrix} V_1 \\ \vdots \\ V_{N_b} \end{bmatrix} \qquad \breve{V} = \begin{bmatrix} V_1 & & \\ & \ddots & \\ & & V_{N_b} \end{bmatrix} \tag{23}$$
$$V = colsp\left[Kr\{A,R,q\}\right] \rightarrow colspan\left[V\right] \subset colspan\left[\breve{V}\right]$$

where $Kr\{A,R,q\}$ is the q column Krylov subspace of the complete system ($A = G^{-1}C$ and $R = G^{-1}B$). The block-wise congruence transformation is (see Figure 3)

$$\hat{G}_{ij} = V_i^T G_{ij} V_j \qquad \hat{C}_{ij} = V_i^T C_{ij} V_j \qquad \hat{B}_i = V_i^T B_i \qquad \hat{L}_j = L_j V_j \tag{24}$$

It should be noticed that the above projection matrix $\breve{V}$ has N_b (number of blocks) times more columns than the original projector. This leads to an N_b times larger reduced system. On the other hand, this technique maintains the block structure of the original system and gives us some flexibility when choosing the size of the reduced model depending on the block layout and relevance. The reduced system will be able to match up to N_b times q block moments of the original complete transfer function (see [20] for details) under the best conditions (i.e. with very weak entries in the off-diagonal blocks). Under the worst conditions, only q block moments are matched, i.e. the same number than in the *flat* reduction.

This technique is applicable to the global system, composed of the individual blocks and their connections (including both resistive as well as capacitive or inductive couplings between the blocks). The BSP technique therefore preserves the block structure of the system. However, the inner structure of the blocks themselves is lost since the procedure turns any non-empty block in the original system into a full block, but it is still possible to identify the blocks and relate them to the original device or interaction

block. Nevertheless, if any block is empty in the global system matrix, it remains empty after reduction, increasing the sparsity.

4.2 PMTBR in Block Structure MOR

Any projection-based MOR procedure can be extended in the BSP manner to maintain the hierarchical structure of a system. In the case of the PMTBR algorithm, additional characteristics of the procedure can be further taken advantageous of in the current framework.

If the system has some internal structure, then the matrix Z that is computed from the vector samples of the global system can be split into blocks. The estimated Gramian can be written block-wise as

$$\begin{bmatrix} Z_1 \\ \vdots \\ Z_{N_b} \end{bmatrix} \rightarrow ZZ^H = \begin{bmatrix} Z_1Z_1^H & \dots & Z_1Z_{N_b}^H \\ \vdots & \ddots & \vdots \\ Z_{N_b}Z_1^H & \dots & Z_{N_b}Z_{N_b}^H \end{bmatrix} = \bar{X} \tag{25}$$

But if we expand the matrix Z into diagonal blocks

$$\check{Z} = \begin{bmatrix} Z_1 & & \\ & \ddots & \\ & & Z_{N_b} \end{bmatrix} \rightarrow \check{Z}\check{Z}^H = \begin{bmatrix} Z_1Z_1^H & & \\ & \ddots & \\ & & Z_{N_b}Z_{N_b}^H \end{bmatrix} = \check{X}. \tag{26}$$

From (25) it can be seen that $Z_iZ_i^H = \bar{X}_{ii}$, i.e. the matrix $\check{X} = \check{Z}\check{Z}^H$ is a block diagonal matrix whose entries are the block diagonal entries of the matrix $\bar{X}$. Under a good quadrature scheme, the matrix $\bar{X}$ converges to the original X, and therefore $\check{X}$ will converge to the block diagonals of X. This means that the dominant eigenspace of $\check{X}$ converges to the dominant eigenspace of the block diagonals of X. We can then apply an SVD to each block of the Z matrix

$$Z_i = V_iS_iU_i \;\rightarrow\; \check{X}_{ii} = \bar{X}_{ii} = V_iS_i^2V_i^T \tag{27}$$

