Efficient Techniques for Modeling Chip-level Interconnect, Substrate and Package Parasitics

Peter Feldmann Sharad Kapur David E. Long
Bell Laboratories, Lucent Technologies

Abstract

Modern IC design requires accurate analysis and modeling of chip-level interconnect, the substrate and package parasitics. Traditional approaches for such analyses are computationally expensive. In this paper, we discuss some recent novel schemes for extraction and reduced order modeling that help overcome this computational bottleneck.

1 Introduction

In recent years, increasing operating frequencies caused by faster digital chips and the emergence of integrated Radio Frequency (RF) applications has introduced a number of new design modeling challenges. In particular, parasitic effects of chip-level interconnect, the substrate and package can no longer be ignored, or modeled in a simple manner. Moreover, at these high frequencies, the signal wavelength becomes comparable to the size of circuit structures. In some cases, lumped element circuit models cease to be accurate and distributed effects need to be captured. In addition, high frequencies cause complex interactions between structures that traditionally could be analyzed separately. As a consequence, traditional extraction tools using rule-based schemes are becoming inadequate and must be replaced by methods that model electromagnetic effects more accurately. Unfortunately, even when accurate extraction can be done, the sheer size of the resulting models makes circuit level analysis impractical.

Typical simulations problems consist of analyzing a large transistor network combined with the extracted on-chip interconnect, parasitics, models of passive linear components, models for the package and board level interconnect. For example, the accurate simulation of a Low Noise Amplifier (LNA) requires in addition to the modeling of the transistor circuit, incorporating parasitics due to the interconnect and substrate and package. A simpler model would not reliably predict the stability of the amplifier. These type of simulations cause a number of difficulties. First, lumped element networks produced by chip-level layout extraction tools can result in millions of capacitive, resistive and inductive elements. On-chip passive elements (such as inductors) and packages have significant distributed effects and are naturally modeled in the frequency domain by a transfer function matrix. Of all the general circuit-level analysis methods, only the method of Harmonic Balance can efficiently handle a mixture of time-domain and frequency-domain methods. Unfortunately, this solution technique is applicable to a small class of problems, and most digital circuits are analyzed in the time-domain with Spice-like tools.

In this paper, we describe recent novel techniques which help in modeling increasingly complex circuits. We first discuss efficient algorithms for solving large linear systems associated with electrostatic and electromagnetic simulation of layout structures required in the extraction process. Next, we describe techniques which generate reduced order models from the results of the extraction. These reduced order models can be used in a higher-level, non-linear, time domain (Spice) or frequency domain (Harmonic Balance) simulation. Finally, we demonstrate such analyses with a number of examples.

2 Static and full-wave extraction

Extracting compact, accurate linear models for packages, interconnect, and components plays a significant role in modern IC designs. Models can be extracted in a variety of ways, but for the high accuracy that demanded by the critical sections, only numeric simulation at the electromagnetic level suffices.

Simulation methods can be broadly divided into two classes. Methods in the first class use differential equation formulations. Finite-element (FE) [2], finite-difference (FD) [15], and finite-difference time-domain (FDTD) [16] approaches all fall into this class. Methods from the second class use integral equations. The method-of-moments (MoM) approach [6] is based on integral equations. While the integral formulation leads to dense matrices, it allows us to apply Green's theorem, reducing volume integrals to surface integrals. This can reduce the matrix dimension significantly since the discretization only involves surfaces such

as the boundary of a conductor or the interface between two dielectrics. For typical IC, board, package or MCM simulations, where the material variation is usually simple (layered dielectric media), the integral approach has become very popular due to the use of surface discretizations. However, the matrix dimension can still easily be in the thousands for complex problems. Traditional direct methods for matrix solution cannot be used to solve such large systems and in recent years a number of novel schemes have been developed for the rapid solution of the matrices associated with extraction.

Maxwell's equations govern the behavior and performance of integrated circuits. At lower frequencies, in an electro or magnetostatic regime, these equations reduce to Laplace's equation. At higher frequencies, a full electromagnetic simulation must be performed.

For example, in the standard problem of capacitance extraction in three dimensions, we compute the charge density ρ by solving the following integral equation:

$$\phi(\mathbf{r}) = \int_S G(\mathbf{r}, \mathbf{r}')\rho(\mathbf{r}')dS'. \quad (1)$$

where ϕ is the potential, and G is the Green's function which evaluates the potential at the position $\mathbf{r}$ due to a point charge at $\mathbf{r}'$. For example, in free-space $G(\mathbf{r}, \mathbf{r}') = 1/(4\pi\epsilon_0||\mathbf{r} - \mathbf{r}'||)$ and ϵ_0 is the permittivity of vacuum. ICs are typically embedded in layered dielectric media. In such layered media, analytic expressions for the Green's functions do not exist. These Green's functions are characteristic to a particular silicon process and can be numerically precomputed and stored for efficiency [17].

