Scalable Parallelization of Harmonic Balance Simulation

David L. Rhodes[1] and Apostolos Gerasoulis[2]

[1] US Army CECOM/RDEC, Fort Monmouth, NJ 07703, USA
d.l.rhodes@ieee.org
[2] Rutgers University, Piscataway, NJ 08854, USA
gerasoul@cs.rutgers.edu

Abstract. A new approach to parallelizing harmonic balance simulation is presented. The technique leverages circuit substructure to expose potential parallelism in the form of a directed, acyclic graph (dag) of computations. This dag is then allocated and scheduled using various linear clustering techniques. The result is a highly scalable and efficient approach to harmonic balance simulation. Two large examples, one from the integrated circuit regime and another from the communication regime, executed on three different parallel computers are used to demonstrate the efficacy of the approach.

1 Introduction

The harmonic balance simulation method is widely used in the arena of large-signal, steady-state analysis of nonlinear circuits. It is often times applied to high frequency electronics since it directly accommodates frequency domain models [8, 9, 4]. Direct support for frequency domain models allow harmonic balance to be preferable in situations where distributed elements or frequency-dependent effects are important, for example, where accurate dispersive transmission line analysis (*e.g.* due to skin-effects, dielectric properties) is necessary. Such conditions are typical in radio-frequency (rf) and microwave circuits and increasingly arise in high-speed, deep sub-micron integrated circuits (IC) as well.

Nonlinear circuit simulations, of any type, are computationally intensive and therefore there have been several efforts to speed them via the use of parallel computation. Since linear matrix operations lie at the foundation of many nonlinear circuit simulation techniques, including harmonic balance, it is important to consider a variety of previous circuit simulation parallelization efforts. For example, several non-harmonic balance analog circuit simulators have been speed-up by parallelization of their inherent linear matrix operations [11, 3, 7]. A parallelization of the (linear part of the) harmonic balance method has been demonstrated [13]. While this effort provided 97% processor efficiency this approach—as well as the others cited—all display limited *scalability*.

Scalability is a measure of the ability of the parallelization technique to *efficiently* make use of an increasing number of available processing elements (PEs).

In the ideal, speed-up scales linearly with an increasing number of PEs since *potential* computational power is increasing linearly. Thus, a method which maintains close to linear speed-up over a wide range of available PEs is called scalable. Since we are interested in an efficient technique for today's parallel processors which contain tens to hundreds of PEs, scalability is an important concern.

The method in [13] parallelized the entire analysis of the linear portion of the harmonic balance method on a per-frequency basis, and thus the scalability of this approach is completely limited by the number of frequencies required in the analysis—adding PEs beyond the number of frequencies for the particular input file causes no additional speed up. Alternatively, methods which parallelize the underlying mathematics as it arises in circuit simulation have not, as of yet, achieved reasonable scalabilty. For example, the results in [11, 3] show efficiencies that are rapidly falling off even for ten PE computing systems. A peak efficiency of about 38% for 8 processors (about $3\times$) has been shown, but speed up then *decreases* beyond 8 processors [7].

A different parallel approach which leverages *circuit substructure*, as we do here, was demonstrated for nonlinear transient domain analysis using the waveform relaxation technique [16]. This approach yielded somewhat better results (efficiencies) than the mathematically oriented approaches but also shows signs of scalability limitations. For example, efficiencies of about 30–60% are shown for ten processors (about $6\times$ speed up), while the method here achieves about 40% efficiency on 64 processors (*i.e.* $25\times$ speed-up) and about 25% efficiency for 128 processors ($32\times$ speed-up). This is for the harmonic balance technique, of course, and not waveform relaxation as in [16].

2 Harmonic Balance Technique

While an overview of the harmonic balance method is given with the aid of Figure 1, please see the references for a full treatment of the technique [8, 9, 4]. The harmonic balance technique divides the circuit or system description into two portions: *linear* and *nonlinear*, interconnected at the *interface* (a set of circuit nodes). The 'balance' then entails iterative adjustment (using an optimization procedure) of the harmonic voltages at the interface until the harmonic currents as given by the linear and nonlinear sides 'agree'. This is done for a fixed set of frequency points, or harmonics, as dictated by the input parameters (the term frequency will be used henceforth for consistency). Sometimes only a few frequency points are used, say for nonlinear amplifier studies, while in other cases hundreds to thousands of frequency points are used, *e.g.* when complete time domain information is important. The 'analysis' on each side of the interface is quite different, but in each case we are interested in computing the frequency-domain currents at the interface given the frequency-domain voltages.

