CAMBRIDGE

Cambridge University Press
978-0-521-86581-4 - Advanced Model Order Reduction Techniques in VLSI Design
Sheldon X. - D. Tan and Lei He
Excerpt
More information

1 Introduction

1.1 The need for compact modeling of interconnects

As VLSI technology advances into the sub-100nm regime with increased operating frequency and decreased feature sizes, the nature of the VLSI design has changed significantly. One fundamental paradigm change is that parasitic interconnect effects dominate both the chip's performance and the design's complexity growth. As feature sizes become smaller, their electromagnetic couplings become more pronounced. As a result, their adverse impacts on circuit performances and powers will become more significant. Signal integrity, crosstalk, skin effects, substrate loss and digital and analog substrate couplings are now adding severe complications to design methodologies already stressed by increasing device counts. It was observed that today's high performance digital design essentially becomes analog circuit design [24] as there has been a need to observe a finer level of detail.

In addition to dominant deep submicron effects, the exponential increase of device counts causes a move in the opposite direction: we need to increase the increasing design abstraction levels to cope with the design capacity growth. It was widely believed that behavioral and compact modeling for the purpose of synthesis, optimization, and verification of the complicated system-on-a-chip are viable solutions to address these challenging design problems [66].

In this book, we focus on the compact modeling of on-chip interconnects and general linear time invariant systems (LTI) because interconnect parasitics, which are modeled as linear RLCM circuits[1.1], are the dominant factors for complexity growth. Unchecked parasitics from on-chip interconnects and off-chip packaging will de-tune the performance of high-speed circuits in terms of slew rate, phase margin and bandwidth [2]. Reduction of design complexity especially for the extracted high-order RLCM networks is crucial for reducing the explosive design productivity gap in the nanometer VLSI design and verification.

This book does not by any means intend to be comprehensive. The absence of coverage of work by other researchers should not diminish their contributions.

1.1 *M here means the mutual inductances.*

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-86581-4 - Advanced Model Order Reduction Techniques in VLSI Design
Sheldon X. - D. Tan and Lei He
Excerpt
More information

1.2 Interconnect analysis and modeling methods in a nutshell

Compact modeling of passive RLC interconnect networks has been a research intensive area in the past decade due to increasing adverse deep submicron effects and interconnect-dominant delays in current high-performance VLSI designs [23, 72]. A number of projection-based model order reduction (MOR) based techniques have been introduced [32, 33, 38, 85, 91, 113, 114] to analyze the transient behavior of interconnects.

An asymptotic waveform evaluation (AWE) algorithm was first proposed [91, 92] where explicit moment matching was used to compute the dominant poles via Pade approximation. The AWE algorithm used the moment concept to control and measure the accuracy of the reduced and the original system, and this was advantageous over many previous black-box fitting methods. Also the AWE method shows that the wildly popular interconnect delay model, The Elmore delay model, is just the first order of moments of a circuit [28]. The success of the AWE method led to intensive research efforts on model order reduction of interconnect circuits.

The AWE method is numerically unstable for higher-order moment approximation. The approximation is carried out around $s = 0$. The same authors introduced some remedial methods to overcome this problem by frequency shifting and expanding around $s = \infty$ [92], but a more effective method involved carrying out multiple-point expansions along the imaginary axis (called frequency hopping) and combining the expansion results at higher computing costs [19].

A more elegant solution to the numerical problem of AWE is to use projection-based model order reduction (MOR) methods, which are based on implicit moment matching. The main idea is to project the explicit moment space into an orthonormal subspace, called Krylov subspace. The projection process basically preserves the moment information, but the Krylov vectors contain much less numerical noise compared with the explicit moments owing to the generation of Krylov subspace. The Pade via Lanczos (PVL) method was the first projection-based method [32], where the Lanczos process, which is a numerically stable method for computing eigenvalues of a matrix, was used to compute the Krylov subspace. Feldmann also proved that the reduced system implicitly matches the original system to a certain order of moments. Later on, the PVL method was extended to deal with multiple input and multiple output cases by MPVL [33], and to deal with circuits with symmetric matrices by SyPVL algorithm [38]. The Krylov subspace can also be generated by the Arnoldi process, which is based on the so-called orthogonal projection. Examples include the Arnoldi method [114] and Arnoldi transformation method [113]. But Arnoldi methods only match the half order of the moments or block moments for the same reduced order.