In order to be numerically solved, the integral equation is discretized by various methods such as collocation or Galerkin methods [11] or higher order Nystrom schemes [8]. The problem is then reduced to the solution of a linear system of equations of the form:

$$A\rho = \phi. \quad (2)$$

For full-wave solutions, the formulation and numerical discretization is a little more involved. The time-harmonic Maxwell's equations are solved using the electric field integral equation [10]. Again, the problem is reduced to the solution of a linear system of the form:

$$B(j\omega)X(\omega) = V, \quad (3)$$

where

$$B(j\omega) = (-\Omega - j\omega A - \frac{1}{j\omega}\Phi), \quad (4)$$

is a matrix composed of a sparse Ohmic interaction (resistive) matrix Ω, a dense vector potential (inductive) matrix A and a dense scalar potential (capacitive) matrix Φ. V is the applied (potential) stimulus and $X(\omega)$ are the unknown currents.

2.1 Fast Algorithms for Dense Matrix solution

For many years, the use of integral equation based formulations was limited because of the following problem: The matrices A and B associated with the electrostatic and electromagnetic simulation are dense. Dense linear algebra is computationally *expensive*. Direct solution of the these linear systems using Gaussian elimination requires $O(N^2)$ storage and $O(N^3)$ time and is impractical for large problems. Fortunately, typical systems arising from integral equations are well conditioned and can be solved by Krylov-subspace iterative schemes such as GMRES [14]. Iterative solvers require application of the matrix A to a sequence of recursively generated vectors. The dominant costs become the $O(N^2)$ time and space required for constructing and storing the matrix and the $O(N^2)$ time required for each matrix-vector product. Unfortunately, for modern state-of-the-art systems, the problem can easily result in having many thousands to a million unknowns, obviating the use of such methods.

Most of the algorithms in recent years have focussed on the numerical compression (or sparse representation) of these dense matrices. Instead of having to represent the matrix with $O(N^2)$ elements with a sparse representation using only $O(N)$ or $O(NlogN)$ numbers. This is achieved by exploiting the structure arising from the physical properties of the problem.

FastCap [11] and FastHenry [7] employ an algorithm called the Fast Multipole Method [4]. The fast multipole method (FMM) [4], while originally developed for particle simulation problems, can be combined with iterative techniques to solve the dense matrices arising from integral equations. This method is based on the observation that the field due to a charge (or current) source decays smoothly with distance and can be dramatically compressed using multipole expansions. While the algorithm is fast, because this implementation of the FMM is tailored to the free-space $1/||r - r'||$ kernel, dealing with common situations such as layered dielectrics is difficult. Another approach called the Precorrected-FFT [13] combines the use of the Fast Fourier Transform with an interpolation based approach for the rapid computation of the matrix vector products. This algorithm has a much smaller memory requirement than the FMM.

More recently, IES^3 ("ice cube"), an Integral Equation Solver for three-dimensional problems [9] was shown to have significant performance advantages over competing approaches. IES^3 compresses dense matrices, by exploiting the smooth variation in interaction strength. By an appropriate ordering of elements, large parts of the interaction matrix become numerically low-rank and can be compactly represented using their singular value decomposition (SVD). The time and memory required to compress the matrix is only $O(N \log N)$, where N is the matrix dimension.

The time required by a matrix-vector multiply using the compressed representation is also $O(N \log N)$. Combining the compressed representation with an iterative solver such as GMRES results in an efficient method for solving the integral equation. In Figure 3 we present a timing comparison of IES^{3}and direct methods for the simulation of an RF test socket (in Figure 1).

Figure 1 illustrates the compressed representation via a rank map for the capacitance extraction of an RF test socket. The rank map shows the partitioning of the matrix into submatrices and the rank of each submatrix. The structure of this rank map is similar for most extraction problems and results in nearly linear storage requirements.