where S_i is real diagonal, and V_i and U_i are unitary matrices. The dominant eigenvectors of V_i in 27 corresponding to the dominant eigenvalues of S_i, can be used as a projection matrix in a congruence transformation over the system matrices for model order reduction. The elements of S_i can also be used for *a priori* error estimation in a way similar to how *Hankel Singular Values* are used in TBR procedures. Of course, the convergence of these singular values, and therefore the error bounds, depends on the *strength* of the coupling and the interconnections, but it is supposed that the impact of the systems (placed in the block diagonals) in the global behavior *dominates* the impact the couplings and interconnections may have in such global behavior. Using these block projectors V_i, a structure preserving projector for the global system can be built (23) which will capture the most relevant behavior of each block (revealed by the SVD) with respect to the global response (recall that Z is composed of sample vectors of the complete system). This approach provides us with more flexibility when reducing a complete system composed of several blocks and the interactions between them, as it allows to control the reduced size of each device via an error estimation on the global response.

5 Parametric Block Structure MOR

From the two-level hierarchical description of a system it is possible to have some extra block information that allows us to perform a more efficient MOR. But the behavior of the individual blocks that compose the system is subject to the effect of process variations, both geometrical and electrical. Such variations, as previously pointed out, also affect the interactions and couplings between these blocks. Any system-wide EM simulations must address these effects. Therefore, the variability study must be done over the complete system, and after model generation, a two-level parametric system will be obtained, with the block matrices in the block diagonals and the interactions between them in the off-diagonals. All these blocks will be functions of the relevant process and geometrical parameter.

$$G = \begin{bmatrix} G_{11}(\lambda_{\{11\}}) & \dots & G_{1N_b}(\lambda_{\{1N_b\}}) \\ \vdots & \ddots & \vdots \\ G_{N_b1}(\lambda_{\{N_b1\}}) & \dots & G_{N_bN_b}(\lambda_{\{N_bN_b\}}) \end{bmatrix} \quad C = \begin{bmatrix} C_{11}(\lambda_{\{11\}}) & \dots & C_{1N_b}(\lambda_{\{1N_b\}}) \\ \vdots & \ddots & \vdots \\ C_{N_b1}(\lambda_{\{N_b1\}}) & \dots & C_{N_bN_b}(\lambda_{\{N_bN_b\}}) \end{bmatrix}$$

$$B = \begin{bmatrix} {B_1}^T & \dots & {B_{N_b}}^T \end{bmatrix}^T \qquad L = \begin{bmatrix} L_1 & \dots & L_{N_b} \end{bmatrix} \tag{28}$$

where $\lambda_{\{ij\}}$ represents the subset of parameters affecting block G_{ij} in (28) (it is supposed that some parameters are local, and thus only affect some localized blocks). From (28) is clear that we have a parametric system depending on $\lambda = \bigcup_{i=1,j=1}^{N_b} \lambda_{\{i,j\}}$. Therefore we can apply parametric MOR reduction. Note that any parameter affecting several blocks (diagonal blocks and their interactions) is treated as a single parameter (this avoids the treatment of the same parameter affecting different systems as several different ones).

In this circumstances, BSP techniques can be applied in order to maintain the system structure. This is possible as long as the selected pMOR technique is based on a projection scheme, which is the case for most of the existing procedures (as already presented in Section 3). The extension is very simple: obtain a suitable basis for projection from the ***complete system***, and then split and expand it into a block structure preserving projector. If the basis spans the most relevant behavior of the parametric system, then the expanded BSP projector will capture those as well.

All the advantages and disadvantages mentioned in Section 4 hold here. But there is an extra and important advantage in the parametric case: **the BSP technique maintains the block parametric dependence**, i.e. if a block C_{ij} depends on a subset of parameters $\lambda_{\{ij\}}$, then the reduced block $\hat{C}_{ij} = V_i^T C_{ij} V_j$ will depend on the same parameter subset and no other. This fact has inherent advantages in terms of storage and use of the sensitivities, as the reduced sensitivities are even sparser than the nominal matrices. On the other hand, as previously discussed, some pMOR algorithms yield very large ROMs, and therefore their combination with BSP techniques will lead to an extremely large ROM.

