Attachment 3

Parallel Adaptive Finite Element Software for Semiconductor Device Simulation*

Robert W. Dutton, Kincho H. Law, Peter M. Pinsky,
Narayana R. Aluru, Bruce P. Herndon

Stanford University
Stanford, CA 94305-4055

Abstract

Parallel algorithms and fully functional application codes for 2D and 3D device analysis of semiconductor devices have been demonstrated. Advanced modeling based on a hydrodynamic formulation (HD) of the semiconductor transport equations and using a Galerkin Least Squares Finite Element Method (GLS-FEM) has demonstrated nearly ideal parallel performance for 2D MOS and Bipolar transistor applications across Intel (512 node, Delta) and IBM (16 node, SP/1) machines. Parallelization of conventional drift-diffusion (DD) based device solvers has broken new ground in both direct and iterative solvers. A well-known application code, PISCES, has been parallelized and ported across Intel, TMC, and IBM architectures with best results to date that now approach 6.5 GFlops sustained performance on a 128 node IBM SP/2. A prototype 3D code (STRIDE) which uses iterative methods has parallelized preconditioners for ILU(0), ILU(1), and ILUV and achieved excellent benchmarks on both the Intel and IBM machines. A 4.9 million grid problem run on the Intel Delta machine achieved 20% efficiency using 512 nodes and convergent solutions for a highly nonlinear bipolar transistor problem in 20 minutes per bias point. In support of both 2D and 3D TCAD applications, a new geometry-based structure generator called VIP3D was created. Quad- and oct-tree utilities were developed and used to support the gridding of complex IC structures benchmarked in this work. Results of industrial impact and collaborative interactions are also discussed.

I. Introduction

The multi-dimensional analysis of semiconductor devices--both in two- and three-dimensions--provides the backbone of advanced technology development. Yet these analysis capabilities are computationally demanding and to date the infrastructure to support large scale 3D models has been insufficient to result widespread industrial use. At the same time, the scaling of IC devices and technology into the deep submicron regime in support of giga- and tera-flop applications increasingly demands such 3D technology computer-aided design (TCAD).

This contract targeted the development of advanced finite-element method (FEM) and other robust device analysis capabilities that can support parallel computational strategies to overcome analysis time and resource constraints. The leverage provided by powerful new parallel computers can thereby reduce the development time and costs for new advanced devices by orders of magnitude. Moreover, the advanced models and numerically stiff partial differential equations (PDE) used in this work serve as a key set of benchmarks for parallel computing that test the machines and algorithms in the context of practical applications.

---

*Excerpts from a final report on ARPA Contract # DAAL03-91-C-0043.

Hence, the objectives of this work focus on the complete parallelization of semiconductor device analysis codes and the benchmarking of their capabilities on state-of-the-art parallel computers. This work involves a range of supporting tasks and software technology. For example, new FEM technology as well as parallel element matrix assembly, nonlinear and linear solvers and supporting gridding technology all need to be developed and made functional in working applications. The approach taken in this work centers on three major areas of development:

1. Galerkin-Least-Squares finite-element methods have been used successfully for computational fluid dynamics. In this project, a similar formulation of the semiconductor equations is being developed. In addition, we are developing error estimators for the coupled system of elliptic Poisson and the hyperbolic advective-diffusive equations.

2. Both iterative and hybrid (direct/iterative) solution techniques are being pursued in order to enhance the robustness of simulations over the complete range of biasing and device configurations. This includes parallel preconditioners and iterative solvers for 3D problems; for 2D problems we are exploring robust iterative techniques.

3. Quad-tree/oct-tree-based gridding schemes are used to support the solver technologies and provide flexibility for adaptation and parallelization. Device geometry comes from a semiconductor wafer representation (SWR) and the methodology supports non-planar surfaces as well as dynamically changing geometry as a result of process simulation (a follow-on ARPA contract called the SPRINT-CAD).

During the course of this work there have been major results achieved in all these areas. The parallel benchmarks achieved on three commercial machines have now clearly demonstrated the viability of 3D TCAD using parallel computers. In addition, there has been rapid and substantial growth in both industrial use of the hardware technology and vendors now appear ready to support commercialization. The following sections are organized to discuss each of the areas listed above. The body of the report gives an overview of the work and major benchmarks as they relate to parallel computing. Finally, there is a section related to technology transition, both in terms of users and potential vendors.