During the balancing iterations, each of the linear- and nonlinear-sides must compute frequency-domain currents given frequency-domain voltages at the interface and at external ports. For the **linear-side**, this is computed as a matrix multiplication (at each frequency) of the frequency-domain voltages with the

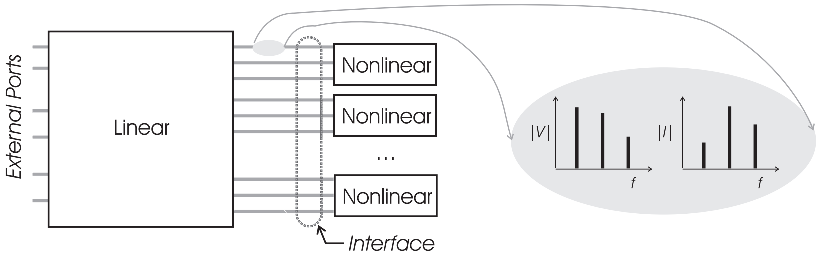

Fig. 1. Visualization of the harmonic balance method

equivalent ('reduced') *admittance matrix* of the linear portion of the circuit. Thus, the linear portion of the circuit may be reduced to its equivalent admittance matrix at each frequency point *once* and then subsequent 'analyses' during the balance process are merely matrix multiplications.

In order to compute the equivalent admittance matrix (at each frequency) of the linear side, circuit nodes that are internal to the linear side and not connected to either of the interface or external ports must be reduced. As it turns out from Kirchoff's currect/voltage laws (KCL and KVL respectively), Gaussian elimination becomes the basic step of such an equivalent reduction. Assuming the typical matrix notation [8, 9, 4], an internal node, k, may be equivalently reduced in the admittance matrix using the expression:

$$Y^f_{i,j} = Y^f_{i,j} - \frac{Y^f_{i,m} Y^f_{m,j}}{Y^f_{k,k}} \qquad \forall m \neq k \quad \text{and} \quad \forall i, j, f \tag{1}$$

where $\mathbf{Y}^f$ is the admittance matrix for a single frequency point f. Assuming that the reduction is to a constant number of nodes, this reduction can be seen to be of *cubic* order; *i.e.*, as linear circuit nodes are added, computational time for the reduction to a fixed-sized admittance matrix increases in a cubic fashion.

On the **nonlinear-side**, each nonlinear model appears connected individually at the interface. Figure 1 only shows example models which have three nodes (as typical of transistors), but any number greater than or equal to two is possible. As is the case for the linear side, nonlinear model evaluation requires computation of frequency-domain currents given the frequency-domain voltages; these models typically make use of Fourier domain transform techniques to allow time-domain modeling.

The overall current at each interface node (at each frequency) is the sum of current contributions from both the linear and nonlinear sides. In accordance with KCL, this current should be zero. Thus an appropriate error function widely used in harmonic balance uses the L2 norm:

$$error = \left(\sum_{j \,\in\, \text{interface nodes}} \; \sum_{k \,\in\, \text{frequency}} I^2_{j,k} \right)^{1/2} \tag{2}$$

where $I_{j,k}$ is the sum of all currents entering the interface node indexed j at frequency k. Finally, a conjugate-gradient optimization routine adapted from MINPACK [1] performs the 'balance' by iteratively adjusting the frequency-domain voltages until an acceptably small current *error* is reached. Of course, at each frequency-domain voltage revision, the linear and nonlinear evaluations described must be used.

3 A Scalable Parallelization Technique

The harmonic balance algorithm, irrespective of parallelization, is then:

Form linear matrix and reduce to nodes of external ports and nonlinear models
Guess harmonic voltages at interface
forever
  Compute harmonic currents from linear side (matrix multiplication)
  Compute harmonic currents from nonlinear side (model evaluation)
  Evaluate *error* (Eq. 2)
  if *error* is acceptably small **then done**
  Update harmonic voltage guess (via conjugate-gradient optimization)

Within this analysis framework, several parallelization techniques are possible. In general, since only a few components are interconnected in typical circuits, even as the circuit itself becomes large, the admittance matrices—or other expressions of circuit equations—tend to be *sparse* [14]. Even with this sparseness characteristic, parallelization of the matrix solutions in other circuit simulators has met with limited scalability [11, 3]. On the linear side only, parallelization of the entire process of model evaluations, matrix fill and reduction was shown effective, but that approach has scalability limited to the number of frequencies to be analyzed [13].

