Power analysis of large interconnect grids with multiple sources using model reduction

Eli CHIPROUT* Tuyen NGUYEN*

Abstract

Power simulation of metal interconnect voltage dissipation in large on-chip VLSI networks requires long CPU times on SPICE-like simulators. Even standard model reduction techniques are not capable of reducing this bottleneck significantly. A new technique allows for the simulation of these large networks with multiple input sources with few generated moments. The algorithm is an extension of the general multi-input multi-output (MIMO) model reduction approach and requires only a 12-20 Taylor series moments for networks with hundreds of unique input sources as compared to the MIMO approach which requires hundreds.

1. Introduction

Due to the increasing electromagnetic effects of on-chip metal interconnect in VLSI circuits, interconnect modeling and analysis have become prominent in the CAD community in the last few years. Most of the focus however has been on delay or noise analysis [1][2] which have been impacted by self or mutual capacitance and inductance.

Interconnect modeling tends to generate very large linear circuits composed of inductors, capacitors and resistors as well as independent voltage/current sources. Simulating these large networks had become a bottleneck for normal nonlinear circuit simulators. In this context, model reduction (MR)[3][4][5] has been useful. MR has been effective in allowing the reduction of these networks of thousands of nodes to networks of only tens of state variables.

Recently, power dissipation concerns in VLSI design have led to the modeling and simulation of large linear interconnect grids for interconnect power analysis and optimization. Power dissipation has become a serious concern on chip and a large part of the power dissipation takes place in the interconnects. Further, local voltage spikes or ebbs on the Vdd (or Gnd) supply lines can slow down or speed up devices that were designed to operate at fixed voltage supplies. This in turn can cause timing errors that were unanticipated by assuming constant power supplies.

The large power networks generated in modeling power supplies however, are not amenable to standard model reduction because they have thousands of independent sources which represent the circuits drawing current from the network. Applying the standard model reduction algorithms on these networks has not been effective because model reduction time is directly proportional to the number of input sources. Therefore these power networks are traditionally simulated with SPICE[6] which requires long CPU times.

We present a new algorithm which extends the model reduction approach to the time-domain analysis of large linear grids with multiple independent sources. The sources, like the network itself, are also expanded in a Taylor series and incorporated directly in the model reduction. This has the consequence of allowing the solution of the network with only one matrix inversion and a 12-20 back solves. Results obtained show that this method is accurate to the criterion required by power design and optimization.

The first part of this paper will describe how power networks come to be generated and discuss the issues of interest. Part 2. will overview model reduction while part 3. will describe the new approach. Finally, an example will show the application of the technique.

2. Power Analysis of interconnects

There are two sources of power dissipation in a VLSI circuit: the devices and the interconnect. Typically, in a standard chip design, groups of devices performing a particular sub-function are grouped together with small local interconnect into a module called a macro. Following this, different functional macros are physically placed and interconnected with longer metal interconnect lines, called global interconnects, into the final chip assembly. Each macro, in addition to being fed by the signal lines from other macros, is fed on the top by Vdd and Gnd buses. A small example macro is shown in Fig. 1.

When full-chip power analysis is performed, each macro is simulated by itself with typical and expected input vectors and its typical (time-dependent) average current draw at its “pins” is calculated (assuming unlimited power supply and therefore a stable Vdd supply). The interest of the designer is to estimate whether all of the macros operating together with their global interconnects will tax the Vdd network to such an extent as to cause power surges. In order to arrive at a

*The authors are with the IBM Austin research laboratory, Austin, Texas 78758, USA.

conclusion, each macro is modeled as one or more simple time-dependent current draws using the results of the stand-alone simulation. The global interconnects are extracted and modeled as R(L)C networks and the macro current sources are attached (Fig. 2). This results in a completely linear network. Simulating this network will give the true Vdd and Gnd supplies as functions of time and allow, if necessary, another iteration on the macro simulation with the more accurate power supply picture.

Realistic networks can contain hundreds or thousands of independent networks and thousands or tens of thousands of interconnect nodes making it a simulation bottleneck on standard simulators. We show next why standard model reduction cannot be applied to such networks.