II. Overview of Parallel Device Analysis

The analysis of semiconductor devices involve the solution of both carrier transport and electrostatics that result from boundary conditions imposed on the device. The Boltzmann Transport Equation (BTE) is the most general representation of the carrier dynamics and can be solved by a variety of methods, depending on simplifications used. Here we focus on the use of assumptions regarding carrier statistics and integration over momentum space that reduces the general BTE to a set of PDEs. Specifically, the two classes of transport considered here are: 1) the hydrodynamic (HD) formulation which involves the conservation of carrier number, momentum, and energy and 2) the drift-diffusion (DD) formulation where only the first moment of the BTE is used. The number of PDEs involved in each case is quite different; for the HD formulation, there are different levels of model complexity. The key point for discussion in the context of this project is the fact that over a range of models and number of supporting PDE systems for semiconductor device analysis, parallel solvers and complete applications have been demonstrated.

Both the HD and DD formulations are used in this work. In an effort to explore and understand the broadest needs and requirements for TCAD applications, both the FEM and a more traditional finite volume approach were considered. In the later case, Stanford has developed the well-known PISCES code for two-carrier analysis (holes and electron) based on the DD formulation. PISCES was later adapted for execution on message-passing distributed-memory parallel computers. This adaptation was modeled upon the earlier work of Lucas [1][2].

The HD analysis capabilities are built on both a new formulation of the semiconductor problem using Galerkin Least Squares (GLS) FEM and supporting matrix free solver technology that has been applied to Euler and Navier Stokes Analysis (ENSA) of computational fluid dynamics (CFD) problems. The overall capabilities in support of FInite Element Solver for TCAD Applications--will henceforth be referred to as FIESTA.

Problem decomposition is a crucial component in coarse-grain message-passing parallel PDE solvers. To achieve good parallel efficiency, the grid structure describing a particular problem domain must be equitably divided among the processors. Lucas' early 2D work used direct linear solution methods and focused primarily on well-structured problems for which straightforward coordinate bisection and nested dissection algorithms sufficed. This work was extended to 3D grids using iterative linear solution techniques by Wu, et.al [3]. The current parallel version of PISCES, PISCES-MP, has been used to explore several domain decomposition methods for unstructured grids. The method of choice is the Recursive Spectral Bisection (RSB) algorithm of Pothen, et.al. Due to extreme ill-conditioning inherent in the problem discretization, direct linear methods continue to be used and have been extensively benchmarked.

III. FIESTA and Benchmarks for the HD Formulation

There are classes of problems where time dependence of the solution and abrupt spatial variations require special consideration and the use of more robust numerical techniques. The field of computational fluid dynamics (CFD) is one such application area. In the TCAD domain, there are both process and device analysis problems where stiffness in time and space is important. In the context of this project, we have focused on analysis of carrier transport in ultra-small devices (i.e. deep submicron MOS and bipolar transistors) where spatially abrupt carrier distributions are of primary importance. For finite volume approaches, substantial effort has been invested in developing upwinding techniques such as the Scharfetter-Gummel approximation [5][6]. While FEM has been applied to semiconductor device analysis [7], the problems with current discretization in 2D have resulted in a hybrid implementation. The objectives of this task were two-fold: 1) to implement the semiconductor device equations in the context of advanced FEM formulation and 2) demonstrate both parallelization and computational benchmarks based on such a prototype code.

A Space-Time and Galerkin Least Squares FEM formulation is developed for the electron and hole HD device equations, and a Galerkin FEM is developed for the Poisson and lattice thermal diffusion equations. One major challenge and accomplishment of this work has been to symmetrize the HD system of equations by employing generalized entropy functions [9]. GLS FEM formulation based on the symmetrized system of equations is shown to satisfy the Clausius-Duhem inequality or the second law of thermodynamics, which is the basic stability requirement for nonlinear system of equations [10]. The governing equations are nondimensionalized to improve conditioning of the system of equations and a staggered approach is employed to treat the coupled HD, Poisson, and lattice equations [11]. As mentioned earlier, since the GLS-FEM has been initially developed in support of CFD applications, shock capturing operators naturally allow for highly nonlinear source terms that occur in the semiconductor problem.

The implementation has followed an SPMD (simple program, multiple data) paradigm to allow for generality as well as parallelization. Again, using leverage provided by work in the CFD community, the ENSA (Euler-Navier Stokes Analyzer) code was chosen as the basis for the solver [8]. The code provides flexibility due to many years of development and applications. Moreover, it runs efficiently on distributed memory, message-passing architectures by exploiting a matrix-free GMRES iterative solution method. Further details of the implementation are provided in references [12] [13][14][15].