To ensure the passivity of the reduced model further, it was shown in [63] that congruence transformation can preserve the model's passivity if the system matrices are in a *passive form*. Later PRIMA [85] used the Krylov subspace vectors to form the projector for the congruence transformation, which leads to passive models with

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-86581-4 - Advanced Model Order Reduction Techniques in VLSI Design
Sheldon X. - D. Tan and Lei He
Excerpt
More information

the matched moments in the rational approximation paradigm. Projection-based methods, however, have several drawbacks. First, they are not efficient for circuits with many inputs and output terminals. This reflect in the fact that the reduction cost is tied to the number of terminals; the number of poles of reduced models is also proportional to the number of terminals. Second, PRIMA-like methods do not preserve structure properties like reciprocity of a network. Third, it is difficult to apply PRIMA-like methods to model very high frequency circuits where the circuit parameters are frequency dependent or where only measure data are available in terms of scattering-parameters.

Another approach to circuit-complexity reduction is by means of local node reduction. The main idea is to reduce the number of nodes in the circuits and approximate the newly added elements in the circuit matrix in reduced rational forms. The major advantage of these methods over projection-based methods is that the reduction can be made in a local manner and no overall solutions of the whole circuit are required (with some circuit realization or synthesis techniques), which makes those methods very amenable to attacking large linear networks. This idea has been explored by approximate Gaussian elimination for RC circuits [27], by the TICER program [107], which is also based on Gaussian elimination but only keeps first two moments, and by the extension of TICER method into RLC circuits [3]. The rational approximation is also explored by the direct truncation of the transfer function algorithm (DTT) [56] for tree-structured RLC circuits and by an extended DTT method for non-tree structured RLC circuits [125].

Recently, a more general topology-based node-reduction method was proposed [96, 98], in which nodes are reduced one at a time (topologically, it is called Y-Δ transformation) and the generated admittance in the reduced network is represented as an order-reduced rational function of s. This method is equivalent to symbolic Gaussian elimination (s is the only symbol) but the reduction is made on circuit topologies only, which is equivalent to the nodal analysis (NA) formulation of a circuit only. The stability is enforced by Hurwitz polynomial approximation. But this method only works for linear circuits with limited element types (RCLK-VJ) and cannot be applied to reduce general linear circuits due to NA formulation requirement. A more general multi-node or block version of Y-Δ transformation, named *hierarchical model order reduction or HMOR* was proposed by Tan [121,124]. Since a number of nodes can be reduced at the same time, this method essentially leads to the general node-reduction based hierarchical model order reduction.

The third major development for model order reduction of LTI systems is by means of control-theoretical-based truncated balance realization (TBR) methods, where the weak uncontrollable and unobservable state variables are truncated to achieve the reduced models [81, 87, 89, 131]. The TBR methods can produce nearly optimal models but they are more computationally expensive than projection-based methods. Also, TBR can produce the passive models by so called *positive real TBR methods* [87,131]. Recently, empirical TBR method, named poor man's TBR, was proposed to improve the scalability of the TBR methods, which shows the connection with the generalized projection-based reduction methods [89].

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-86581-4 - Advanced Model Order Reduction Techniques in VLSI Design
Sheldon X. - D. Tan and Lei He
Excerpt
More information

1.3 Book outline

Mode order reduction of time-invariant linear systems is still an active research area. There are many excellent books covering the classic MOR methods such as moment matching, Krylov subspace projection-based methods and node-reduction based methods [14, 20, 72, 97].

In this book, we look at some important developments in this area since the PRIMA method was introduced in 1997, but we make no attempt to be comprehensive. Instead, we primarily present several important methods that give new perspectives on model order reduction techniques in terms of improved efficiency, accuracy, and more compact reduced model sizes over the existing projection-based methods. For instance, we look at truncated balanced realization-based methods, the hierarchical model order reduction method, MOR methods for linear circuits with multiple terminals, MOR methods for highly inductive circuits, general passivity enforcement and circuit realization techniques, and terminal reduction methods. In the following, we give the outline of this book.

- Chapter 2 will review the concepts of model order reduction, moment matching, and classic explicit moment-matching methods like AWE for model reduction. Then we will review Krylov subspace projection-based model order reduction techniques, such as the projection-based MOR methods, which are still the most widely used reduction techniques. We will present the basic concepts of Krylov subspace, passivity, numerical algorithms, such as Arnoldi and Lanczos methods for obtaining orthogonal Krylov basis and reduction matrices, and the PRIMA method. We also present some important theoretical results regarding the Krylov subspaces. Some of the concepts introduced here will be used throughout this book.
- Chapter 3 studies the SVD-based model order reduction technique based on the classic control theory, called truncated balanced realization (TBR), which leads to more compact models than the Krylov subspace projection MOR methods but at much higher computation costs. We will review the basic concepts of truncated balanced realization methods in terms of controllability and observability from the control-theory perspective. We then present the positive real TBR methods which can produce the passive models. After this, the empirical TBR method named poor man's TBR is also presented, which can scale to reduce large circuits. Finally, some numerical and implementation issues with TBR methods are discussed.
- Chapter 4 presents a new passive TBR method, called *PriTBR*, for interconnect modeling. Different from existing passive truncated balanced realization (TBR) methods where numerically expensive Lur'e or algebraic Riccati equations (ARE's) are solved, the new method performs balanced truncation on the linear systems in descriptor form by solving generalized Lyapunov equations. Passivity preservation is achieved by congruence transformation instead of simple truncations. The PriTBR method can be applied as a second stage model