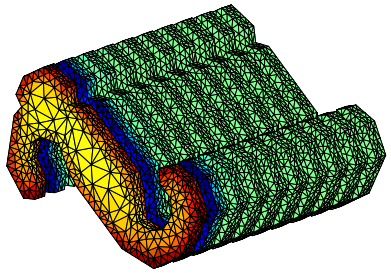

Figure 1. Charge distribution on an RF socket

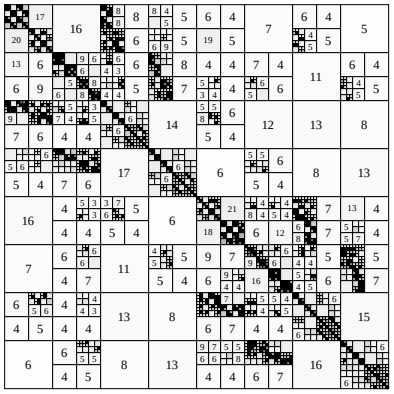

Figure 2. Compression of the matrix

3 Reduced Order Modeling of Linear Systems

The large linear systems resulting from the extraction are described using an inordinate amount of modeling data. Their representation must be considerably compressed in order to render them useful for circuit analysis. Such compression can be achieved through reduced-order modeling.

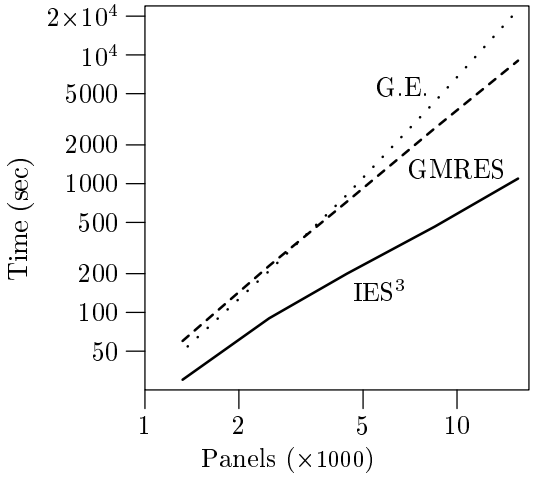

Figure 3. IES^3 vs direct schemes

These linear systems are described using the state-space equations

$$B(s)X(s) = RI(s), \quad O(s) = L^T X(s). \tag{5}$$

Here, $X(s)$ is a vector of state variables, $B(s)$ is a matrix that describes the state equations, $I(s), O(s)$ are vectors representing the system inputs and outputs. L and R are constant incidence matrices that specify the location of the input/output ports. From the circuit analysis point of view, only the behavior of the linear system at its ports is of interest. This behavior is fully described by the system transfer function, $H(s)$

$$O(s) = \underbrace{L^T B(s)^{-1} R}_{H(s)} I(s). \tag{6}$$

The reduced-order model should capture with sufficient accuracy the input-output behavior of the original linear system in the desired frequency range. In addition, the reduced-order model should have efficient representations in both the time and frequency domains.

Extraction programs model chip-level interconnect parasitics as large, lumped-element, linear networks, consisting of up to millions of capacitive, resistive, and even inductive elements, (RLC circuits in general).

The equations describing the RLC circuits can be brought to the form

$$(G + sC)X(s) = RI(s), \quad O(s) = L^T X(s). \tag{7}$$

where G and C capture the dissipative and reactive parts of the system, respectively. These matrices are typically too large to be of practical use in the analysis of the complete circuit. In general, a reduced-order model preserves

the same form of the equations, however, involving much smaller matrices.

$$(\hat{G} + s\hat{C})\hat{X}(s) = \hat{R}\hat{I}(s), \quad \hat{O}(s) = \hat{L}^T\hat{X}(s). \qquad (8)$$

The smaller matrices are obtained by projection of the original system matrices into "well-chosen" subspaces. Depending on the choice of the spaces and of the projection, model reduction methods can achieve different desired properties.

The PVL (Padé via Lanczos) family of algorithms [1, 3] first transforms the system matrix $G + sC$ into $I + sA$ by a change of variables. Here, $A = (G + s_0C)^{-1}C$, and s_0 is the frequency-domain expansion point. The new system matrix is projected from the left and the right onto the block-Krylov subspaces spanned by $[R, AR, A^2R, \ldots]$, and $[L, A^TL, (A^T)^2L, \ldots]$, respectively. The projection of the problem onto the block-Krylov subspaces is calculated implicitly by a block-Lanczos algorithm. The transfer function of resulting reduced system is shown to represent a matrix-Padé approximation of the original transfer function. By the definition of the Padé approximation, the Taylor expansions of the original and the approximant coincide up to the $2n^{\text{th}}$ term.

$$H(s) - \hat{H}(s) = O((s - s_0)^{2n}); \qquad (9)$$

The Padé approximation computed by PVL-type methods is optimal in the sense that it generates the largest achievable number of matched Taylor coefficients. Unfortunately, the Padé approximation does not always maintain certain desirable properties of the original circuit such as stability and passivity. These properties can often be regained by simple post-processing.