Testing of the FIESTA code involves two aspects: 1) application-oriented evaluation of the results and 2) benchmarking of the parallel performance. It is well-known that n+_n_n+ diodes exhibit thermal non-equilibrium effects near the so-called drain terminal and are widely used as test problems. In this work, both 1D and 2D versions of the n+_n_n+ diode have been analyzed and results compared with others in the literature as well as an implementation of the energy transport formulation in the PISCES code (2ET) [16]. The primary focus of this project was directed toward the parallel benchmarks discussed below. The simulation of both submicron MOS and bipolar transistors was also demonstrated [12]. Again, the results showed excellent smoothness of solutions with no signs of numerical instability or difficulty in convergence. One feature of the GLS approach is the FEM discretization of time as well as space. In the work reported here, the use of the time-dependent aspect of the formulation provides guaranteed stability at the expense of an artificial time-stepping of the solution. On the other hand, in anticipation of the application to TCAD process simulation which is definitely time dependent, the exploration and parallelization of the method provides an excellent platform for further use in the SPRINT-CAD project.

The parallel benchmarking of FIESTA on both the Intel and IBM parallel machines provides clear evidence of the potential in support of other TCAD applications. From a practical engineering perspective, for a 2D bipolar problem with 22,000 grid points obtained on the Caltech Delta, the results show the ability to achieve an accurate solution in 40 minutes on a 512 node parallel machine. The same problem (and program) when run on an IBM RS 6000/560 fully-loaded workstation took 40 hours. This parallel improvement has both quantitative and qualitative implications--the ability to real-time engineering and innovation depends critically on obtaining timely feedback of information. More recent benchmarks obtained on the 16 node IBM SP/1 machine at Stanford [17] give the CPU time comparisons made for several interesting device applications on serial and parallel computers. For complicated device structures that require large grid sizes, the serial computers are not only inefficient for practical engineering simulations but are also inadequate. The dramatic reduction in CPU times observed with just 8 processors on IBM SP/1 provides a unique opportunity to perform large-scale device simulations to study device characteristics in ultra small structures. These results again reinforce the engineering importance of the FIESTA code demonstrations and benchmarking. A 16 node machine represents a very practical configuration from an industrial perspective. The parallel performance improvements of more than an order-of-magnitude provide the essential enabling technology that supports real engineering applications.

To summarize the contributions of the FIESTA code and benchmarks, this work has broken new ground in both application of the GLS-FEM technology and in demonstrating near ideal parallel performance improvements based on the ENSA solver implementation. The application of FIESTA to submicron MOS and bipolar semiconductor devices has been demonstrated.

IV. Parallel Direct and Iterative Solvers with benchmarks for the DD Formulation

The drift-diffusion (DD) formulation is the most widely used engineering approach for semiconductor device analysis. In contrast to the previous section where advanced FEM technology was used, the section focuses on the parallelization of DD-based application codes with a special emphasis on the parallel linear solvers for sparse matrices. Both incomplete nested dissection and recursive spectral bisection algorithms were integrated into the solvers to perform domain decomposition. By using "plug-in" modules for the domain decomposition, the code can easily use the latest and most promising partitioning algorithms available. Thus, efforts could be concentrated on parallelizing the existing semiconductor device analysis applications and their linear solvers.

The PISCES code (release 2B, version 9009) was used as the test vehicle for the 2D parallelization. PISCES is an widely-used, industry-standard code with over 15 years of development history. The code encompasses many areas of interest in both physics and computational mathematics. Almost all aspects of this complex code have been parallelized including physical model evaluations, matrix formation and assembly, non-linear solution, and linear solution. A complete discussion of the creation of the parallel application, PISCES-MP, has been presented in [13][18][19]. To facilitate the parallelization of PISCES, the existing public domain linear solver was replaced with a parallel direct linear solver based on the work of Lucas [1]. The resulting parallel application has been ported to three parallel platforms: Intel, Thinking Machines, and IBM. The distributed-memory, message-passing parallelization methodology used made porting among the architectures straightforward. In fact, a major achievement of the PISCES-MP project has been to understand in detail, by means of consistent benchmarking, the aptitudes of the various parallel machines. Moreover, since PISCES-MP is a fully-functional application with broad acceptance in the IC technology community, these results carry user-side credibility on promoting and accelerating use of parallel machines. Recent benchmarks will be presented in a paper by Dr. Bruce Herndon.