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-86581-4 - Advanced Model Order Reduction Techniques in VLSI Design
Sheldon X. - D. Tan and Lei He
Excerpt
More information

order reduction to work with Krylov subspace methods to generate a nearly optimal reduced model from a large scale interconnect circuit while passivity, structure, and reciprocity are preserved at the same time.

- In Chapter 5, we present the hierarchical model order reduction method, named *HMOR*, which is based on multiple-point expansion. We will review basic steps of the hierarchical reduction technique and describe the flow of the multiple-point expansion based on hierarchical reduction. The concept of symbolic analysis based on a determinant decision diagram (DDD) will be reviewed; this is the core algorithm for the HMOR. We also discus the new pole search algorithm and some important properties of hierarchical reductions such as structure preserving and numerical stability for tree-like circuits.
- Chapter 6 first reviews a terminal reduction algorithm named *SVDMOR*, which performs the reduction on the input and output position matrices of a transfer function matrix. Then we present another general terminal reduction algorithm, named *TermMerg*, to efficiently reduce the terminal number of general linear interconnect circuits with a large number of input and output terminals considering delay variations. TermMerg can reduce many similar terminals and keep a small number of representative terminals. It can also work with passive model reduction algorithms to generate passive compact models. This is in contrast to SVDMOR, which may not produce passive models. After terminal reduction, traditional model order reduction methods can be applied and achieve more compact models and improve simulation efficiency.
- Chapter 7 deals with a new inductance modeling technique, vector potential equivalent circuit and its application in the HMOR. We will discuss the concept of VPEC models and VPEC-based model mutual inductance sparsification technique. Some theoretical results of passivity of VPEC models and its application in the hierarchical modeling reduction will be presented.
- Chapter 8 presents a structure-preserving projection-based MOR method. It starts with the SPRIM method, which is the first structure-preserving reduction algorithm based on 2×2 partitioning of circuit matrices. Then we present a general block structure-preserving projection-based model order reduction technique, called *TBS*, which is an extension of the SPRIM based algorithm. The new algorithm can preserve the structure of the reduced circuits, which makes it easier and more efficient to realize the reduced circuits. Also, we show that by partitioning the original circuits into many disjoint subcircuits, not only can we preserve sparsity of the reduced circuits, but we can also match more poles of the original systems, thus improving the model accuracy.
- Chapter 9 introduces another generalized block structure-preserving reduced order interconnect macromodeling method (BSPRIM). The new approach extends the structure-preserving model order reduction (MOR) method SPRIM [37] into more general block forms. The chapter first shows how a SPRIM-like structure-preserving MOR method can be extended to deal with admittance RLC circuit matrices and show that the $2q$ moments are still matched and symmetry is preserved. It then shows that $2q$ moment match-

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-86581-4 - Advanced Model Order Reduction Techniques in VLSI Design
Sheldon X. - D. Tan and Lei He
Excerpt
More information

ing can't be achieved when the RLC circuits are driven by both current and voltage sources. Using BSPRIM improves SPRIM by introducing the re-orthonormalization process on the partitioned projection matrix. The BSPRIM method can deal with more circuit partitions and can perform the general block structure preserving MOR for circuits formulated in impedance and admittance forms. The reduced models by the proposed BSPRIM will still match the $2q$ moments and preserve the circuit structure properties, like symmetry, as SPRIM does.