However, it was shown in Section 3.3 that the ROM sizes obtained with the Variational PMTBR method are usually less sensitivity to the number of parameters, and such method is an extension of the PMTBR framework to handle parametric systems;

Algorithm I: Block Structure Preserving VPMTBR

Starting from a Block Structured System C, G, B, L with N_b blocks:

1: Select a quadrature rule of K points in the space $[s,\ \lambda]$

2: For each point compute: $z_k = \left(s^{(k)}C(\lambda^{(k)}) + G(\lambda^{(k)})\right)^{-1} B$

3: Form the matrix columns $Z = [z_1 \dots z_K]$

4: Split it into N_b blocks, according to the system structure

$$Z = \begin{bmatrix} Z_1 \\ \vdots \\ Z_{N_b} \end{bmatrix}$$

5: For each block Z_j obtain the SVD: $Z_j = V_j S_j U_j$

6: For each matrix V_j drop the columns whose singular values falls bellow the desired global tolerance

7: Build a Block Structure Preserving Projector from the remaining columns

$$\breve{V} = \begin{bmatrix} V_1 & & \\ & \ddots & \\ & & V_{N_b} \end{bmatrix}$$

8: Apply $\breve{V}$ in a congruence transformation on the Block Structured System C, G, B, L

the main difference is that the sampling scheme for obtaining the matrix whose columns span the desired subspace is extended to the multidimensional space of the parameters and the frequency, the rest of the procedure being exactly the same.

For this reason, the Variational PMTBR framework can be easily extended and combined with the BSP methodology, by direct use of the technique presented in Section 4.2 in the variational case. The advantages of the block size control and error estimation provided in such case are still valid, although in this case, as in [17], only a bound on the expected error can be given. This block-wise control is very useful when the various component models of a complete system have very different relevant rank: if the same ROM size is applied to every block, the reduction may grow unnecessarily large. In contrast, the complexity of the proposed methodology is exactly the same as that for the non-structure-preserving techniques. The only difference is that the SVD (or orthonormalization in the moment matching approaches) must be done block-wise in order to avoid numerical errors (e.g. the expansion of a orthogonal matrix to a block diagonal does not guarantee the orthogonality of this new basis). This can become an advantage, because for some blocks the number of vectors needed is lower, so less computational effort is required in orthonormalization steps.

6 Results

To illustrate the proposed procedure we present results from two examples to which several pMOR techniques were applied. These include [17] denoted as VPMTBR, [13]

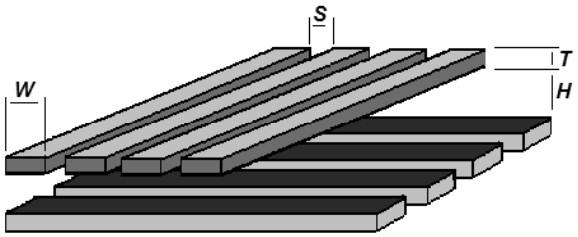

Fig. 4. Bus topology for Example 1.

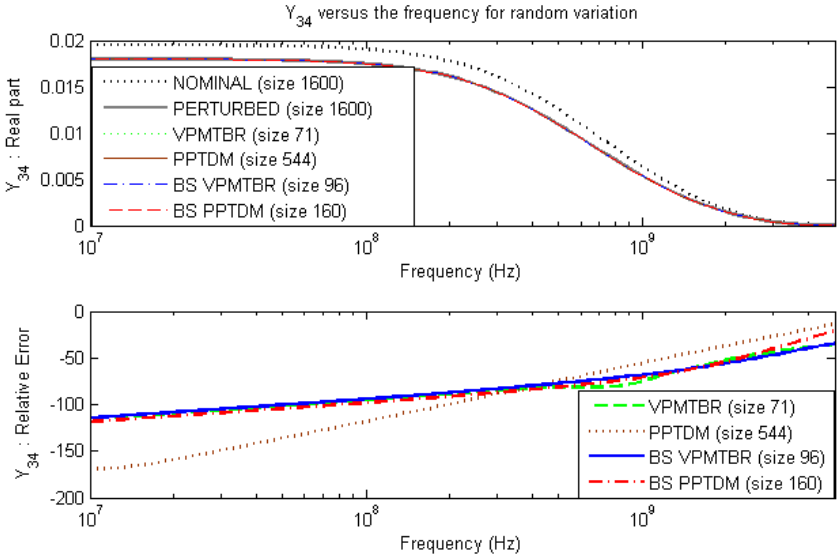

Fig. 5. (Up) Y_{34} versus the frequency for Example 2 for the nominal, pertubed and parametric ROMs with random parameter variation set. (Down) Relative Error of the ROMs w.r.t. the perturbed response.

denoted as PPTDM, and two Block Structure preserving methods: *Algorithm I*, denoted as BS VPMTBR, and block structure based on [13], denoted as BS PPTDM. The non-reduced model response will be denoted as Original or Perturbed, depending on whether a parameter variation has been applied.