The solution of 3D device analysis problems is generally believed to be beyond the practical capabilities of direct solvers. The STRIDE code has been developed as a test vehicle to explore parallel iterative solvers in the context of the DD formulation [2][3]. While all the code development was done on Intel computers, during the last few months of this research contract, core routines were ported to and benchmarked on the SP/1 as well. A broad range of parallel matrix solution methods and preconditioners that work well for the semiconductor problem have been demonstrated. Concurrent ILU(1) and ILUV preconditioners have been developed and demonstrated to be effective where concurrent ILU(0) is insufficient. Results indicate that ILUV is more promising for difficult problems and can achieve convergence with much less fill-in than for ILU(0). These results indicate that for ill-conditioned matrices, the condition number of the preconditioner can become a dominant factor in determining effectiveness. While the primary results in this study were demonstrated using the STRIDE code with tensor product grid, a prototype version of an object-oriented sparse matrix solver was also demonstrated. The single processor version achieved identical performance to the FORTRAN version used with STRIDE. In addition, benchmarks on an industrial matrix problem provided by IBM will be discussed in the final section of the report.

The STRIDE benchmarks on parallel iterative solvers for the DD formulation of a bipolar transistor with 4.9 million grid was demonstrated on the Caltech Delta machine. For the 512 node machine, a maximum sustained performance of 1.7 GigaFlops was achieved, leading to recognition in the Grand Challenges competition in 1992. Of equal and even greater importance is the fact that convergence of these huge 3D bipolar problems was achieved in an average of 20 minutes per bias point. This is indeed an encouraging result that suggests scaling to still larger problems will be very practical based on more powerful processors than are currently available. One rather surprising result was the fact that the growth of computational time with grid size was a very modest 1/3-power law. This is a result of the careful management of communication requirements and the fact that with larger problems, the processors have more work to keep them busy and hence are communicating less.

An important postscript to the benchmarking of both PISCES-MP and STRIDE codes is the fact that memory size and its full utilization are key aspects of achieving high performance parallel computing. Namely, having high-performance computational nodes with commensurate per-node memory is of prime importance to effectively realize parallel computing's potential. In terms of the benchmarks themselves for a fixed problem size and moving to more parallel processors, there comes a point of diminished returns where less work is given to each processor and the communications requirements (and their frequency) start to overcome benefits of having more processors. Hence, the results presented in this section, while indeed promising, cannot be considered definitive in showing scaling limits for parallel applications in semiconductor device analysis. Quite to the contrary, much of the limitations revealed here press the need for more robust grid generation and the more complete restructuring of the applications themselves. Finally, there are indeed opportunities to improve the domain decomposition process, both for static and dynamic grid conditions.

V. Integrated Geometry and Grid Services in Support of 3D TCAD

In order to support 3D TCAD applications using parallel computational capabilities, there are a number of infrastructure issues that need to be addressed in both specifying 3D structures and in gridding them for simulation. The full extent of the challenge involved in both these tasks was not apparent at the outset of this project. While the STRIDE examples with millions of grid in 3D had been achieved early on in the project, the task of generalizing these capabilities to nonplanar surfaces proved to be major effort of its own proportions. Moreover, in the early stages of the project, there was considerable optimism that a standardized semiconductor wafer representation (SWR) for 3D would progress at a rate that would interleave with requirements for this project. In fact, over the course of this project, the SWR efforts became stalled at both a committee level and in terms of actual prototyping. The results presented here have achieved significant advances in developing both a unified representation and supporting tools for both 2D and 3D gridding services in support of a TCAD framework, despite the missing infrastructure in terms of a full SWR. Nonetheless, it is important to note that such development efforts are beyond the scope and manpower of this project.

The specification of solid geometry services to support a fully 3D wafer representation were initially proposed (SWR, Version 1.5) and subsequently revised in the course of this project [20]. Based on this information model, a commercial solid modeler from Spatial Technology called ACIS was used to create a fully functional 3D geometry server that was used for two applications--as a 3D structure generator as well as a support utility for 3D gridding of nonplanar structures. The gridding aspects are discussed shortly. The purely geometric use of the server resulted in a new application code for creating 3D Virtual Integrated Processes (VIP3D) starting with IC mask layout information and a minimal set of process parameters such as layer thicknesses and shape factors (i.e. gate spacers, locally oxidized isolation shapes, etc.). The results of such a purely geometric representation of a 3D IC structure are illustrated Attachment 1 that shows a four transistor static memory (SRAM). It can be seen that complex features of the design are readily mimicked in the computational prototype. This SRAM design came from a commercial development project at Cypress Semiconductor. In fact, a simplified SRAM design was used as a vehicle to test a suite of TCAD tools to extract IC behavior at the circuit level based on layout information and simple technology parameterization [21]. Among the most mature and promising of the applications that can directly use results of VIP3D are: 1) interconnect analysis of complex 3D structures based on the multiple analysis technique used in FASTCAP [22] and 2) mixed-mode analysis of arbitrary technology cross-sections where 2D device analysis such as PISCES can be used in concert with the SPICE circuit analysis program [16] practical test conditions can be imposed on multi-device cross-sections as well. These demonstrations are part of the ongoing applications of the prototype VIP3D code. The industry response to these demonstrations has been highly enthusiastic. It is recommended that further attention and effort be applied in creating more complete infrastructure for 3D TCAD based on this prototype.

