MODEL ORDER REDUCTION: AN ADVANCED, EFFICIENT AND AUTOMATED COMPUTATIONAL TOOL FOR MICROSYSTEMS

T. BECHTOLD, A. VERHOEVEN, E. J. W. TER MATEN
NXP Semiconductors, Research, DMS - Physical Design Methods, High Tech Campus 48, 5656 AE Eindhoven, The Netherlands
E-mail: {Tamara.Bechtold,Arie.Verhoeven,Jan.ter.Maten}@NXP.com

T. VOSS
Delft university of Technology, Mekelweg 2, 2628 CD Delft, The Netherlands
E-mail: {T.Voss@tudelft.nl

The goal of mathematical model order reduction (MOR) is to replace the non-automatic compact modeling, which is the state of the art in simulation flow of microelectronic and micro-electro- mechanical systems (MEMS). MOR offers a possibility of automatically creating small but very accurate models which can be used within system level simulation. The main challenges in integrating model order reduction as a standard tool in the current simulation flow are to be able to reduce non-linear ordinary differential equation systems (ODEs) and differential algebraic equation systems (DAEs), which arise from either spatial discretization of partial differential equations (PDEs) or from electronic circuit equations. We present a methodology for applying MOR to linear and non-linear ODEs and DAEs and numerical results for several MEMS and microelectronic devices.

Keywords: Model Order Reduction; Microsystem Simulation

1. Introduction

The decreasing size and growing complexity of micro-electronic systems impose new challenges on the designers and simulation tools [1]. The main requirement on modern simulation tools for microsystems is the automatic macromodeling with very high accuracy. In Fig. 1 (left) an ”ideal” design flow is shown, which is at present hindered by the lack of modeling continuity.

Luckily, modern mathematical methods, known as model order reduction [2], show most promising results to-wards filling this ”gap”. The main advantage of mathematical model order reduction over ”classical compact modeling” is that it is formal, robust and can be automatized. Model order reduction starts with either an ODE or an DAE system (see Fig. 1 (right)). Numerical solution of time-dependent
2. Principle of Model Order Reduction

Both, the ODE systems, which arise from spatial discretization of 1st order time-dependent PDEs, and the DAE systems, which describe the dynamics of electrical
circuits at time $t$ can be described by:

$$\frac{d}{dt} q(x, t) + j(x, t) + B u(t) = 0,$$

(1)

with the difference that for “real” DAEs the partial derivative $q_x$ is singular. In case of circuit equations for example, the vector-valued functions $q(x, t)$ and $j$ represent the contributions of respectively, reactive elements (such as capacitors and inductors) and of nonreactive elements (such as resistors) and all time-dependent sources are stored within $u(t)$. In case of linear or linearized models, (1) simplifies to:

$$C \cdot \dot{x} + G \cdot x = B \cdot u(t)$$

$$y = L \cdot x$$

(2)

where $x(t) \in \mathbb{R}^n$ is the state vector, $u(t) \in \mathbb{R}^m$ is the input excitation vector and $y(t) \in \mathbb{R}^p$ is the output measurement vector. $G, C \in \mathbb{R}^{n \times n}$ are the symmetric and sparse system matrices and for DAEs, $C$ is singular. $B \in \mathbb{R}^{n \times m}$ and $L \in \mathbb{R}^{p \times n}$ are the user-defined input and output distribution arrays. $n$ is the dimension of the system and $m$ and $p$ are the numbers of inputs and outputs. The idea of model order reduction is to replace (2) by the system of the same form but with much smaller dimension $r \ll n$:

$$C_r \cdot \dot{z} + G_r \cdot z = B_r \cdot u(t)$$

$$y_r = L_r \cdot z$$

(3)

which can be solved by suitable DAE or ODE solver and will approximate the input/output characteristics of (2). A transition from (2) to (3) is formal and is done in two steps. The first is the transformation of the state vector $x$ to the vector of generalized coordinates $z$ and truncation of a number of those generalized coordinates, which leads to some (hopefully) small error $\epsilon$:

$$x = V \cdot z + \epsilon$$

(4)