6.1 Example 1 - Coupled Buses

This example, depicted in Figure 4, is composed of 16 blocks: 2 buses of 8 parallel lines each (each line modeled as an RC ladder of 100 segments) are on different metal layers, and cross at a square angle. The inputs and outputs are taken at the edges of each line of the first bus, so the system will have 16 ports. In this case there is no interconnection, just coupling effects. Each line is assumed coupled to the previous and the next line of their bus, and to every line of the other bus in the crossing area. Each

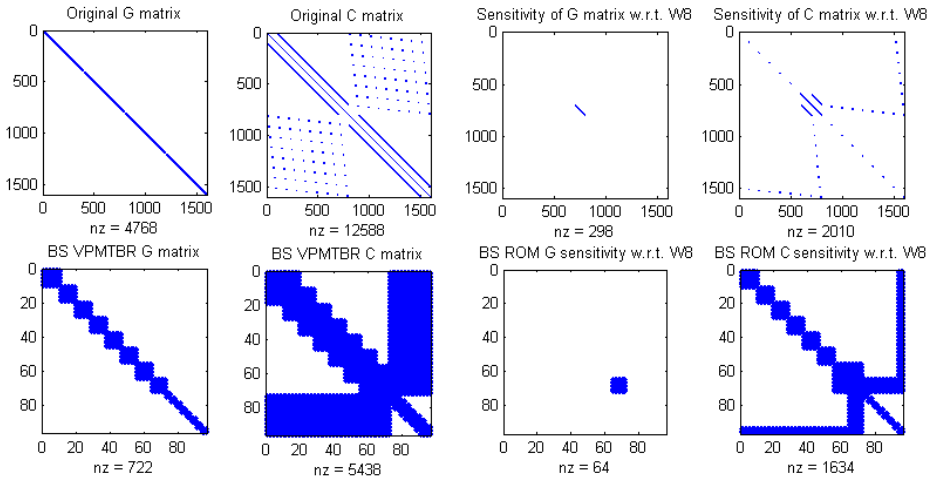

Fig. 6. Example 1 - Structure of the matrices for the original system (up), and the ROM obtained with BS VPMTBR (down). Note the different dimensions (nz is the number of nonzero elements in matrix).

Table 1. Characteristics of the pMOR methods applied in Example 1

MOR Method	Size	NNZ (G C)	Sparsity Ratio
NONE	1600	4768 12588	0.0018 0.0049
VPMTBR	71	5041 5041	1.000 1.000
PPTDM	544	295936 295936	1.000 1.000
BS VPMTBR	96	722 5438	0.078 0.590
BS PPTDM	160	1600 17200	0.062 0.672

line has its width (W) as a parameter, which implies 16 independent parameters. The width variation affects the line model, as well as the in-bus coupling (width variation also affects the interline spacing), and the inter-bus coupling (the crossing area varies).

Figure 5 shows the frequency response of the nominal system, the pertubed response of the non-reduced system, and the responses of ROMs for VPMTBR, PPTDM, BS VPMTBR and BS PPTDM. Again, the main characteristics of the resulting ROMs are shown in Table 1. The PPTDM based algorithms result in very large ROMs even for small number of moments to match (2 w.r.t. the frequency and 2 w.r.t. each parameter). For these reasons each block moment from PPTDM was truncated to 10 vectors to keep the size manageable (otherwise no reduction would be possible). While this seems to produce acceptable results, there is little control over the result. On the other hand, the PMTBR based techniques leads to more compressed ROMs, as the SVD reveals the most relevant vectors. In the case of the BS VPMTBR, the control of each block allows different reduction sizes for each bus: since the ports of the 2^{nd} bus are not taken into account, less effort is needed to capture its behavior. In fact, the models for the 1^{st} bus

