Model Reduction for VLSI Physical Verification

N.P. van der Meijs
Delft University of Technology, Dept of ITS/EE, Mekelweg, 2628 CD Delft, The Netherlands, nick@cas.et.tudelft.nl

Abstract

In this paper, we will first introduce physical verification of Very Large Scale Integration (VLSI) circuits and put the challenges in perspective from a computational electro-magnetics point of view. We argue that any models obtained from electromagnetic (electrostatic) analysis should be as simple as possible for a given accuracy. We will then focus on two aspects of VLSI physical verification, namely Finite Element based interconnect resistance extraction and Boundary Element Based interconnect capacitance extraction. We will specifically highlight some associated model order reduction issues, some of them only arise upon a specific interpretation of the FEM and BEM solution schemes.

Introduction

VLSI (Very Large Scale Integration) circuits consist of a small (approx. 1cm x 1cm) silicon chip. At one side of the chip the so-called active devices (typically for signal amplification and switching) are created inside the top few micrometers of the silicon, and on top of it an intricate interconnection pattern is created. Such chips may be very complex: the basic length scale of the devices is usually around 0.1 ... 0.2 μm for the most advanced technologies and millions (>50 million for the Pentium IV) of such devices can be integrated on one chip. The interconnect pattern is also very complex: around 8 metalization layers criss-crossing over the chip, totaling several km's in length, and with hundreds of millions of connections between the layers (through the insulating intermediate layer between each pair of interconnect layers). See Figure 1 for an example of such an interconnect pattern. The figure shows a SEM image where the insulating, dielectric layers have been etched away to show the 3D structures.

The on-chip interconnections are not ideal: they e.g. exhibit non-negligible parasitic resistive, capacitive and inductive behavior. During the chip design process, care must be taken to account for their effects upon the performance and functionality of the chip. Indeed, these interconnection parasitics tend to determine the performance. With each new technology generation, their relative influence (compared to the intrinsic performance potentially offered by the active devices w./o. considering the interconnect parasitics) is rising, and for current state-of-the-art technologies it is even dominating.

However, during the chip design phase, only simple interconnect models can be used. Accurate models would be far too complicated: the associated amount of detail would effectively inhibit any design optimization procedure because of the required amount of computing resources, and it would totally obfuscate conceptual design. Therefore, chip design is always followed by a so-called physical verification step, in which any design flaws are to be discovered before the design is released for manufacturing. If necessary, the verification results are used to fine-tune or change the design. In practice, design and verification form an iterative, hopefully converging, procedure.

Physical verification by necessity employs much more accurate and detailed interconnect models than being used during design. Although this is feasible during verification, it still is a daunting task because of the complexity of scale (e.g. km's of interconnect) on the one hand and the sub-micron scale geometric resolution on the other hand. In principle, one has to solve Maxwell's equations for structures of which Figure 1 only shows 1 millionth part. Fortunately, however, many simplifications can be made and most often it is allowed to first compute a so-called "equivalent circuit-model". This is essentially the circuit as intended by the designer, but augmented ('back-annotated') with lumped parasitic resistors, capacitors and inductors. Subsequent verification (either static analysis or simulation) will then show (lack of) compliance to the specifications.

Given the scale of the problem, the equivalent circuit model (obtained by a process called 'layout to circuit extraction') easily overwhelms the capacity of the actual verification task. Therefore, so-called *reduced order modeling* or *model order reduction* is of paramount importance. This activity aims at producing circuit models as simple as possible that still accurately predict the actual chip behavior. Obvious issues that are involved with this process include stability and/or passivity of the resulting models, accuracy and required computing resources (memory and CPU time).

In this paper, we will focus on Finite Element based interconnect resistance extraction and 3D Boundary Element Based interconnect resistance extraction. We will specifically highlight some associated model order reduction issues, some of them only arise upon a specific interpretation of the FEM and BEM solution schemes. Our goal is to show our modeling philosofy, hoping for exposing cross-fertilization opportunities.