An alternative method, first implemented in PRIMA [12] is to project the original $G + sC$ system matrix both from right and left onto the same block-Krylov subspace $[R, AR, A^2R, \ldots]$. This method generates a sub-optimal Padé-type approximation, matching only half of the possible Taylor coefficients for a given matrix size. In return, the passivity of the reduced-order model is preserved.

Other model-reduction methods have also been proposed. By considering multiple expansion points $s_1, s_2, \ldots$, with their corresponding system matrices, $A_1, A_2, \ldots$, one generates a subspace spanned by union of block-Krylov basis vectors. This family of methods called rational-Krylov schemes is described in [5] and, while being relatively expensive to apply, has the potential of producing the most compact models.

As an example we show the reduction of a package model. The package is originally modeled by a linear circuit consisting of more than 4000 RLC elements. The simulation of the integrated circuit, in this case a low-noise-amplifier, in its true environment was extremely slow due to the large package model network. By applying PVL-based model-reduction with post-processing for passivity, a model consisting of only 80 state variables was produced. Figure 4 shows the transfer function of one package pin, calculated exactly and from the reduced-order model. There is almost no loss of accuracy due to model reduction despite the two order-of-magnitude decrease in complexity.

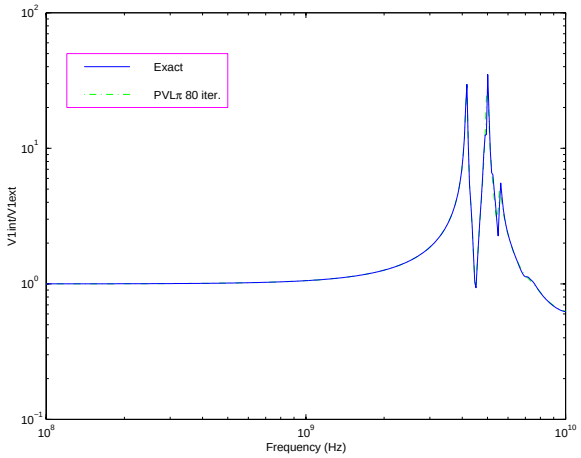

Figure 4. Reduced-order model of a package

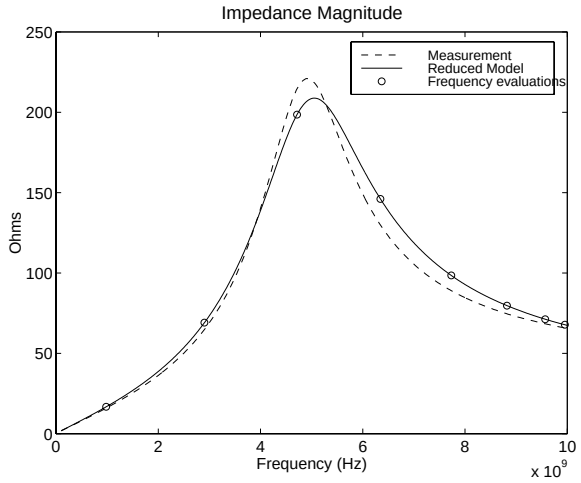

Figure 5. Full-wave reduced inductor model

The linear systems resulting from the time-harmonic Maxwell equations, as solved by IES^3, are of a similar form but slightly different in nature,

$$(-\Omega - j\omega A - \frac{1}{j\omega}\Phi)X(\omega) = R, \qquad (10)$$

Note that the equations are only defined on the imaginary axis of the complex plane ($s = j\omega$). The very large and dense matrices that describe the transfer function are now frequency dependent. However, for typical RF structures

(which are usually a small fraction of the wavelength), this dependence is weak. These matrices can be projected in subspaces spanned by system solutions (and transpose solutions) at multiple frequency points, resulting in a significantly more compact transfer function representation with excellent approximation properties. For systems that are not dependent on frequency, a multipoint Pade-like approach will exactly capture exactly two moments at each point. For systems that have weak frequency dependence this level of accuracy degrades only slightly. In Figure 5 we compare simulations based on reduced-order full-wave modeling of a spiral inductor (solid line) to exact transfer function evaluations (circles). We also present comparison to measurement (dotted line). In this example, the the dimension of the dense matrices was 2200. The dimension of the reduced-order model matrices was only 12. Eight evaluation points (circles) were chosen for system solves. Excellent agreement is observed between the exact transfer function evaluations and reduced order simulations. Clearly, if one needs a detailed simulation it is much more efficient to do it (at the expense of a few system solves) via a reduced model.