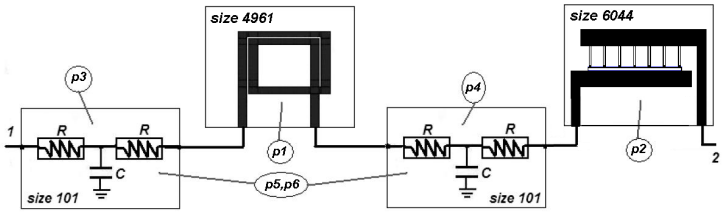

Fig. 7. Interconnection scheme for Example 2, with original sizes and parameter indication.

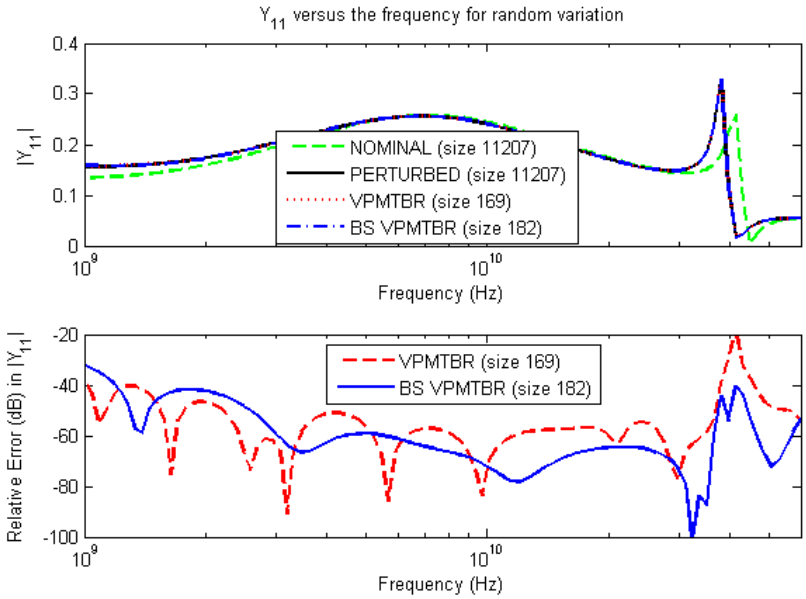

Fig. 8. Example 2 - (Up) Magnitude in dB of Y_{11} versus the frequency of Example 1 for the nominal, the pertubed and the parametric ROMs for a random parameter variation. (Down) Relative Error (in dB) of Y_{11} for the ROMs w.r.t. the perturbed response.

are of sizes 8 to 10, while models for the 2^{nd} bus are all size 3. The ability to control reduction locally is clearly an advantage of the method. The effect of this control can be seen in Figure 6, which shows the structure for the original (nominal and one sensitivity) matrices (up), and the structure obtained with BS VPMTBR reduction.

6.2 Example 2 - EM based models

The second example system is composed of four blocks: a *Multiple Input Multiple Output* (MIMO) RC ladder of size 101, with 2 ports, a MIMO EM based model of a planar Spiral Inductor of size 4961, with 2 ports, another RC ladder of size 101 and 2 ports, and an MIMO EM-based model for a metal-insulator-metal (MIM) capacitor. The four systems are connected in series as shown in Figure 7, so the global inputs and outputs are taken in the first port of the first RC and the second port of the CMIM model.