The time-independant $V \in \mathbb{R}^{n \times r}$ is called the transformation or projection matrix, as (4) can be seen as the projection of the state vector onto some low-dimensional subspace defined by $V$. Note that the spatial and physical meaning of $x$ is lost during such projection. In the second step, (2) is multiplied from the left hand side with another matrix $W^T \in \mathbb{R}^{r \times n}$, so that $C_r = W^T C V_r$, $G_r = W^T G V_r$, $B_r = W^T B$ and $L_r = E V$. Note that the number of inputs and outputs in the reduced system (3) is the same as in (2). The model order reduction process is schematically shown in Fig. 2.
The above principle of projection can also be applied directly to the second order ODE systems [19], which may arise from spatial discretization of the 2nd order time-dependent PDEs (e.g. the equation of motion, which is often solved in MEMS simulation). Furthermore, it can be applied to the nonlinear system (1). This, however, is much more complicated and does not necessarily result in reduction of computational time, as will be shown in the following chapters.

3. Methods

In Fig. 3 a number of state-of-the-art methods for model order reduction of (1) is displayed.

3.1. Linear MOR

The control theory methods (mostly used are Truncated Balanced Realization, Hankel Norm Approximation and Singular Perturbation Approximation) offer a
good mathematical basis and a global error estimate. For MOR of smaller-size linear ODE systems they have been successfully used since many years [18]. Unfortunately, their computational complexity is of $\mathcal{O}(n^3)$, as for the construction of $V$ and $W$ the singular value decomposition (SVD) of the large-scale system matrices is required. Their generalization to DAEs has been explored in [7].

In order to overcome the pore scaling of SVD-based methods, Krylov subspace methods (also known as moment matching methods) [5] are mostly used for MOR of large-scale linear ODEs and DAEs. They are based on writing down the transfer function of (2) in the Laplace-domain, developing it into Taylor series around a frequency point of interest (single or multiple frequency points might be chosen for “classical” or rational Krylov methods) and truncating the terms of the higher order. The left over terms define a reduced model’s rational transfer function ($r$ terms define a reduced model of order $r$ for one-sided projections and of order $2r$ for two-sided projections). The Taylor coefficients are called moments of the transfer function and can not be explicitly computed in a stable way. Rather, the bases for the Krylov subspaces are computed and stored within $V$ and $W$. The main disadvantage of Krylov methods is the lack of global error estimate and the fact that the preservation of stability and passivity of the reduced model is not guaranteed in general. In order to overcome this bottle-necks without loosing the speed-up, a research is going on, on how to combine the advantages of the SVD-based and Krylov-based methods.

An SVD-Krylov method [9] is based on iteratively updating the interpolation frequencies (and so the transformation matrices $V$ and $W$) for the rational Krylov method by using the eigenvalues of the reduced model in each iteration. After convergence, a stability of the reduced model is guaranteed, as well as it’s error minimality. The Poor man’s TBR has been proposed in [8]. It is based on the singular value decomposition of the projection matrix $V$, which is computed by rational Krylov. In this way, an additional reduction of model size compared to rational Krylov is achieved and at the same time the error estimate property of TBR is inherited. Low-rank Gramian approximation aims at speeding up the solution of the sparse generalized Lyapunov equations. Hereby, the Gramians are approximated through the low-rank matrices. In [6] the applicability of balanced truncation on parallel computers was extended to sparse systems with up to $\mathcal{O}(10^9)$ states. The method inherits the preservation of stability and passivity and the global error bound of “classical” TBR. Fig. 4 summarizes the properties of the available algorithms.
Control theory methods (SVD-based)
- global error estimate
- preservation of stability and passivity
- fully automatic
- computational effort $O(n^3)$

Moment matching (Krylov-based)
- low computational effort
- no global error estimate
- manual selection of $r$
- preservation of stability and passivity only in some cases

Hybrids (SVD-Krylov poor man’s TBR low rank gramians)
- global error estimate
- computational effort $O(n^2)$
- preservation of stability and passivity
- still under development

Advantages Disadvantages

---

Fig. 4. Methods for model order reduction of linear dynamic systems.