4 Conclusion

In this paper we discussed a number of schemes for overcoming the computational complexity involved in the simulation and modeling of modern ICs, parasitics of packages, chip-level interconnect and the substrate. We described a number of algorithms, based on dense matrix compression, for efficient capacitance and full-wave extraction. We also discussed a variety of Krylov-subspace based approaches for constructing reduced order models of large RLC circuits and full-wave distributed structures. Finally, we discussed a few of the open problems associated with the passive reduction of frequency dependent matrices that arise in simulation of distributed structures.

This compact full-wave representation, however is only useful in frequency domain analysis. The generation of a time-domain model from a frequency-domain matrix transfer function requires the solution of a realization problem. There are a number of ad-hoc approaches for doing this. For example, it is easy to construct a polynomial representation of the frequency dependent reduced model and to use a larger companion matrix to reduce it to a frequency independent form (7). It is also relatively easy to enforce stability by projecting unstable poles out of the model. Unfortunately, passivity is hard to preserve because of the underlying "non-passive" polynomial interpolation that is required to obtain the standard state space representation (7). The realization problem is difficult and has been extensively studied in the control literature. Currently, most practical solutions are of a rather ad-hoc nature and the problem can be considered an open.

References

[1] P. Feldmann and R. Freund. Efficient linear circuit analysis by Padé approximation via the Lanczos process. In *Proceedings of the European Design Automation Conference*, Sep 1994.

[2] R. L. Ferrari and P. P. Silvester. *Finite Elements for Electrical Engineers*. Cambridge University Press, 1996.

[3] R. W. Freund and P. Feldmann. Reduced-order modeling of large passive linear circuits. In *Proceedings of the International Conference on Computer-Aided Design*, pages 280–287, Nov 1996.

[4] L. Greengard and V. Rokhlin. A fast algorithm for particle simulations. *Journal of Computational Physics*, 73(2):325–348, Dec. 1987.

[5] E. J. Grimme. *Krylov Projection Methods for Model Reduction*. PhD thesis, University of Illinois at Urbana-Champaign, Urbana, Illinois, 1997.

[6] R. F. Harrington. *Field Computation by Moment Methods*. IEEE Press, New York, 1991.

[7] M. Kamon, M. J. Tsuk, and J.White. FASTHENRY: A multipole-accelerated 3-D inductance extraction program. *IEEE Transactions on Microwave Theory and Techniques*, 42(9):1750–1758, Sept. 1994.

[8] S. Kapur and D. E. Long. High-order Nyström schemes for efficient 3-d capacitance extraction. In *38th International Conference on Computer Aided Design*, Nov 1998.

[9] S. Kapur and D. E. Long. IES^3: Efficient electrostatic and electromagnetic simulation. *IEEE Journal of Computation Science and Engineering*, 5(4):60–67, Oct. 1998.

[10] S. Kapur, D. E. Long, and J. Zhao. Efficient full-wave simulation in layered, lossy media. In *Proceedings of the IEEE Custom Integrated Circuits Conference*, May 1998.

[11] K. Nabors, F. T. Korsmeyer, F. T. Leighton, and J. White. Preconditioned, adaptive, multipole-accelerated iterative methods for three-dimensional first-kind integral equations of potential theory. *SIAM Journal on Scientific and Statistical Computing*, 15(3):713–735, May 1994.

[12] A. Odabasioglu, M. Celik, and L. T. Pillegi. Prima: Passive reduced order macromodeling algorithm. *IEEE Transactions on Computer-Aided Design*, 17(8):645–654, Aug. 1998.

[13] J. R. Philips and J. White. A precorrected-FFT method for capacitance extraction of complicated 3-D structures. In *Proceedings of the International Conference on Computer-Aided Design*, Nov. 1994.

[14] Y. Saad and M. H. Shultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 7(3):856–869, July 1986.

[15] G. D. Smith. *Numerical Solution of Partial Differential Equations: Finite Difference Methods*. Oxford University Press, 1986.

[16] A. Taflove. *Computational Electrodynamics: The Finite-Difference Time-Domain Method*. Artech House, 1995.

[17] J. Zhao, S. Kapur, D. E. Long, and W. W.-M. Dai. Efficient three-dimensional extraction based on static and full-wave layered Green's functions. In *Proceedings of the 35th Design Automation Conference*, June 1998.