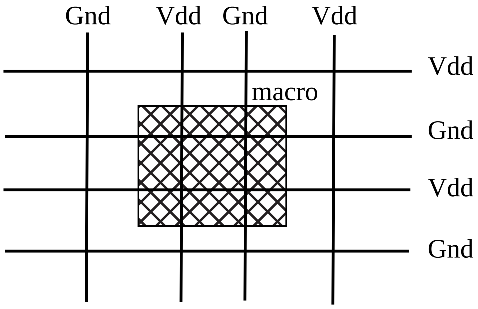

Figure 1: *An example macro fed by Vdd and Gnd buses at four input points on the grid.*

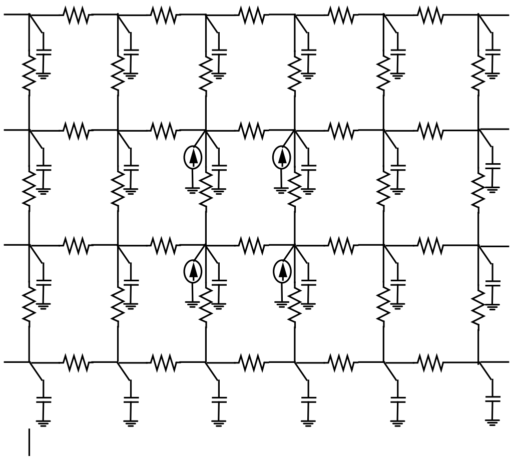

Figure 2: *A model of the macro drawing current from the power network at the four points.*

3. Model Reduction

The network in Fig. 2 (with or without inductors) can be formulated in terms of a set of MNA[7] network equations in the form of:

$$s\boldsymbol{C}\boldsymbol{x} = -\boldsymbol{G}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u}_p$$
$$\boldsymbol{i}_p = \boldsymbol{L}^{\boldsymbol{T}}\boldsymbol{x} \qquad (1)$$

where $\boldsymbol{C}$ and $\boldsymbol{G}$ represent the frequency independent NxN conductance and susceptance MNA matrices, $\boldsymbol{x}$ is the vector of node voltages and inductor/source currents, $\boldsymbol{u}_p$ and $\boldsymbol{i}_p$ are the currents/voltages at the input and output ports respectively, and $\boldsymbol{B}$, $\boldsymbol{L}$ are the matrices mapping the input ports (internal states) to the internal states (output ports).

In on-chip power analysis, matrices $\boldsymbol{C}$ and $\boldsymbol{G}$ are large and consist of thousands of state variables. The typical model reduction algorithm takes a linear network in the frequency domain form of (1) and generates the Krylov subspace:

$$span\{\boldsymbol{G}^{-1}\boldsymbol{B}, \boldsymbol{G}^{-1}\boldsymbol{C}\boldsymbol{G}^{-1}\boldsymbol{B}, \ldots, (\boldsymbol{G}^{-1}\boldsymbol{C})^n\boldsymbol{G}^{-1}\boldsymbol{B}\ldots\} \qquad (2)$$

using it in a transform on the system in (1) to yield a much smaller system in the same form as the original system:

$$s\boldsymbol{C}_r\tilde{\boldsymbol{x}} = -\boldsymbol{G}_r\tilde{\boldsymbol{x}} + \boldsymbol{B}_r\boldsymbol{u}_p$$
$$\tilde{\boldsymbol{i}}_p = \boldsymbol{L}_r^{\boldsymbol{T}}\tilde{\boldsymbol{x}} \qquad (3)$$

The transform can be a Lanczos type transform[5] or other transforms such as Arnoldi or others.

However, the bottleneck in such approaches is the number of sources in the vector $\boldsymbol{u}_p$. The longer this vector (i.e. the more unique input vectors exist) the bigger the Krylov subspace to be used because the bigger the $\boldsymbol{B}$ matrix. For each extra column in the $\boldsymbol{B}$ matrix, another forward-backward LU substitution is required in order to obtain a new Krylov vector. Therefore the time of the model reduction algorithm is directly proportional to the number of inputs/outputs (ports) and typically networks with only a small number of ports are reduced.

Another, older, approach to model reduction makes use of a explicit rational Padé approximation[4]. The Taylor series moments of (1) are generated at the selected outputs $\boldsymbol{i}_p(s)$:

$$\boldsymbol{L}^{\boldsymbol{T}}(\boldsymbol{G} + s\boldsymbol{C})^{-1}\boldsymbol{B}\boldsymbol{u}_p = m_0 + sm_1 + \ldots + s^n m_n \qquad (4)$$

This in turn allows one to form an approximation of the form:

$$\boldsymbol{i}_p(s) = \sum_i \frac{k_i}{s - p_i} \qquad (5)$$

This procedure, while useful for many circuits, fell out of favor for the general case because of its lack of numerical ability to extract more than 6-10 poles (p_i).