The gridding support for 3D TCAD is a major challenge. As indicated earlier, the specification of a viable information model for 3D has received only modest support through a SEMATECH Lithographer's Workbench (LWB) project. In fact, in a recent workshop sponsored by SEMATECH, a proposal for implementing an SWR Version 2.0 was turned down by the sponsoring companies. A major problem with both the Version 1.5 document and the Version 2.0 proposal is the over specification of nonessential features which make implementation difficult. In this project, two distinctly different implementations of 2D and 3D gridding capabilities have been created using a subset of the SWR, Version 1.5 specification. In both cases, a tree-based data structure (quad-tree for 2D and oct-tree for 3D) has been used. The key feature of the tree-based approach of interest for this project is the opportunity to use the tree as a logical structure for refinement as well as coarsen of the grid. In the context of implementing the gridders based on a client-server architecture, both codes have used geometry operations to handle various issues related to interfaces and surfaces. The following discussion summarizes each of the gridders used and developed in this work.

The FOREST gridder was developed in support of 2D process and device simulation research funded by the Semiconductor Research Corporation (SRC). As stated above, the quad-tree approach is used and a custom geometry server capability is directly integrated into the code. Both static and dynamic (moving) grids are handled by FOREST. While it is beyond the scope of this project to provide detailed documentation of the code and algorithms, several papers [23][24] that summarize the capabilities. Of special interest in the context of this project is the fact that problems suitable for parallel analysis with PISCES-MP can be generated either using FOREST or the 2D/3D gridder described next. Because FOREST was very specifically targeted to support 2D analysis, it is more aggressive in its algorithms and its use of grid to resolve fine features in IC cross-sections.

The CAMINO gridder has evolved out of the thesis work of Yang [25] with the specific intent to handle complex 3D structures. The gridding and refinement algorithms are based on quad- and oct-tree methodology. The development of a level-control function [23] was one key innovation needed to avoid the excessive refinement of grid in areas not of practical interest. The second major contribution has been the implementation of a delta-zone at surface regions in order to support refinement at non-planar surfaces and still maintain the tree data structure without excessive refinement [26]. In contrast to the built-in geometry services used in FOREST, the CAMINO program uses the same ACIS-based server described earlier as part of VIP 3D. This choice in implementation was targeted in testing the approach and algorithms and not for efficiency. In the course of industrial collaboration with IBM as discussed in the next section, the use of ACIS is now being reconsidered. In fact, jointly with IBM a different solid geometry modeler was interfaced to further test and benchmark CAMINO. As part of follow-on work in the SPRINT-CAD project, an efficient special purpose geometry server will be implemented in CAMINO.

In summary, several key utilities and framework services were developed as part of this project. While it was initially anticipated that standardized SWR services could be used in this project, this turned out not to be possible owing to factors outside the scope and resources of this project. However, the use of three prototype codes in support of this project have clearly demonstrated the viability of both the client-server model and the constituent geometry and grid information models. The VIP 3D code based on purely geometric specification of 3D device structures has been shown to be a very practical and powerful technique. In addition, the FOREST and CAMINO tools have created a solid foundation for semi-structured 2D and 3D gridding respectively using a tree-based algorithmic approach. Both these tools are capable of supporting hierarchical grid refinement and will be developed as part of the follow-on SPRINT-CAD project.

VI. Technology Transition including Industrial Applications

The application of TCAD to the design and manufacturing of ICs has demonstrated major benefits and cost savings that have been quantified by virtually all manufacturers in the industry. The use of 1D and 2D simulators has been the workhorse of the industry. Yet the need for more powerful 3D tools has emerged as a pressing one with the drive for higher frequency performance and scaling of technology to deep submicron structures. In the course of this project, both practical examples and industrial interactions have helped to drive the development and to quantify the output. Several of these points have been mentioned in the previous sections and are now discussed in greater detail.