FEM for Interconnect Resistance Modeling as Model Order Reduction

Matrix Formulation of the FEM Without going into detail (but see, e.g. [8]), we state that the finite-element method discretizes the interior of the interconnections and produces the following system of equations:

$$\begin{bmatrix} \mathbf{A}_{tt} & \mathbf{A}_{it} \\ \mathbf{A}_{ti} & \mathbf{A}_{ii} \end{bmatrix} \begin{bmatrix} \mathbf{x}_t \\ \mathbf{x}_i \end{bmatrix} = \begin{bmatrix} \mathbf{b}_t \\ \mathbf{b}_i \end{bmatrix} \tag{1}$$

Here, $\mathbf{A}$ is a symmetric, positive definite and sparse matrix, $\mathbf{x}$ is a vector of unknown potentials and $\mathbf{b}$ is a vector of currents fixed by the boundary conditions that are partitioned into blocks associated with the terminals

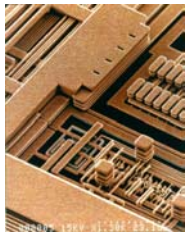

Figure 1: *Copper interconnect pattern (IBM)*

(subscript t) and the internal variables (subscript i), where $\mathbf{b}_i$ is zero. We can then Gaussian-eliminate $\mathbf{x}_i$ since we seek a compact relation between the terminal currents and voltages only as follows:

$$\mathbf{x}_t = \left[\mathbf{A}_{tt} - \mathbf{A}_{it}\mathbf{A}_{ii}^{-1}\mathbf{A}_{ti}\right] \mathbf{b}_t = \mathbf{Y}\ \mathbf{b}_t. \tag{2}$$

Finally, we can create an admittance network where an admittance between nodes i and j has a value $G_{ij} = -Y_{ij}$. Thus, we directly obtain an equivalent circuit model of the layout, without solving a set of field problems. However, it is even better to actually already begin with a circuit interpretation of $\mathbf{A}$ in Equation (1). Such an interpretation of the FEM [8] initially produces a large but sparse resistance network that models the resistive interconnections in detail but that is subsequently reduced by a set of node eliminations. Stated in this way, solving the FEM equations is in fact a kind of model order reduction. We will show that this perspective offers some interesting opportunities for improving the versatility of the FEM method.

The (Gaussian) node elimination step with G_{ij}^k the admittance between node i and j before the elimination of node k, and G_{ij}^{k+1} after the elimination of node k (we assume that node k is eliminated in step k), can be written as follows:

$$G_{ij}^{k+1} = G_{ij}^k + \frac{G_{ik}^k G_{jk}^k}{\sum_{x \neq k} G_{kx}^k}. \tag{3}$$

Equation (3) states that the star network of the node that is eliminated is replaced by a full network (a *clique*) on the nodes that were previously connected to node k. It also shows that a node k can be eliminated as soon as all the admittances G_{kx} connected to node k are known. Such early elimination of each node significantly reduces the amount of storage required when compared to first building the complete network and then doing the eliminations. Early elimination is in fact of paramount importance when large layouts must be handled.

The admittances connected to a node are all known when all the finite elements incident to that node have been assembled. If the elements are processed systematically from one side of the layout to the other, e.g. in a kind of scanline fashion, the resulting ordering is inefficient: it typically produces too many fill-ins.

To improve the efficiency, the "as soon as possible" elimination strategy must be adapted. Then, one of several heuristics (note that the problem of computing the best ordering is NP complete) can be applied. One simple but reasonably effective heuristic is the minimum degree ordering heuristic: First the node with the lowest degree is eliminated, then the next lowest, and so on. This heuristic is motivated by the fact that elimination of a low degree node can only produce a limited amount of fill-in.

Min-degree ordering is not directly possible for layout to circuit extraction, because it is really undesirable to work with the full admittance matrix from a memory perspective. However, consider introducing a short priority queue [5] of a fixed, adjustable size. Whenever a node is ready for elimination, it is not immediately eliminated but only inserted into this queue and when the queue becomes full, a node in the queue with the lowest degree is (Gaussian-)eliminated and removed from the queue to make room for the next insertion. Using only a short priority queue, CPU time can be reduced by more than an order of magnitude[7].

