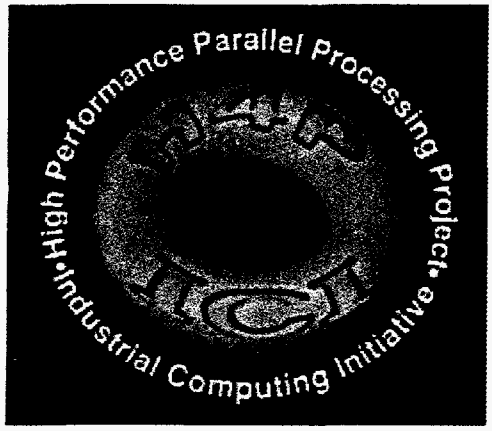

High Performance Parallel Processing Project

Industrial Computing Initiative

Progress Reports for Fiscal Year 1995

February 9, 1996

UCRL-ID-123246

Lawrence
Livermore
National
Laboratory

This report has been reproduced
directly from the best available copy.

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information
P.O. Box 62, Oak Ridge, TN 37831
Prices available from (615) 576-8401, FTS 626-8401

Available to the public from the
National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Rd.,
Springfield, VA 22161

Work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract W-7405-ENG-48.

High Performance Parallel Processing Project
Industrial Computing Initiative
Progress Reports for Fiscal Year 1995

The High Performance Parallel Processing Project (H4P)

Alice Koniges, Project Leader
Lawrence Livermore National Laboratory

Overview and History

The High Performance Parallel Processing Project (H4P) is a package of 11 individual CRADAs (Cooperative Research and Development Agreements), plus hardware (the Cray Research CRAY T3D sited at Lawrence Livermore National Laboratory). This innovative project established a three-year multi-party collaboration that is significantly accelerating the availability of commercial massively parallel processing (MPP) computing software technology to U.S. government, academic, and industrial end-users. It has been historically known as the "SuperCRADA," since it is a piece of a $40M set of computing-related agreements with Lawrence Livermore National Laboratory (LLNL), Los Alamos National Laboratory (LANL), Cray Research Inc. (CRI), and other industrial partners announced in 1994 by the Secretary of Energy, Hazel O'Leary. This project brought the first 128 processing elements (PEs) of the T3D to LLNL. (There is a similar set of individual CRADAs and a T3D at LANL.) The second half of the T3D (another 128 PEs, bringing the total to 256) came to LLNL as part of the Computations Department Director's Initiative and UC funding. Now these projects share the full 256-PE machine.

There are a total of nine LLNL principal investigators (PIs) from various directorates associated with the project, and additional FTEs for graphics, machine management, and project management. Each of the PIs has a matching FTE from an industrial partner. Further, two of these projects have an additional CRADA with CRI, and an associated CRI partner.

The purpose of the CRADAs is to take laboratory software technology, some of it with roots in weapons programs, advance the technology, and make it available to U.S. industry with joint licensing agreements. In return, the project offers the core Laboratory programs a means to enhance their code development activities with leading edge high-performance computing capability and a test suite of unclassified industrial applications to benchmark their computations. It is important to note that each of the projects is not just moving code to the private sector, but developing true MPP (massively parallel processor) applications which allow for realistic geometries and three-dimensional simulations. In general, each of the projects is part of a major LLNL code system effort, and the CRADA money provides an additional FTE for parallel code development.

As a result of this project, LLNL is designated by CRI as one of five Parallel Applications Technology Program (PATP) sites. This gives LLNL additional FTEs from CRI to help with the project.

The financial information for the project is summarized in Table 1. The projects and their industrial partners are summarized in Table 2.

An Environment for Supercomputing Applications

The CRAY T3D is an MPP supercomputer with a peak speed of 150 Mflops per node and a communication bandwidth between nodes of 300 Mbytes/s in each direction. The node architecture consists of two EV4 DEC alpha processing chips clocked at 150 Mhz, with 64 Mbytes of memory. The complexity of programming on such an environment with current technology is akin to, but even more involved than, the transition to vector supercomputing made in the previous decades. However, the payoff in terms of large memory (e.g., three-dimensional) modeling and real-time clock speeds which are expected to reach teraflop performance in the

Table 1. Manpower H4P funding in K dollars.

	FY95	FY96	FY97
H4P—CRADA Research PIs	$1,381	$1,655	$1,720
H4P—T3D—infrastructure	498	350	381
H4P total manpower	1,879	2,005	2,101

Table 2. H4P Projects.