- Chapter 10 examines some effective methods to enforce the passivity of models and optimize models. Model passivity and passivity enforcement are widely used for building realizable models from direct measurements and simulation data for radio-frequency and microwave applications. They are also used in HMOR and TermMOR methods. The studied methods include convex-based passivity enforcement and optimization and least-square-based methods for active model optimization.
- Chapter 11 studies the problem of realizing a reduced model into a SPICE-compatible netlist. This process is called model realization. We first present a traditional one-port network synthesis technique, Brune's method, for realizing a passive circuit from its mathematical model. The concept of traditional network realization will be covered. We then present a general multi-port network-based realization technique, which includes a one-port realization based on a Foster's method and general multiple-port impedance realization based on one-port realization techniques. We also discuss how to realize general non-symmetric circuits.
- Chapter 12 presents a novel compact reduced modeling technique to reduce interconnect circuits with many external ports called *TermMOR*. The proposed method overcomes the difficulty associated with subspace projection-based MOR methods for reducing circuits with many ports. The new method can lead to much smaller reduced models for a given frequency range or much higher accuracy given the same model sizes than subspace projection-based methods. Like HMOR, the TermMOR method is a closed-loop method, as it can produce models matching the desired frequency range precisely.
- Chapter 13 presents an approach to enforcing the passivity of a reduced system of general passive linear time-invariant circuits. Instead of making the reduced models passive for infinite frequencies, the method works on the signal waveform driving reduced models. It slightly shapes the waveforms of the signal such that the resulting signal spectra are band limited to the frequency range in which the reduced system is passive. As a result, the reduced models only need to be band-limited passive (also called conditionally passive), which can be achieved much more easily than traditional passivity for a reduced system, especially for one with many terminals or requiring wide band accuracy (more poles). We propose to use spectrum truncation via FFT and IFFT and low-pass-filter-based approaches for transient waveform shaping processing. We analyze the delay and distortion effects caused by using low-pass filters and present methods

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-86581-4 - Advanced Model Order Reduction Techniques in VLSI Design
Sheldon X. - D. Tan and Lei He
Excerpt
More information

to mitigate the two effects.

1.4 Summary

In this chapter, we first present the cases for compact modeling for interconnect circuits. We then briefly survey the previous developments on this topic and present what will be covered in this book for each chapter.

Throughout the book, numerical examples are provided to shed light on the discussed topics to help the reader gain more insights into the discussed algorithms. All our treatments of many topics may not be mathematically rigorous. Instead, we try to present the topics from a typical computer-aided design (CAD) engineer's perspective and try to help reader to apply those techniques to solve real VLSI design problems and develop more efficient simulation tools.

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-86581-4 - Advanced Model Order Reduction Techniques in VLSI Design
Sheldon X. - D. Tan and Lei He
Excerpt
More information

2 Projection-based model order reduction algorithms

Compact modeling of passive RLC interconnect networks has been an intensive research area in the past decade owing to increasing signal integrity effects and interconnect-dominant delay in current system-on-a-chip (SoC) design [72].

In this chapter, we briefly review the existing modeling order reduction (MOR) algorithms for linear time-invariance (LTI) systems developed over the past two decades in the electrical computer-aided design community. Since compact modeling of TLI systems is a well researched and studied field, many efficient approaches have been proposed over the years. Given the space in this book, we cannot review all of them and neither do we attempt to be complete in our review. Instead, we mainly review the Krylov subspace projection-based model order reduction methods, which are widely used MOR methods and are closely related to the rest of this book. Although there exists an excellent and detailed treatment of Krylov subspace projection-based methods already [14], for the completeness of this book, we still present some basic concepts, algorithms and important results for Krylov subspace projection-based MOR methods. We try to present them in a way that can be easily understood from the practical application point of view.

2.1 Moments and moment-matching methods

In this section, we briefly review the concepts of time-domain moments, the Elmore delay and Pade-approximation-based moment-matching method, which are important concepts for subspace projection-based model order reduction methods.

2.1.1 *Concept of moments*

In the s domain, the transfer function of a linear network $H(s)$ is defined as the ratio of the output to the input under zero initial conditions:

$$H(s) = \frac{Y(s)}{X(s)}. \tag{2.1}$$

If the input is the impulse function $\delta(t)$, its Laplace transformation is 1. So the transfer function is also the impulse response at the port. If we expand $H(s)$ around

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-86581-4 - Advanced Model Order Reduction Techniques in VLSI Design
Sheldon X. - D. Tan and Lei He
Excerpt
More information

$s = 0$ by the Taylor series expansion, we have

$$H(s) = \sum_{k=0}^{\infty} m_k s^k, \tag{2.2}$$

where

$$m_k = \frac{1}{k!} \times \frac{\mathrm{d}^k H(s)}{\mathrm{d}s^k}\bigg|_{s=0}. \tag{2.3}$$

where the kth coefficeint of $H(s)$, m_k, is called the kth moment.