Articulation Nodes Delayed Frontal Solution conveniently co-operates with another heuristic, called "Introduction and Preservation of Articulation Nodes". For any long piece of interconnect, the current will flow along its length, and perpendicular to the length there will be an equipotential. Now, assume a finite element triangulation with edges along such equipotential lines perpendicular to the current flow. Upon interpretation of such a FEM discretization as a resistance network, there will not flow any current between any pair of nodes joined by such an "equipotential edge". Hence, the two nodes may be contracted without introducing error.

Such contracted nodes are called 'articulation nodes', as the FEM graph would break in unconnected parts upon their removal. When they are deliberately inserted as often as possible, they can help reducing the fill-ins when they are eliminated later than all other internal nodes—fill-in can not occur between sections joined by articulation nodes. In effect, the articulation node mechanism partitions the elimination problem into many smaller subproblems. This helps for the performance because of the super linear computational complexity of Gaussian elimination. Articulation nodes can also be introduced for smaller good conducting metalization regions joining several higher resistive interconnections, where the whole metal region forms an approximate equipotential region. Articulation nodes are illustrated in Figure 2. Reference [10] shows that the time complexity changes dramatically, and becomes approximately linear in the size of the layout.

RC Model Reduction Gaussian elimination can also be applied to networks specified in the Laplace (or frequency) domain. Then, the initial circuit might consist of both resistors and capacitors, with $Y_{ij}(s) = 1/R_{ij}$

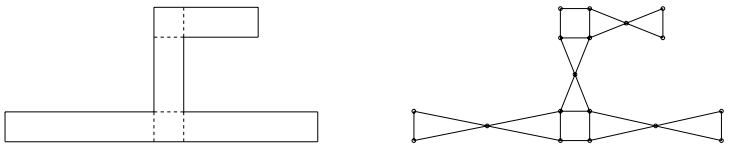

Figure 2: *Piece of interconnect subdivided into rectangles, FEM graph with articulation nodes corresponding to equipotential lines in the layout.*

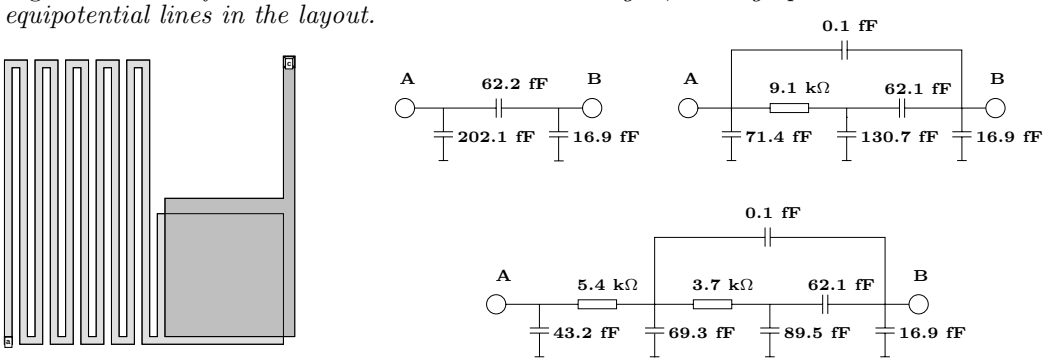

Figure 3: *Layout of a poly resistor (grey) on top of which a metal plate capacitor (black) and 3 extracted circuit models for f_s = 1 MHz, f_s = 100 MHz and f_s = 200 MHz.*

and $Y_{ij}(s) = sC$, respectively. Upon elimination of a node, higher order powers (or moments) of s arise [2]. In practice, it is sufficient to truncate the moment expressions at some low order, since a short series typically is accurate enough for RC circuits. Moment matching (i.e. Pade) techniques may be used to find an approximation of the low-order poles, facilitating e.g. back transformation to the time domain.