At first, it might appear that the harmonic balance iterations will dominate computation time as these are looped. However, as mentioned earlier the linear portion of the computation, represented entirely on the first line of the algorithm, actually dominates many harmonic balance calculations. As the circuit grows, linear fill and reduction tend towards cubic order, while the balance remains close to linear. This isn't always true of course, but many rf/microwave circuits do not exhibit strong interaction among nonlinear elements (unlike digital switching circuits)—although there are certainly exceptions. Thus, in many cases the linear part of the analysis dominates.

Obviously, a critical element for any parallelization technique is that sufficient parallelism be *exposed*, with a granularity that is not too small. **Granularity** may be loosely defined as the ratio of useful computation to inter-PE communication time. It is therefore a function of the particular multi-processor computing resource, even for a fixed problem—thus a technique is needed which both exposes a good deal of parallelism with a granularity useful for the computing resource at hand. The method here is just such a technique, leveraging circuit substructure to expose medium grain parallelism.

An introductory example will help clarify the general technique; first consider the linear part of harmonic balance by the example circuit, composed using subcircuits, as shown in Figure 2. In Figure 2.A, the circuit A has an instance of circuit $B1$ and $B2$ in it. In turn, $B1$ has instances of $C1$, $C2$ and $C3$ in it while $B2$ contains $C4$ and $C5$. The dark circles are circuit nodes, shown with their node labels.

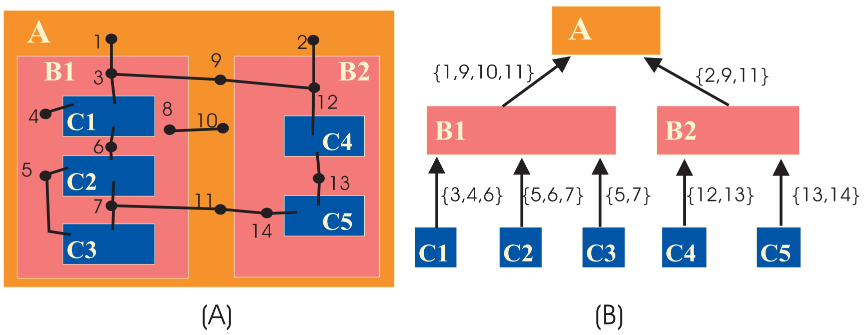

Fig. 2. (A) Linear subcircuit structure; (B) computational structure resulting from the method, where arrows represent data-dependencies

A typical analysis method would fully elaborate the circuit, forming a very large, sparse matrix [14] for the entire circuit description including all circuit nodes in all subcircuits. Alternatively, each subcircuit's internal nodes can be reduced prior to fill into the next level; with this technique, a full elaboration never exists. In this sense, such a method takes advantage of the sparsity present in circuit descriptions but without use of explicit sparse matrix techniques. Figure 2.B shows the computational flow of this approach for the example in Figure 2.A. The computation represented by each box is: (i) **matrix fill** in which requires admittance matrix evaluation of local circuit element models and use of subcircuit admittance matrices in the fill, and (ii) **matrix reduction** which then reduces internal circuit nodes to allow presentation of the reduced admittance matrix to the next level. Note that the boxes in Figure 2.B represent the *computation* required for this subcircuit and the arc represents the passing of admittance information to the next hierarchical level. Arcs are labeled using the node names for the equivalent, reduced subcircuit as it must appear at the higher level. Since subcircuit matrix information is required prior to computation, this approach gives rise to a directed, acyclic graph (dag), actually a special form called an **in-tree**, of computations.

Note that the linear matrix reduction computation shown in Figure 2.B is for a single frequency, each frequency point in the analysis requires an independent computation, so the whole linear computation becomes a **forest** of in-trees. Inclusion of nonlinear models within the hierarchy is handled by 'carrying' the nodes attached to any nonlinear model 'up' the hierarchy. Such a procedure results in exactly the form Figure 1 while leaving internal linear-only node reduction in hierarchical form as in Figure 2.

For nontrivial circuits, this method exposes a good deal of parallelism, which exists across siblings for the in-tree at each frequency and across the trees themselves. The method in [13] can be viewed as a subset of this approach in that it parallelizes across entire frequency trees only. Since all communicated results are only needed by one consumer (the graph is an in-tree), a message-passing approach is appropriate. The current implementation is message-passing based and uses the Message Passing Interface (MPI) standard [6, 15].