4. Model reduction for power

In order to solve the power problem more effectively, we need to note some important points about the problem. Power fluctuations of interest are typically low frequency. There is not much interest in the high frequency (more that 100 times the clock frequency) activity which is typically randomly balanced out by other high frequency activity in the circuit and does not cause much timing variation. What is of interest is prolonged (over one or more clock cycles) current draw that will have impact on the operating speeds of the macros over a one or more clock periods. This automatically implies that only a few low frequency poles in a model reduction are of interest.

As well, in order to be effective in simulation time, the ideal is to look at small time windows of suspected activity in some part of the network and therefore the simulation time for any window need not be over more then one or two clock cycles and several small windows of simulation time can be combined. Also, random time windows can be chosen to see if general activity is disrupting expected supply levels.

As a final point, the sources of each of the macros can be modeled for the low frequency purposes of power as piecewise linear sources.

These being the case, it is possible to use a variation of the explicit Padé model reduction described above.

First we begin with the system description in (1) and note that the input function $\boldsymbol{u}_p$ consists of multiple independent sources, each of which is piecewise linear. In the frequency domain these can then be modeled as a sum of delayed ramps:

$$\boldsymbol{u}_p(s) = \sum_i \frac{r_i \exp(-t_i s)}{s^2} \tag{6}$$

where the r_i is the slope of ramp *i* and t_i is the delay of ramp *i* from t=0. The Taylor series expansion of $\boldsymbol{u}_p$ is given by:

$$\begin{aligned} \boldsymbol{u}_p(s) = \frac{1}{s^2}\sum_j \sum_i r_i \frac{(-t_i s)^j}{j!} &= \\ \frac{1}{s^2}(\boldsymbol{u}_0 + s\boldsymbol{u}_1 + s^2\boldsymbol{u}_2 + \dots) & \end{aligned} \tag{7}$$

It is important to note that this series represents the combined input of all of the voltages to the system. Given the Taylor series of these combined sources, a Taylor series of the entire network response can be constructed as a Taylor series of $x(s)$ by including the sources as part of the model reduction:

$$(\boldsymbol{G} + s\boldsymbol{C})^{-1}\boldsymbol{B}\boldsymbol{u}_p = \frac{1}{s^2}(m_0 + sm_1 + \dots + s^n m_n) \tag{8}$$

In a normal model reduction algorithm, the sources are abstracted to impulse or step sources and the resultant solved output is convolved with any arbitrary input. In this new algorithm the sources can be numerous and varied. The series is premultiplied by a $1/s^2$ term in order to make them a more accurate response to input terms that are of the same kind in $\boldsymbol{u}_p(s)$. This yields a relationship with the moments of the input sources:

$$\begin{aligned} (\boldsymbol{G} + s\boldsymbol{C})(m_0 + sm_1 + \dots + s^n m_n) &= \\ \boldsymbol{B}(\boldsymbol{u}_0 + s\boldsymbol{u}_1 + s^2\boldsymbol{u}_2 + \dots) & \end{aligned} \tag{9}$$

which then gives a recursive relationship for the moments of the outputs:

$$\boldsymbol{G}m_0 = \boldsymbol{B}\boldsymbol{u}_0 \text{ and} \tag{10}$$

$$\boldsymbol{G}m_i + \boldsymbol{C}m_{i-1} = \boldsymbol{B}\boldsymbol{u}_i \tag{11}$$

Since the inputs of the macros are repetitive the number of ramps is always limited.

Note, that the output can be separated into:

$$\begin{aligned} x(s) = \left(\frac{1}{s^2}m_0 + \frac{1}{s}m_1\right) &= \\ (m_2 + sm_3 + \dots + s^n m_n) & \end{aligned} \tag{12}$$

The first parentheses contain what can be converted to a time domain ramp and a time domain step function. The next parenthesis can be approximated as (5) by an explicit Padé approximation. This approximation is obtained by mapping the Taylor series to a rational function approximation and then solving two sets of linear equations, one for the coefficients of the denominator and one for those of the numerator, and subsequently reducing it into a pole residue in the form of (5)[4].

The pole/residue form in the frequency domain is translated into the time domain as:

$$\sum_i k_i e^{p_i t} \tag{13}$$

One point of interest to note is that because transient, possibly nonmonotonic, sources are now introduced into the model reduction, there is no guarantee that for real-pole systems the generated reduced order poles will not be complex, (even though the network may contain only R's and C's!). In fact, for any realistic example the response will contain complex poles.