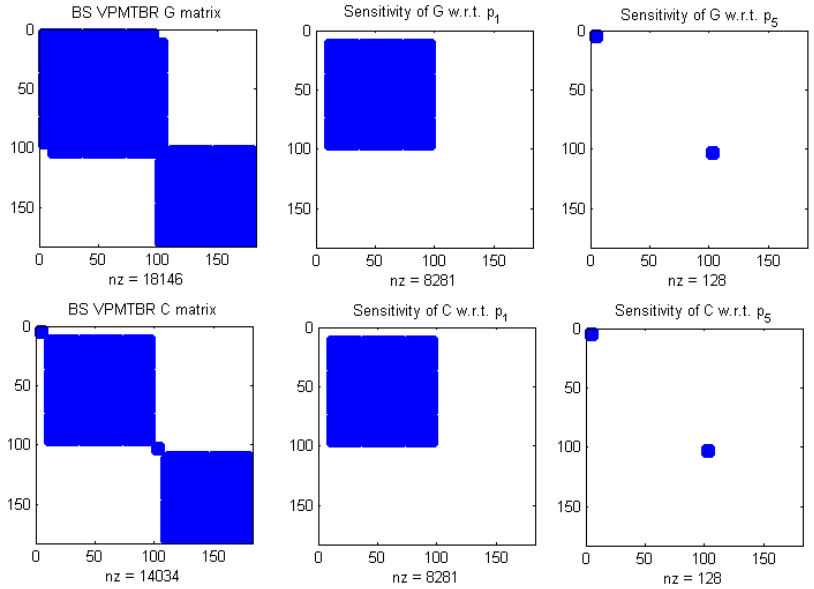

Fig. 9. Example 2 -(Up) Structure of the nominal G matrix (left), and the structure of the sensitivities of G w.r.t. p_1 (centre, affecting the block related to the spiral), and w.r.t. p_5 (right, affecting the RC ladders) for the BSP-based reduction. (Down) Same matrices, but for the dynamic part.

Table 2. Example 2 - Characteristics of the pMOR methods applied

MOR Method	Size (Blocks)	NNZ (G C) (Sparsity Ratio)	Max. RE	Generation Cost
NONE	11207 $(101, 4961, 101, 6044)$	49305 13708 (0.00039 0.00011)	0	*none*
VPMTBR	169 (169)	28561 28561 (1.00 1.00)	$-18.7dB$	$90 Samples(s+\lambda)$ $SVD(n \times 169)$
BS VPMTBR	182 $(8, 91, 8, 75)$	18146 14034 (0.55 0.42)	$-31.9dB$	$90 Samples(s+\lambda)$ $SVD(n \times \{8, 91, 8, 75\})$

The system depends on six parameters, affecting different blocks. Figure 8 shows the frequency response of the self-admittance Y_{11} of the nominal system, the pertubed response of the non-reduced system, and the responses of the PMTBR-based models (the PPTDM and BS PPTDM models do not produce competitive results sizewise, and therefore were omitted). Table 2 shows the main characteristics of the obtained ROMs. The BS VPMTBR yields a slightly bigger ROM, but it maintains the block structure, both of the nominal matrices and the sensitivities (see Figure 9), of the original system, and is able to control the size of each reduced block depending on its relevance on the global response. Furthermore, the block parameter dependence is clearly maintained. Figure 9 shows the structure of the matrices obtained with BS VPMTBR, for the nominal matrices (left, G up and C down), where the effect of the block order control can be seen,

and for two sensitivities, one affecting only the block related to the Spiral (centre), and other affecting the RC ladders (right), which are extremely sparse. On the other hand, the *flat* reduction via VPMTBR yields full matrices, both for the nominal and the sensitivities. The accuracy is also better for the BSP based approach, and the procedure requires similar computational effort.

7 Conclusion

In this paper we have presented a block structure-preserving parametric model order reduction technique, as an extension of existing parametric MOR techniques, in order to improve the reduction when a two-level hierarchical structure is available in the system description. This type of structure is common in coupled or interconnected systems, and can lead to simulation advantages. The methodology presented here is general in the sense that it can be used with any projection parametric MOR technique to maintain the two-level hierarchy and the block-parameter dependence. The presented extension of the PMTBR-based procedures into the Block Structure Preserving framework, allows more control on the reduction, provided by the inclusion of estimated error bounds on the single blocks oriented to the global response.

Acknowledgment

The authors gratefully acknowledge the support from the EU, and would like to thank the CHAMELEON RF consortium, and in particular Wim Schoenmaker (MAGWEL), Nick van der Meijs and Kees-Jan van der Kolk (TU Delft), and Daniel Ioan and Gabriela Ciuprina (PU Bucharest) for many helpful discussions, and for providing some of the simulation examples.