3.2. Nonlinear MOR

We said that the goal of model order reduction is to produce a lower dimensional system that has approximately the same response characteristics for all the inputs as the original system. For the linear systems of the form (2) and their reduced systems (3) this was the case. However, the direct projection of the general non-linear system (1) leads at present mostly to the reduced models, which approximate the original one only for a certain single input. Furthermore, the extraction of the reduced order model requires the simulation of the original one. Sometimes, the simulation of the reduced model might require even longer CPU time than the simulation of the full-scale model, which is of course, not what we want. Hence, the reduced non-linear models are at best, meaningful for re-use.

The idea behind Proper Orthogonal Decomposition (POD) [11] is to directly project the original nonlinear system (1) onto some subspace with smaller dimension. As this, however, does not lead to the reduction of the computational time, Missing Point Estimation technique (MPE) [12] is used to speed up the simulation.

Empirical balanced truncation [15] is an extended version of a POD method. Instead of creating the reduced subspace with only one relevant input and initial state, several training trajectories are created and the reduced subspace is built in a similar way as in the BTR method.

The idea behind the Trajectory Piecewise Linear (TPWL) [10] method is to linearise (1) several times along a training trajectory (corresponding to some typical input). The local linear reduced systems (can be created with arbitrary linear
MOR method) are then used to create a global reduced subspace. The final TPWL model is constructed as a weighted sum of all local linearised reduced systems.

The idea behind Volterra series [14] is to construct a bilinear system, which approximates the first moments of the nonlinear system. Then linear model reduction techniques are used to create a reduced bilinear system which matches as many moments of the original system as possible.

Fig. 5 summarizes the properties of the available algorithms. In [16] a review of some additional methods is given.

<table>
<thead>
<tr>
<th>Nonlinear MOR methods</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>POD</td>
<td>high accuracy</td>
<td>no speed up (MPE can help)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>no global error estimate</td>
</tr>
<tr>
<td>Empirical Balanced Truncation</td>
<td>god approximation</td>
<td>most expensive model extraction</td>
</tr>
<tr>
<td></td>
<td>different inputs/initial values possible</td>
<td>no speed-up (MPE can help)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>no global error estimate</td>
</tr>
<tr>
<td>TPWL</td>
<td>cheap reduced model evaluations</td>
<td>high memory consuming</td>
</tr>
<tr>
<td></td>
<td></td>
<td>bad accuracy for highly nonlinear systems</td>
</tr>
<tr>
<td>Volterra Series</td>
<td>moment matching</td>
<td>increased dimension of the state vector</td>
</tr>
<tr>
<td></td>
<td>full-system simulation is not necessary</td>
<td>not applicable to DAEs</td>
</tr>
</tbody>
</table>

Fig. 5. Methods for model order reduction of nonlinear dynamic systems.

4. Applications

The test-cases below are academic examples of electrical circuits, which closely resemble the industry-relevant models. They have been defined as MOR case studies for the European project COMSON [13].

4.1. Transmission Line (linear DAE system)

Fig. 6 shows an academic model of transmission line. This very simple model has been chosen, because it resembles the interconnect modeling and can be effectively used for testing new MOR algorithms. It consists of scalable number of RLC ladders and after modified nodal analysis results in a linear DAE system of the form (2).
4.2. Diode Chain (nonlinear DAE system)

Fig. 7 shows an academic, highly non-linear model of a diode chain. It consists of scalable number of diodes and is described by the following equations:

\[
\begin{align*}
V_1 - U_{in}(10^9t) &= 0, \\
i_s - g(V_1, V_2) &= 0, \\
g(V_1, V_2) - g(V_2, V_3) - CV_2 - \frac{1}{R}V_2 &= 0, \\
&\quad \left\{ \begin{array}{ll}
g(V_2, V_3) = \frac{V_2 - V_3}{R} - 1 & \text{if } V_2 - V_3 > 0.5 \\
0 & \text{otherwise}
\end{array} \right.
\end{align*}
\]

\[
g(V_{N-1}, V_N) - g(V_N, V_{N+1}) - CV_N - \frac{1}{R}V_N &= 0, \\
g(V_N, V_{N+1}) - CV_{N+1} - \frac{1}{R}V_{N+1} &= 0,
\]