In the period from Fall 1992 through Spring 1993, the ALADDIN-CAD project (AnaLysis of ADvanced Devices based on Industry-Networked TCAD) had been a consortia driven effort to apply and test early prototypes of parallel software for TCAD. The final report on that project [27] provided documentation of both the technical achievements and the industrial interactions that lead to follow-on collaborations are key parts of the work reported here. Specifically, the following companies and their application domains each contributed to the testing and benchmarking of results from this follow-on project:

- **Cypress Semiconductor**--SRAM development based on 3D geometry modeling and prototype components of VIP3D.
- **Hewlett-Packard**--High-speed device design and optimization with emphasis on optoelectronics and high-speed components for data communications--test examples for PISCES-MP came from these collaborations.
- **IBM**--Testing and applications of virtually all aspects of this project. The donation of a 16 node SP/1 to Stanford in 1993 for research collaboration allowed key benchmarks to be quickly developed and reported.
- **Intel**--Application of the 3D STRIDE code for parasitic analysis was an early example of code transfer to industry and its ongoing industrial use.
- **National Semiconductor**--High voltage semiconductor analysis and design has provided another example and benchmark for the PISCES-MP code.

The details of Cypress Semiconductor's application of 3D geometry modeling are proprietary. On the other hand, the results shown in Attachment 1 clearly indicate that practical examples can now be created and used effectively in industrial practice. In fact, Dr. Tony Alvarez, VP of Research and Development, has provided us with key feedback about both the benefits and an estimate of the quantitative benefits of this work. As mentioned in the previous section, while the VIP3D software is still very much in a prototype stage, it is clear that further development and application are of the highest potential. In fact, a unique strength of the approach used in this work is its applicability to a wide range of technologies (i.e. OEIC, MEMS, etc.) in addition to the primary demonstration in the arena of VLSI silicon.

The interactions with HP, Intel, and National Semiconductor each represent a sampling of diverse applications and the unique leverage possible with large scale device simulations. In the cases of both HP and National, the device structures each put major demands on analysis capabilities of PISCES-MP. The fact that the scaling of problems sizes beyond 100,000 grid level and the robust and efficient performance of the direct solver used are most encouraging. In fact, the direct solver technology has been sought after by vendors not only from the TCAD domain but also those involved in parallel applications to mechanical engineering applications.

The Intel applications of STRIDE on their parallel machine and in support of industrial 3D IC design considerations has been a major benefit of the ALADDIN-CAD project that has carried over into this work. Dr. Francisco Leon, Program Manager of Process and Device Modeling at Intel has given clear evidence of the ongoing impact of the work. Of special interest and importance is the collaboration related to 3D geometry-based modeling. In spite of the difficulties arising out of the delayed development of the SWR, Intel has been a key driving force in looking for viable technical approaches that meet the long-term needs for 3D TCAD. It is expected that follow-on efforts related to the SWR in the context of the SPRINT-CAD project will continue to be productive.

The interactions with IBM are of special importance. In some sense, the impact may show the greatest potential for long-term growth. Not only has the collaborative research been an excellent model for interaction but, additionally, that IBM has chosen to further leverage the research efforts through substantial equipment donations. In the case of this project, the donation by IBM of a 16 node SP/1 computer has made it possible to accelerate not only the development and benchmarking of parallel TCAD software, it has set the stage for other industrial and research applications of the technology. As noted in both sections III and IV of the report, excellent benchmark results for both FIESTA and PISCES-MP have been achieved on the SP/1 at Stanford as well as the SP/2 at the NASA/NAS facility. In fact, over both this project and the ALADDIN-CAD efforts, the availability of both a local machine at Stanford and more powerful scaled-up version at NASA/NAS has been a highly effective means to develop benchmark and further test the limits of research codes.

The breadth of the IBM interactions, stimulated through this project as well as ALADDIN-CAD are beyond the scope of the present report. However, a few more highlights will serve to illustrate their importance and potential impact:

- Collaborative efforts to apply both more advanced HD modeling and algorithms for parallelization of the FIELDAY program.

- Benchmarking of large sparse problems provided by IBM to test Stanford's iterative solvers based on parallel preconditioning.

- Collaboration in porting the CAMINO 3D gridding tool into the IBM environment, based on their internal solid geometry support utilities, and to test its application for industrial problems.

- Porting of IBM's FIELDAY program to Stanford for application in the ARPA-related project--both SPRINT-CAD and a new Computational Prototyping project.

- Assistance in adapting FIELDAY to support the mixed-mode capabilities for coupled device-circuit analysis in conjunction with a Stanford-modified version of UC Berkeley SPICE.