Even though linear computation is the dominant and more interesting part of the overall technique, next consider opportunities for parallelization in the balance portion of the method. First, the ('forever') loops themselves are serially dependent—*i.e.* for some iteration i, the frequency-domain voltage update is done in loop $i-1$, while the update itself cannot be done until the i^{th} *error* is computed. Thus, only parallelization within a single step is possible (assuming that the basic algorithm itself is not changed). Within this loop the nonlinear current computations dominate—the linear current computations are small matrix multiplications (at each frequency), and the error evaluation and frequency-domain voltage updates are also small calculations. The optimization update, a non-sparse, second-order, conjugate-gradient matrix computation is also not readily parallelizable and also is of small size. This matrix size is exactly

$$2 \times \textit{number-frequencies} \times \textit{number-nodes-in-interface} \tag{3}$$

where the factor of 2 arises because the frequency-domain voltages are complex-valued numbers. More importantly, this calculation does not usually tend to be a significant contributor to the iteration computation time.

Fortunately, the nonlinear current computation is readily parallelizable, by parallelizing the evaluation of each nonlinear device, as can be visualized in Figure 1. Thus, the non-linear analysis becomes a straightforward parallelization of the nonlinear model evaluations; the next section describes the background and specific technique for parallelizing the forest of in-trees that arises from the linear computation.

4 Allocation and Scheduling

For the linear part of the method, the forest of in-trees must be allocated and scheduled on the number of PEs to be used—this resource constrained scheduling problem is readily recognized as NP-hard/complete [5]. A further complication is that computation times (runtimes) are not known. Therefore an allocation/scheduling approach which *does not require* graph node runtimes is used. It is interesting to note that node computation times for realistic inputs do in fact vary widely—by two to three orders of magnitude—giving rise to a very 'irregular' scheduling problem.

Two alternative approaches are possible; first a quick runtime estimation function could be developed and used in a more robust scheduling method. Secondly, since runtimes are usually not highly dependent on frequency, all nodes of the first frequency in-tree could be evaluated providing accurate runtime estimates for subsequent in-trees (*i.e.* re-schedule in-trees after the first using a scheduling method which requires runtime estimates). Of course, the scheduling time must be included in the net parallel efficiency achieved—meaning that only 'low-order' methods are possible. Based on observations and execution trace information, it appears that employing such a low-order schedulers will not improve the (good) results already achieved. Note that the overheads we observed in execution traces (not shown for space reasons) are consistent with other investigations [17, 2, 10].

The method makes use of a **linear clustering** applied to the in-tree structure. Linear clusters are just single dependency chains of the in-tree, each of these is then statically allocated to a PE by assigning each in order to a PE, modulo the number of PEs—this allocation is a *wrap allocation of linear clusters*. Within this static allocation, local scheduling of each PE is dynamic, each PE loops first from the lowest to highest frequency and then from lowest (tree) level to highest looking for a 'ready' computation.

The first results obtained using this method revealed that load-balancing remained problematic. Obviously there is no assurance that a balanced allocation will result from this, and in fact without node runtimes or at least estimates, no allocation method could make such a guarantee. To remedy this situation, **rotated allocation** across 'frequency' in-trees is used. That is, a single in-tree is clustered and wrap allocated, but then this static assignment is 'rotated' across frequencies (modulo the number of PEs). The culmination of these approaches is the final allocation/scheduling technique and that for which results are subsequently shown.

5 Examples and Results

In order to validate and demonstrate the approach, three different large-scale, parallel processors were used, with up to 128 PEs, an order of magnitude higher than presented in previous circuit simulators [3, 16]. These are: Cray's T3E, IBM's SP2, SGI's Origin-2000. Vendor supplied MPI routines were used in each case. Every effort has been made to ensure that the runtimes presented are accurate, however, there are a few factors that cannot be controlled. Since these machines are multi-user, various types of interference are possible including processor multi-tasking on some machines and communication contention. Nevertheless, multiple executions were done on 'quiet' machines and the data presented is felt to be correct. Two examples are executed on each of the processors with varying numbers of PEs.

The first example comes from the monolithic, microwave integrated circuit (MMIC) domain. It is a linear circuit hierarchically composed with 7 parametric circuit descriptions which, when elaborated, contains 169 subcircuit instances. It is analyzed at 20 frequency points, this gives 3,380 computational nodes total in the in-tree forest when elaborated across both circuit hierarchy and frequency. Figure 3 shows the results when using wrap clustering, (frequency index) rotating allocation, and dynamic local scheduling for the Cray T3E for up to 128 PEs. As can be seen, runtime is consistently improving all the way up to 128 PEs. Figure 4.A and Figure 4.B show similar results for the IBM SP2 for up to 64 PEs and the SGI Origin-2000 for up to 32 PEs. All graphs are plotted in log-log form since the ideal runtime curve becomes linear with a slope of -1, *i.e.* ideal runtime is $x/number\text{-}of\text{-}PEs$ where x is a single PE runtime.