Furthermore, [1] presents a selective node elimination (SNE) method. Starting from an RC network, the method eliminates the least significant nodes such that the resulting reduced RC network accurately models the original network from DC up to a user specified frequency. The significance of a node for the frequency range is estimated from the error that would result from ignoring the higher order moments when the node would be eliminated. That is, for each node k an error weight δ_k is computed based upon the relative strength of the change in the second versus the zeroth and first moment at $\omega_s = 2\pi f_s$, as follows

$$\delta_{ij} = \lim_{s \to j\omega_s} \mid \frac{s^2 \tilde{M}_{ij}^{(2)}}{\overline{M}_{ij}^{(0)} + s\overline{M}_{ij}^{(1)}} \mid = \frac{\omega_s^2 \mid \tilde{M}_{ij}^{(2)} \mid}{(\mid \overline{M}_{ij}^{(0)} \mid^2 + \omega_s^2 \mid \overline{M}_{ij}^{(1)} \mid^2)^{\frac{1}{2}}} \quad , \quad \delta_k = \parallel (\delta_{ij} | i,j \in N_k) \parallel \tag{4}$$

where N_k consists of all the nodes previously connected to k and i and j enumerate all its edge combinations (it is a clique). This error weight is initially (step 0) computed for all nodes. Then, (step 1), the node q with the lowest error weight is eliminated and (step 2) the error weights of all nodes previously connected to q are updated. Step 1 and 2 are repeated until all remaining errors exceed a specified tolerance and the algorithm terminates with the f_s specific circuit model.

This is illustrated in Figure 3. The layout on the left is a physical realization of a RC series circuit that was intended by the designer. The resistor has been folded in order to save space, but this causes some parasitic capacitive coupling between the different branches that renders the high frequency behavior much more complex than that of the simple intended one-pole RC circuit. Even more parasitics, e.g. from coupling to the silicon substrate, occur and they are all properly extracted using the 3D BEM method as explained in the next section. The resistances are computed using the 2D FEM method as described earlier in this section. The result is a relatively complex RC network, that accurately models the dynamic behavior of the layout after fabrication, but that would be too complex for further processing. The circuits on the right of Figure 3 show the results of applying the SNE model reduction method for three different f_s values.

BEM based capacitance extraction

In order to extract the interconnect capacitances, several methods can be employed. These include heuristic, empirical formula based methods as well as numerical methods such as FEM and BEM. BEM (Boundary Element Method) based methods can be applied when good accuracy and efficiency is needed, but not the flexibility of the FEM. Here we will describe some model order reduction aspects of the BEM.

Basically, BEM techniques discretize the outside interconnect surface, and compute the potential induced in each discretization panel resulting from a unit charge in another panel starting from a integral equation as follows: $\phi(x_0) = \int_{allcharge} G(x;x_0)\xi(x)dx$. Here $\phi(x_0)$ denotes the potential in x_0, $G(..)$ is the so-called Green's function which is characteristic for the dielectric properties of the medium and $\xi(x)$ is an assumed charge density. Discretization then leads to a system of equations

$$C_s = A^T G^{-1} A \quad , \quad G_{ij} = \frac{1}{|\Gamma_j|} \int_{\Gamma_j} G(x_j;x_i) d\Gamma_j \tag{5}$$

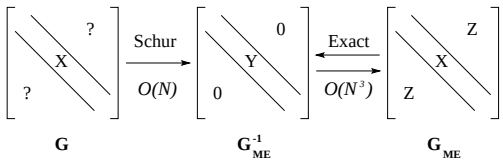

Figure 4: *Illustration of the Schur algorithm matrix extension properties*

where C_s gives the capacitances to be computed, A is a $1-0$ incidence matrix relating conductors to discretization panels and Γ_j denotes integration over panel j.

From Equation (5) a time complexity of $O(N^3)$ can be inferred, where N is the number of panels, which is linear in the size of the layout. Such a time complexity is actually unacceptable in practice, and many approaches for speeding up BEM-based capacitance extraction have been developed. One particularly popular approach is multipole-accelerated GMRES based solution [3]. In the next subsection, we will discuss a direct solution method that simultaneously works as an effective model order reduction technique.

Schur Method for BEM Speedup and Model Order Reduction Matrix G in (5) is in principle a full matrix. However, many entries are small, they correspond to a weak coupling between distant panels. It would be tempting to introduce sparsity by making them zero, but this would render the system non physical and it would compromise the results.