- Collaboration in providing Stanford a license to use the DAMOCLES Monte Carlo device analysis code in support of both the ARPA Computational Prototyping project and the NSF National Center for Computational Electronics (NCCE).

- Collaboration in producing visualization results for the 4.9 million bipolar transistor examples computed on the Intel Delta machine.

These collaborations with IBM were facilitated in large part through the participation of Dr. Ronald Knepper of IBM East Fishkill, an industrial visitor at Stanford's Center for Integrated Systems. The interactions have involved more than a half-dozen collaborators at five locations within IBM. As noted in the ALADDIN-CAD support letter from IBM, the fostering of such a relationship from the initial consortia efforts has been of major benefit to both the Stanford research efforts and to IBM as an industrial partner.

VII. Conclusion

The previous sections have summarized the results of a three-year research effort supported by

ARPA to demonstrate the parallelization of FEM Software for semiconductor device simulation. The results clearly demonstrate both the viability and computational efficiency of such parallel codes applied to the semiconductor domain. In addition, there is clear potential that the availability of such codes can have a long-term impact on the IC industry to achieve greater efficiency and effectiveness in the design of high performance hardware.

The specific achievements of this project can be quickly summarized as follows:

- development and parallelization of a GLS-FEM formulation of the HD transport equations using matrix-free GMRES;
- parallel benchmarks of the GLS-FEM code on Intel and IBM machines for 2D MOS and bipolar examples;
- parallelization of a mature 2D finite volume code PISCES based on a direct solver and its porting across Intel, TMC, and IBM machines;
- benchmarks of the PISCES-MP for 2D examples scaled with available memory to 128 processors and hundreds-of-thousands of grid, achieving 6.5 GFlops on an SP/2;
- implementation of parallel preconditioned iterative solvers using ILU(0), ILU(1), and ILUV nd complete application testing using the STRIDE code based on the DD formulation;
- parallel benchmarks using STRIDE for 3D bipolar examples with 4.9 million grid computed on the 520 processor Delta machine, achieving convergent solution in 20 minutes per bias point;
- development of quad- and oct-tree gridding utilities in support of parallel TCAD applications;
- development of a 3D geometry-based structure generator based on a Virtual Integrated Process (VIP 3D) representation;
- testing of several large-scale industrial examples for PISCES-MP including: CMOS gates, high-voltage and optoelectronic device structures;
- technology transfer of algorithms, code fragments and even full applications to industrial partners.

References

[1] R. F. Lucas, “Solving Planar Systems of Equations on Distributed-memory Multiprocessors,” Ph.D. Dissertation, Stanford University, 1987.

[2] K-C Wu, R.F. Lucas, Z. Wang, and R.W. Dutton, “New Approaches in a 3-D One -carrier Device Solver,” IEEE Trans. CAD, vol. 8, no. 5, pp. 528-537, 1989.

[3] K-C Wu, G. Chin, and R.W. Dutton, “A STRIDE Towards Practical 3-D Device Simulation--Numerical and Visualization Considerations,” IEEE Trans. on CAD, vol. 10, no. 9, pp. 1132-1140, 1991.

[4] H.D. Simon, “Partitioning of Unstructured Problems for Parallel Processing,” Comp.Sys. in Eng., vol. 2, pp. 135-148, 1991.

[5] M. Pinto, C. Rafferty, H. Yeager, and R.W. Dutton, “PISCES-IIB Supplementary Report,” Stanford University, 1985.

[6] R.W. Dutton and Z. Yu, TCAD--Computer Simulation of IC Processes and Devices, Kluwer Academic Publishing, 1993, pp. 373.

[7] E.M. Buturla, J. Johnson, S. Furkay, and P. Cottrell, “A New Three-dimensional Device Simulation Formulation,” NASECODE VI: Proceedings of the Sixth International Conference on the Numerical Analysis of Semiconductor Devices and Integrated Circuits, J.J.H. Miller, Ed.,