4.3. Pyrotechnical Microthruster (linear ODE system)

The Pyrotechnical microthruster is a MEMS device, which was fabricated within a European project Micropyros (founded under IST-99047). It is based on the integration of solid fuel with a silicon micro-machined structure (see Fig. 8).

The heat transfer within the hotplate is described through the following equations:
where $\kappa$ is the thermal conductivity in W/mK and $C_p$ is the specific heat capacity in J/kgK. Assuming that both are temperature independent around the working point, which is realistic, the finite element based spatial discretization of (5) leads to a large linear ODE system (2).

5. Numerical Results

Transmission line model of order 6002 has been reduced with Krylov-based block Arnoldi algorithm from [3] down to 150, with the relative error less than 0.02% (not shown). We have experienced problems, as the MNA system matrices $C$ and $K$ coming from Pstar (NXP in-house circuit simulator) were indefinite. In Fig. 9 (left) the instable reduced order model is shown. After the multiplication of the equations which correspond to the inductor branches and the voltage source branch with $-1$ and so making the system matrices positive semi-definite, stability of the reduced model was gained (see Fig. 9 (right)), as well as the excellent approximation (the relative error, which is less than 0.02% is not shown). This example shows however, that the engineers may easily experience difficulties in an attempt to automatize model order reduction process.

Our experiments show, that the above problems do not appear for reducing the linear ODE systems, which arise from spatial discretization of PDEs. Fig. 10 (left) shows the relative error of the BTA and Arnoldi-based reduction of the microthruster device from order 1071 down to order 7. As expected, the BTA shows smaller error in the transient phase but performs worse for the steady-state.
Fig. 9. Instable frequency response of a reduced transmission line model at $V_{100}$ (left) and the stable reduced model, after conversion of the original model into positive definite DAE system (right).

Fig. 10 (right) shows, how relative error between two successive reduced models can be effectively used as an error estimate [17] for Arnoldi algorithm. In this way, compact model extraction becomes completely automatic.

Fig. 10. Relative Error of the BTA and Arnoldi-based reduction (expansion around $s_0 = 0$) in a single output node of the microthruster model of order 1071 (left). Convergence of the relative error between the two successive reduced models for the microthruster (right).

Lastly, in Fig. 11 we show the relative errors over all output nodes of the diode chain model with order 302. The POD models are, as expected, more accurate, but much slower to simulate than the TPWL models (see the corresponding extraction and simulation times in [20]). A significant speed up has been achieved by combining the POD with MPE.

6. Conclusion and Outlook

It has been shown, that the lack of automatic compact modeling is the main bottleneck in the design flow of today’s micro- and nanosystems. Mathematical model order reduction appears to be a “perfect” tool to solve this problem. In this paper, we have described and demonstrated the methodology for applying model order reduction to microelectronic and MEMS models. Furthermore, we have reviewed
the most promising methods for linear and nonlinear MOR. Although, there are still difficulties to completely atomize the process, which are due to necessity of extraction a full-scale model in a proper form and/or due to missing a reliable global error estimate for Krylov-based methods, it looks like that model order reduction for linear ODE and DAE models is mature enough for common engineering use. This is not the case for more realistic and thus much more complicated nonlinear models. Although, we have presented promising numerical results for an academic model of diode-chain, model order reduction of industry-relevant nonlinear models, which might be subjected to a broad spectrum of excitations, remains a research topic.

Furthermore, as model reduction in its original form does not allow us to preserve parameters within the system, which is essential for a quick design iteration, a parametric model order reduction (see e.g. [21]) should be researched.

References


**Acknowledgment:** We would like to acknowledge the EU Marie Curie RTN project COMSON grant n. MRTN-CT-2005-019417.