The Schur method, on the other hand, creates zeros in the inverse of G. Moreover, it does so by leaving the corresponding entries in G unspecified. More specifically, the Schur method produces the inverse of the so-called maximum entropy extension of a partially specified matrix G. This inverse is the unique matrix that has zeros at the positions that are unspecified in G and upon exact inversion (i.e. $(G_{ME}^{-1})^{-1} = G_{ME}$) coincides with G at the positions that were specified. This is schematically illustrated in Figure 4.

Referring to the physical situation, the specified entries of G correspond to capacitive coupling over distances smaller than a user-specified window distance w only. (In practice, this scheme is actually somewhat more involved, giving rise to a double-band sparsity scheme and the so-called hierarchical Schur algorithm, see [4].) In effect, the Schur algorithm thus is a method for model-order reduction, as only relevant coupling capacitances (associated with short distances) are computed.

This reduced order modeling is provably good in theory, and also performs well in practice. See [6] for details. Hence, circuit simulation runs faster because of the reduced model complexity, but also the runtime of Schur-accelerated BEM is reduced. It becomes, for a fixed window distance, linear in the problem size: $O(Nw^3)$. The memory becomes bounded by $O(w^2)$.

Conclusion

In this paper, we have argued that reduced-order modeling techniques are of great importance for VLSI design. We have demonstrated, using examples from FEM and BEM, that much can be gained by proper interpretation of the original electro-magnetics (electro-statics) problems. For example, The FEM model is efficiently reduced through exploiting both physical phenomena (articulation nodes) and efficient implementation (delayed frontal solution). The BEM is model is efficiently reduced during its construction through an approximative matrix inversion algorithm (the Schur algorithm). When both capacitance and resistance are modeled, frequency dependent behavior is included in the model reduction procedure resulting in a comprehensive overall model. It is hoped that such modeling philosophies can cross-fertilize other application areas. More details of our techniques can be found in [9].

References

[1] P. J. H. Elias and N. P. van der Meijs. Extracting circuit models for large RC interconnections that are accurate up to a predefined signal frequency. In *Proc. 33rd Design Automation Conf.*, pages 764–769, Las Vegas, Nevada, June 1996.

[2] P. J. H. Elias and N. P. van der Meijs. Including higher-order moments of RC interconnections in layout-to-circuit extraction. In *Proc. European Design and Test Conf.*, pages 362–366, Paris, France, March 1996.

[3] K. Nabors and J. White. Fastcap: A multipole accelerated 3-d capacitance extraction program. *IEEE Trans. on Computer-Aided Design*, 10(11):1447–1459, November 1991.

[4] H. Nelis, E. Deprettere, and P. Dewilde. Approximate inversion of positive definite matrices, specified on a multiple band. In *Proc. SPIE 88*, San Diego, California, August 1988.

[5] C. E. Leiserson T. H. Cormen and R. L. Rivest. *Introduction to Algorithms*. MIT Press, Cambridge Massachusetts London, England, 1998.

[6] N. P. van der Meijs. *Accurate and Efficient Layout Extraction*. PhD thesis, Delft University of Technology, Delft, The Netherlands, January 1992.

[7] N. P. van der Meijs and A. J. van Genderen. Delayed frontal solution for finite-element based resistance extraction. In *Proc. 32nd Design Automation Conf.*, pages 273–278, San Francisco, California, June 1995.

[8] A. J. van Genderen and N. P. van der Meijs. Extracting simple but accurate RC models for VLSI interconnect. In *Proc. Int. Symp. on Circuits and Systems*, pages 2351–2354, Helsinki, Finland, June 7-9 1988.

[9] A. J. van Genderen and N. P. van der Meijs. VLSI modeling and verification, May 1994. World Wide Web home page of the modeling and verification project, available at URL http://cas.et.tudelft.nl/space.

[10] A. J. van Genderen and N. P. van der Meijs. Using articulation nodes to improve the efficiency of finite-element based resistance extraction. In *Proc. 33rd Design Automation Conf.*, pages 758–763, Las Vegas, Nevada, June 1996.