Boole Press Ltd., Dublin, 1989.
[8] F. Shakib, T.J.R. Hughes, and Z. Johan, “A New Finite Element Formulation for Computational Fluid Dynamics: X. The Compressible Euler and Navier-Stokes Equations,” Comp. Methods in Applied Mechanics and Eng., vol. 89, pp. 141-219, 1989.
[9] N.R. Aluru, A. Raefsky, P.M. Pinsky, K.H. Law, R.J.G. Goossens, and R.W. Dutton, “A Finite Element Formulation for the Hydrodynamic Semiconductor Device Equations,” Computer Methods in Applied Mechanics and Engineering, vol. 107, pp. 269-298, 1993.
[10] N.R. Aluru,K.H. Law, A. Raefsky, P.M. Pinsky, R.J.G. Goossens, and R.W. Dutton, “Numerical Solution of Two-carrier Hydrodynamic Semiconductor Device Equations Employing a Stabilized Finite Element Method,” to appear in Computer Methods in Applied Mechanics and Engineering.
[11] N.R. Aluru, K.H. Law, P.M. Pinsky, R.J.G. Goossens, and R.W. Dutton, “Space-Time Galerkin/Least Squares Finite Element Formulation for the Hydrodynamic Device Equations,” IEICE Trans. Electron., vol. E77-C, no. 2, pp. 227-235, 1994.
[12] N.R. Aluru, K.H. Law, and R.W. Dutton, “Simulation of the Hydrodynamic Model on Distributed Memory Parallel Computers,” in preparation.
[13] B.P. Herndon, N.R. Aluru, A. Raefsky, R.J.G. Goossens, K.H. Law, and R.W. Dutton, “A Methodology for Parallelizing PDE Solvers: Application to Semiconductor Device Simulation,” VIIth SIAM Conference on Parallel Processing for Scientific Computing, Feb. 1995, San Francisco, CA.
[14] K.H. Law, “Large Scale Engineering Computations on Distributed Memory Parallel Computers and Distributed Workstations,” NSF Workshop on Scientific Supercomputing, Visualization, and Animation in Geotechnical Earthquake Engineering and Eng. Seismology, Nov. 1994.
[15] N.R. Aluru, K.H. Law, P.M. Pinsky, and R.W. Dutton, “An Analysis of the Hydrodynamic Semiconductor Device Model--Boundary Conditions and Simulations,” submitted for publication.
[16] Z. Yu, D. Chen, L. So, and R.W. Dutton, “PISCES-2ET--Two Dimensional Device Simulation for Silicon and Heterostructures,” Stanford University, 1994.
[17] N.R. Aluru, K.H. Law, A. Raefsky, and R.W. Dutton, “FIESTA-HD: A Parallel Finite Element Program for Hydrodynamic Device Simulation,” submitted to parallel CFD'95, California Inst. of Technology, Pasadena, CA, June 26-28, 1995.
[18] B.P. Herndon, A. Raefsky, and R.J.G. Goossens, “PISCES MP--Adaptation of a Dusty Deck for Multiprocessing,” Proceedings of NASECODE-VII, May, 1992.
[19] B. Herndon, A. Raefsky, R. Goossens, and R.W. Dutton, “A Methodology for Parallelizing PDE Solvers: Applications to PISCES,” submitted to the Journal of Computer and Software Engineering, Special Issue on Parallel Computing.
[20] TCAD Framework Group, Semiconductor Wafer Representation Working Group, Semiconductor Wafer Representation Procedural Interface (PI), version 1.5, Technical Report, CAD Framework Initiative Inc., 1993.
[21] K. Wang, F. Rotella, T. Chen, D. Yang, A. Lee, Z. Yu, R.W. Knepper, J. Watt, and R.W. Dutton, “Layout-based Extraction of IC Electrical Behavior Models,” IEDM 1994 Proceedings, December 1994, San Francisco, CA, pp. 209-212.
[22] K. Nabors and J. White, “A Fast Multiple Algorithm for Capacitance Extraction of Complex 3-D Geometries,” Proc. Custom Int. Circuits Conference, San Diego, CA pp. 21.7.1-21.7.4.
[23] Z. Sahul, E. McKenna, R.W. Dutton, “Grid Evolution for Oxidation Simulation Using a

Quadtree Based Grid Generator," NUPAD V Conference, June 5-6, 1994, Honolulu, HI, pp. 155-158.
[24] Z. Sahul, E. McKenna, R.W. Dutton, "Grid Techniques for Multi-Layer Device and Process Simulation," TECHCON'93, Sept. 28-30, 1993, Atlanta, GA.
[25] D. Yang, "Mesh Generation and Information Model for Device Simulation," Ph.D. Dissertation, Stanford University, 1994.
[26] D. Yang, K. Law, and R.W. Dutton, "An Automated Mesh Refinement Scheme Based on Level-Control Function," Proceedings of NUPAD IV, May 31-June 1, 1992.
[27] R.W. Dutton, "Analysis of Advanced Devices Using Industry-Networked Technology CAD (ALADDIN-CAD)," Final Report, California Competitive Technology Program (CompTech Grant# C90-072).