References

1. Antoulas, A.C.: Approximation of Large-Scale Dynamical Systems. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA (2005)
2. Moore, B.: Principal Component Analysis in Linear Systems: Controllability, Observability, and Model Reduction. IEEE Transactions on Automatic Control **AC-26** (1981) 17–32
3. Phillips, J., Daniel, L., Silveira, L.M.: Guaranteed passive balancing transformations for model order reduction. In: 39^{th} ACM/IEEE Design Automation Conference, New Orleans, Louisiana (2002) 52–57
4. Li, J.R., Wang, F., White, J.: Efficient model reduction of interconnect via approximate system grammians. In: International Conference on Computer Aided-Design, San Jose, CA (1999) 380–383
5. Jaimoukha, I.M., Kasenally, E.M.: Krylov subspace methods for solving large Lyapunov equations. SIAM Journal on Numerical Analysis **31** (1994) 227–251
6. Kamon, M., Wang, F., White, J.: Generating nearly optimally compact models from Krylov-subspace based reduced-order models. IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing **47** (2000) 239–248

7. Feldmann, P., Freund, R.W.: Efficient linear circuit analysis by Padé approximation via the Lanczos process. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems **14** (1995) 639–649
8. Odabasioglu, A., Celik, M., Pileggi, L.T.: PRIMA: passive reduced-order interconnect macromodeling algorithm. IEEE Trans. Computer-Aided Design **17** (1998) 645–654
9. Phillips, J.R., Silveira, L.M.: Poor Man's TBR: A simple model reduction scheme. IEEE Trans. Computer-Aided Design **24** (2005) 43–55
10. Elfadel, I.M., Ling, D.L.: A block rational arnoldi algorithm for multipoint passive model-order reduction of multiport rlc networks. In: International Conference on Computer Aided-Design, San Jose, California (1997) 66–71
11. Daniel, L., Siong, O.C., Low, S.C., Lee, K.H., White, J.K.: A multiparameter moment-matching model-reduction approach for generating geometrically parametrized interconnect performance models. IEEE Trans. Computer-Aided Design **23** (2004) 678–693
12. Li, P., Liu, F., Li, X., Pileggi, L., Nassif, S.: Modeling interconnect variability using efficient parametric model order reduction. In: Proc. Design, Automation and Test in Europe Conference and Exhibition. (2005)
13. Gunupudi, P., Khazaka, R., Nakhla, M., Smy, T., Celo, D.: Passive parameterized time-domain macromodels for high-speed transmission-line networks. IEEE Trans. On Microwave Theory and Techniques **51** (2003) 2347–2354
14. Li, X., Li, P., Pileggi, L.: Parameterized interconnect order reduction with Explicit-and-Implicit multi-Parameter moment matching for Inter/Intra-Die variations. In: International Conference on Computer Aided-Design, San Jose, CA (2005) 806–812
15. Zhu, Z., Phillips, J.: Random sampling of moment graph: a stochastic krylov-reduction algorithm. In: Proc. Design, Automation and Test in Europe Conference and Exhibition, Nice, France (2007) 1502–1507
16. Li, Y.T., Bai, Z., Su, Y., Zeng, X.: Parameterized model order reduction via a two-directional arnoldi process. In: International Conference on Computer Aided-Design, San Jose, CA, USA (2007) 868–873
17. Phillips, J.: Variational interconnect analysis via PMTBR. In: International Conference on Computer Aided-Design, San Jose, CA, USA (2004) 872–879
18. Vandendorpe, A., Dooren, P.V.: Model reduction of interconnected systems. In: Proc. of 16th International Symposium on Mathematical Theory of Networks and Systems (MTNS 2004), Leuven, Belgium (2004) THP3–4
19. Freund, R.W.: Sprim: Structure-preserving reduced-order interconnect macro-modeling. In: International Conference on Computer Aided-Design, San Jose, CA. U.S.A (2004) 80–87
20. Yu, H., He, L., Tan, S.X.D.: Block structure preserving model order reduction. In: BMAS - IEEE Behavioral Modeling and Simulation Wokshop. (2005) 1–6
21. Li, R.C., Bai, Z.: Structure-preserving model reduction using a krylov subspace projection fomulation. Comm. Math. Sci. **3** (2005) 179–199