5. Example

We simulated a modeled power supply network containing 1128 capacitors, 2475 independent sources and 45561 resistors extracted from a 1GHz PowerPC microprocessor test chip containing 620 macros designed in CMOS6X (a 0.25 micron, 6-metal, 1.8 volt process) at the Austin research lab. For the frequency of analysis and power estimation time of interest we only extracted 4 poles/residues for each output. Some of these were complex. The poles/residues represent the complete response of the network to the given inputs for the time period of interest. Some typical Gnd and Vdd node outputs are displayed in Fig. 3. Note the sag of some of the Vdd or Gnd lines in part of the clock cycle. A histogram of a particular node's Gnd voltage distribution over time is shown in Fig. 4. As can be seen most of the time is not spent in a true Gnd state.

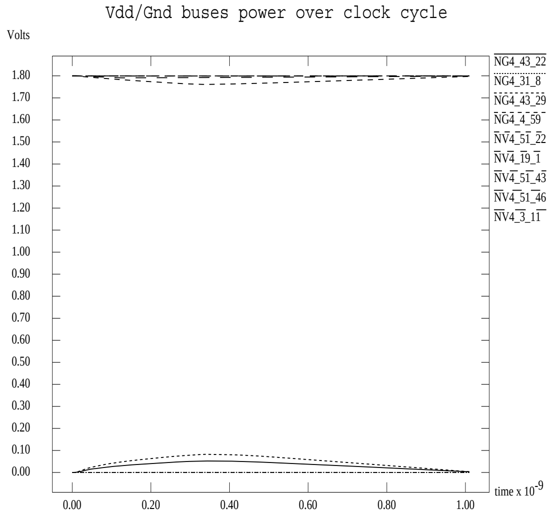

Figure 3: *Nine sampled power grid outputs of example (the lower are Gnd grid points and the higher are Vdd)*

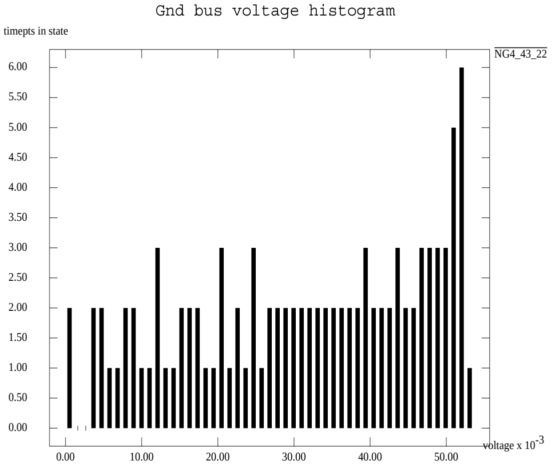

Figure 4: *Distribution of Gnd voltages in clock cycle*

6. Conclusion

We have demonstrated a novel model reduction method that is applicable to power simulation of large linear networks with a large number of independent sources. As compared to the conventional reduction techniques, the new method does not require moments proportional to the number of inputs but assimilates all of the inputs directly into the reduced model. Future research should concentrate on possibly extending this approach to a Krylov based model that is not dependent on explicit Padé approximations which will allow more time accuracy.

Acknowledgments

We would like to that Sani Nassif of the Austin research lab for providing generating the power networks from the microprocessor design.

References

[1] B. Krauter and S. Mehrotra, "Layout based frequency dependent inductance and resistance extraction for on-chip interconnect timing analysis", Proc. Design Automation Conference (DAC), San Francisco, 1998.

[2] A. Devgan, "Efficient coupled noise analysis for full-chip RC interconnect networks", Proc. International Conference on Computer Aided Design (ICCAD), San Jose, 1997.

[3] L. Pillage and R. A. Rohrer, "Asymptotic Waveform Evaluation for timing analysis", IEEE Trans. on Computer-aided design, 9(4):352-366, April, 1990.

[4] E. Chiprout and M. Nakhla, "Asymptotic Waveform Evaluation", Kluwer Academic Press, Norwell, MA, 1994.

[5] P. Feldman and R. W. Freund "Efficient linear circuit analysis by Pade approximation via the Lanczos process", IEEE Trans. on Computer-aided design, 14:639-649, 1995.L.

[6] W. Nagel, "SPICE2, a computer program to simulate semiconductor circuits", Technical report ERL-M520, UC-Berkeley, May 1975.

[7] C. W. Ho, A. E. Ruehli and P. A. Brennan, "The modified nodal approach to network analysis", IEEE Trans. on Circuits and Systems, CAS-22:504-509, June, 1975.