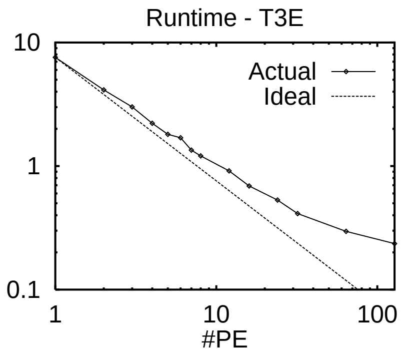

Fig. 3. Ideal and actual runtimes for the mmic circuit on the Cray T3E

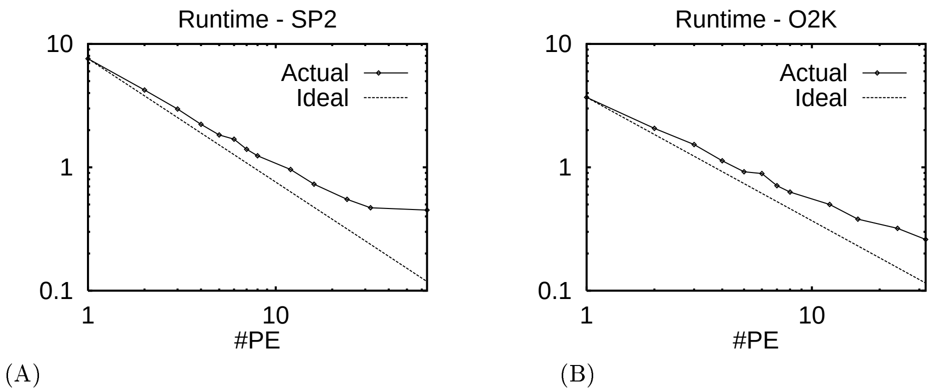

Fig. 4. Runtime data for mmic circuit for the IBM SP2 (A) and SGI Origin 2000 (B)

The second example file represents a set of 8 radios arranged in a circular fashion; each radio has a nonlinear transmitter and receiver, as well as filtering circuitry and an antenna model. A simple dispersive (loss dependent on frequency) model for the atmosphere is included. When elaborated, the circuit contains 16 nonlinear elements—a set of 6 frequency (harmonically related) points are analyzed. Figure 5.A shows the results for the linear part of the method only for this input. The parallelization here is not as good as for mmic on the T3E as shown in Figure 3 but consistent improvement up to 32 PEs is shown. Figure 5.B shows the results for the whole analysis. By examining the runtime difference for a single PE in Figures 5.A and 5.B, it can be seen that the nonlinear balance time is about 15 seconds of the 100 seconds run (this clearly illustrates the earlier statement that linear analysis time tends to dominate a harmonic balance simulation). As can be seen in Figure 5.B, *overall* parallel efficiency is not as good, but this is partly because the circuit only has 16 nonlinear elements, hence for any number of PEs over this amount the nonlinear portion of the method cannot be sped up.

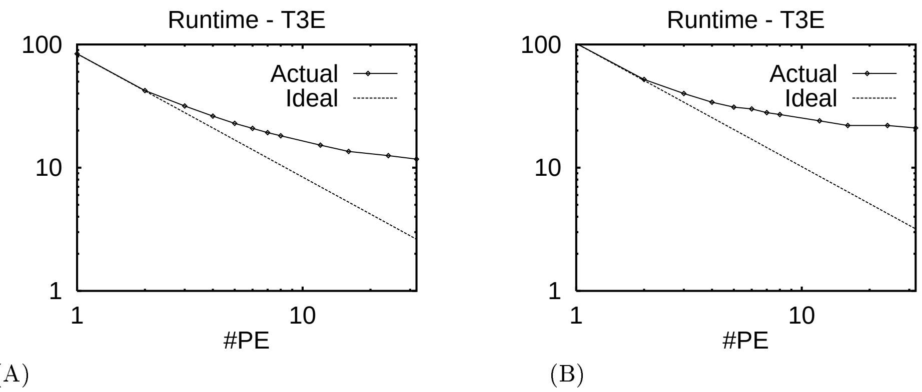

Fig. 5. Ideal and actual runtimes for linear portion only (A) and full analysis (B) for the cosite file executed on the Cray T3E

6 Concluding Remarks

A novel approach to the scalable parallelization of a circuit simulation problem has been developed. This includes a new approach to exposing parallelism as well as application-domain specific methods for allocation and scheduling. Measured speed-ups for three different parallel computers demonstrate the efficacy of the approach. This includes reaching 32 $\times$ speed up (on 128 PEs) which appears to be the highest speed up reported for any type of circuit simulation.