Assuming that $h(t)$ is the corresponding time-domain impulse response, we have

$$H(s) = \int_0^{\infty} \mathrm{e}^{-st} h(t)\mathrm{d}t. \tag{2.4}$$

We rewrite moments defined in (2.4) in terms of $H(t)$ by using the Taylor expansion of e^{-st} in the Laplace transform $H(s)$ and we have

$$\begin{aligned} H(s) &= \int_0^{\infty} h(t)\mathrm{e}^{-st}\mathrm{d}t \\ &= \int_0^{\infty} \left(1 - st + s^2\frac{t^2}{2} + \cdots + s^k\frac{(-1)^k}{k!}t^k + \cdots\right)\mathrm{d}t \\ &= \sum_{k=0}^{\infty} s^k \frac{(-1)^k}{k!}\int_0^{\infty} t^k h(t)\mathrm{d}t. \end{aligned} \tag{2.5}$$

Comparing (2.5) with the definition that $H(s) = \sum_{k=0}^{\infty} m_k s^k$, moments can be rewritten as:

$$m_k = \frac{(-1)^k}{k!}\int_0^{\infty} t_k H(t)\mathrm{d}t, \tag{2.6}$$

or

$$m_0 = \int_0^{\infty} h(t)\mathrm{d}t, \tag{2.7}$$

$$m_1 = -\int_0^{\infty} th(t)\mathrm{d}t, \tag{2.8}$$

$$m_2 = \frac{1}{2!}\int_0^{\infty} t^2 h(t)\mathrm{d}t, \tag{2.9}$$

$$\cdots$$

2.1.2 Elmore delay

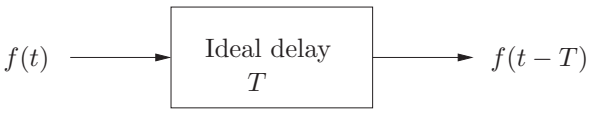

Figure 2.1 The network of an ideal delay of T.

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-86581-4 - Advanced Model Order Reduction Techniques in VLSI Design
Sheldon X. - D. Tan and Lei He
Excerpt
More information

For an ideal delay network as shown in Figure 2.1, the response of the network for an input function $f(t)$ is $f(t-T)$. We take the Laplace transformation of $f(t-T)$ and we have

$$\mathcal{L}(f(t-T)) = \mathrm{e}^{-sT}F(s), \tag{2.10}$$

where $F(s)$ is the Laplace transformation of $f(t)$ and $\mathcal{L}(*)$ is the Laplace transformation operator. So the ideal delay element's transfer function is

$$H_d(s) = \mathrm{e}^{-sT}. \tag{2.11}$$

If we take the derivative of $H_d(s)$ with respect to s, we have

$$\frac{\mathrm{d}H_d(s)}{ds}|_{s=0} = -T. \tag{2.12}$$

As a result, we may use $\frac{\mathrm{d}H_d(s)}{ds}|_{s=0}$ as an approximate for the delay of a general linear network described by $H(s)$, i.e.,

$$T_d \approx -\frac{\mathrm{d}H_d(s)}{ds}|_{s=0}. \tag{2.13}$$

T_d is the so-called Elmore delay [28], which is also the first-order moment in (2.3). So we have

$$m_1 = \int_0^\infty h(t)t\mathrm{d}t = \frac{\mathrm{d}H_d(s)}{ds}|_{s=0}. \tag{2.14}$$

Another popular mathematic interpretation of the Elmore delay is by means of probability perspective. Physically, the delay of a network can be measured using the 50% point delay of the monotonic step response from a unit step input. If $h(t)$ is the unit impulse response, the unit step response is $\int_o^\infty h(t)\mathrm{d}t$. The 50% point delay τ is then defined as

$$\int_0^\tau h(t)\mathrm{d}t = 0.5. \tag{2.15}$$

If we treat $h(t)$, which is assumed to be non-negative for all $t \geq 0$, as the probability density function (p.d.f.), i.e., $\int_0^\infty h(t) = 1$, the Elmore delay, T_d, is essentially the mean under the p.d.f. of $h(t)$, the impulse response of the network

$$T_d = m_1 = \int_0^\infty h(t)t\mathrm{d}t, \tag{2.16}$$

which actually is the first-order moment m_1.

It can be shown that for an ideal delay network or a network whose impulse response is symmetric, the Elmore delay is exactly the actual 50% delay of the network. For practical networks, whose responses are always skewed as shown in Figure 2.2, the Elmore delays are just an estimation. Actually Gupta and Pileggi *et al.* proved that the Elmore delay is the upper bound for general RC circuits [48].

The Elmore delay was first introduced by Elmore in 1948 for estimating the delay of active circuits. It was popularized by Penfield and Rubinstein [100] as it can be computed directly and efficiently for RC trees by using the R and C values

© Cambridge University Press www.cambridge.org