The authors would like to thank Professor Shirley Browne of the University of Tennessee Knoxville for assistance in obtaining execution trace results using the VAMPIR tool [12] and the High Performance Computing Modernization Office of the US DoD for partial project support.

References

1. http://netlib.att.com/netlib/minpack/.
2. Ghieth A. Adandah and Edward S. Davidson. Modeling the communication performance of the IBM SP2. In *Proc. of 10th IEEE Int. Parallel Processing Symp.*, pages 249–57, 15–19 April 1996. Honolulu, HI.
3. C. Chen and Y. Hu. A practical scheduling algorithm for parallel LU factorization in circuit simulation. In *Proc. of IEEE Int. Symp. on Circuits and Systems*, pages 1788–91, 8–11 May 1989. Portland, OR.
4. Benjamin Epstein, Stewart Perlow, David Rhodes, J. Schepps, M. Ettenberg, and R. Barton. Large-signal MESFET characterization using harmonic balance. In *IEEE Microwave Theory and Techniques-S Digest*, pages 1045–8, paper NN–2, May 1988. New York, NY.
5. Michael R. Garry and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W.H. Freeman and Company, New York, NY, 1979.
6. W. Gropp, E. Lusk, and A. Skjellum. *Using MPI: Portable Parallel Programming with the Message Passing Interface.* MIT Press, Cambridge, MA, 1994.
7. Volkhard Klinger. DiPaCS: a new concept for parallel circuit simulation. In *Proc. of IEEE 28th Annual Simulation Symposium*, pages 32–41, 9–13 April 1995. Phoenix, AZ.
8. K. Kundert and A. Sangiovanni-Vincentelli. Finding the steady-state response of analog and microwave circuits. In *IEEE 1988 Custom Integrated Circuits Conf. (CICC)*, 16–19 May 1988. paper 6.1, Rochester, NY.
9. K. Kundert, J. White, and A. Sangiovanni-Vincentelli. *Steady-state Methods for Simulating Analog and Microwave Circuits.* Kluwer Academic Publishers, Boston, MA, March 1990.
10. Yong Luo. MPI performance study on the SGI Origin 2000. In *IEEE Pacific Rim Conf. on Communications, Computers, and Signal Processing (PACRIM'97)*, pages 269–72, 20–22 August 1997. Victoria, CA.
11. K. Mayaram, P. Yang, J. Chern, R. Burch, L. Arledge, and P. Cox. A parallel block-diagonal preconditioned conjugate-gradient solution algorithm for circuit and device simulations. In *IEEE Int. Conf. on Computer-Aided Design (ICCAD'90)*, pages 446–9, 11–15 Nov 1990. Santa Clara, CA.
12. PALLAS GmbH, Brühl, Germany. *Vampir User's Manual, Release 1.1.*
13. D. Rhodes and B. Perlman. Parallel computation for microwave circuit simulation. *IEEE Trans. on Microwave Theory and Techniques*, 45(5):587–92, May 1997.
14. V. Rizzoli, F. Mastri, F. Sgallari, and V. Frontini. The exploitation of sparse-matrix techniques in conjunction with the piecewise harmonic balance method for nonlinear microwave circuit analysis. In *IEEE Microwave Theory and Techniques-S Digest*, volume 3, pages 1295–8, paper OO–5, May 1990. Dallas, TX.
15. M. Snir, S. Otto, S. Huss-Lederman, D. Walker, and J. Dongarra. *MPI: The Complete Reference.* MIT Press, Cambridge, MA, 1996.
16. E. Xia and R. Saleh. Parallel waveform-newton algorithms for circuit simulation. *IEEE Trans. on Computer-Aided Design*, 11(4):432–42, April 1992.
17. Zhiwei Xu and Kai Hwang. Modeling communication overhead: MPI and MPL performance on the IBM SP2. *IEEE Parallel and Distributed Technology: Systems and Applications*, 4(1):9–23, 1996.