Principal Investigator	Project	Technology	Partner(s)	MPP/Code info.
Tomas Diaz de la Rubia, ext. 2-6714* (Chemistry and Material Science)	Shallow-Junction Device Modeling	Microelectronics/ device modeling characteristics at microscopic level	**AT&T**: George Gilmer **CRI**: Kevin Lind	Molecular dynamics and Monte Carlo
Bruce Griffing 2-4498 (Computations)	T3D Gang Scheduler	Software for MPP systems	**CRI:** Steve Luzmoor	GUI design, roll-in/roll-out for MPP
Milo Dorr 3-2423 (Computations)	Nuclear Imaging	Petroleum exploration– nuclear well logging	**Halliburton:** Larry Jacobsen	Codes: Ardra, AMTRAN
Steve Ashby 3-2462 (Computations)	Environmental Remediation	3D subsurface flow in heterogeneous materials	**IT Corp.**: **CRI:** Kevin Lind	Code: ParFlow MPP algorithms for conjugate gradients
Richard Couch 2-1655 (Defense and Nuclear Technologies)	Finite Element Fluids and Structures	Metal forming, manufacturing	**Alcoa**: Don Ziegler **CRI**: (not a CRADA) TBA	Codes: ALE3d and ALEC Domain decompo-sition for MPP
Cliff Shang 2-6174 (Engineering)	Computational Electromagnetics	3D dynamic E&M field solver for laser, radar and antenna design for high clock rate microelectronics	**Hughes Air Craft**: E. Illoken	Parallel mesh generation. Parallel PDE solvers
Jerry Goudreau 2-8671 (Engineering)	Fluid Dynamics, Acoustics, Structures	Acoustical studies of submarines	**Arete Corp.**: F. L. Fernandez	PING compo-nent of DYNA3D. Parallel I/O, MPI implementation
Doug Rotman 2-7746 (Environmental Programs)	Global Atmospheric Chemistry	Effect of new aircraft on environment	**Boeing**: Steve Baughcum	Code: IMPACT stiff, coupled ODEs
Christian Mailhiot 2-5873 (Physics and Space Technologies)	Advanced Materials	Materials design. Semiconductors, metals, surfaces, thin films	**Xerox**: J. Northrup C. Van de Walle	Code: PCGPP conjugate gradients

*LLNL telephones are area code (510)42 followed by extension.

upcoming years makes this type of architecture an ideal test-bed for next-generation supercomputing.

The transition of real applications codes to the MPP environment often requires a rethinking of how to map a physics or engineering problem onto the computer. It is not generally true that the best vector supercomputer algorithm will translate into the best MPP algorithm. This dichotomy occurs at all levels, from posing the equations to the numerical implementation. Additionally, the institutional aspects of running such a computer as a multi-user facility requires new software for code management, scheduling, I/O, graphics, and other areas. The LLNL H4P project is making advances in all of these areas.

Summary of First Year Accomplishments

In the first year of funding, each of the nine projects attained milestones tailored for the particular project. These milestones include developing and testing new algorithms for parallel computing, porting existing MPP codes to the T3D, incorporating new and multi-dimensional physics. By the end of the year, all of the projects were computing on the new T3D platform with at least a portion of these large code systems.

Some highlights of the first year include achievements in speed, algorithm development, and large memory implementations. For example, through the use of single node optimization techniques (see also, "Parallelizing Code for Real Applications on the T3D," A. E. Koniges and K. R. Lind, *Computers in Physics* **9**, 399, 1995) one of our applications code kernels was clocked at 87 Mflops on a 1024-processor T3D. Another project worked on conjugate gradient algorithms, achieving a 100 times speed-up over conventional techniques. A third project pioneered the use of heterogeneous computing by running the primary code application on the T3D while remotely driving the visualization on an SGI workstation.

As a result of this and similar projects, the software environment on the T3D has greatly improved to include better debugging tools and standard parallel programming environments such as MPI (Message Passing Interface). One of our own projects, the dynamic gang scheduler, which will allow roll-in-roll-out of codes to facilitate time-sharing on the MPP platform, is approaching the beta-test phase. As the projects mature, they continue to press the development of high-performance computing in areas such as parallel I/O and heterogeneous environments for computer graphics and graphical user interfaces. The project has paved the way for recognition of the MPP platform as a viable and necessary machine of the future by showing its utility in real-world computations.

In this document, we collect reports from the nine projects and give summaries of their first year of funding. We explain the general approach to the problem, role of the industrial partner, progress, milestones and future plans.

For more information about the H4P, contact Alice Koniges—`koniges@llnl.gov`

Lawrence Livermore National Laboratory
P.O. Box 808, L-630
Livermore, California 94550
(510) 423-7890

The *ParFlow* Project: Enabling Detailed Simulations of Subsurface Flow and Chemical Migration

Steven F. Ashby
Lawrence Livermore National Laboratory

Project Objectives

The goal of the ParFlow project is to apply high performance computing techniques to the three-dimensional modeling of fluid flow and chemical transport through heterogeneous porous media to enable more realistic simulations of subsurface contaminant migration. These simulations will be used to improve the design, analysis, and management of engineered remediation procedures.

General Approach

The ParFlow simulator uses the power of massively parallel processing and the efficiency of state-of-the-art numerical methods to enable detailed simulations of multiphase fluid flow and multicomponent chemical transport. These simulations, which use up to 64M spatial zones, allow one to resolve the fine-scale subsurface heterogeneities that can dramatically impact plume migration, and consequently, the design of optimal remediation strategies. We employ geostatistical techniques to reproduce the effects of these heterogeneities, and our simulation technology supports probabilistic risk assessment (in which multiple realizations of the subsurface are used to quantify the inherent uncertainty). We also advocate a scalable approach to building the conceptual model of the subsurface in a way that is independent of the computational grid on which the problem will be solved. Our code runs on a variety of computers, ranging from a single PC or workstation to massively parallel machines like the CRAY T3D.

Role of the Industrial Partners

We are working with colleagues at International Technology Corporation (IT) to develop and apply the ParFlow simulator to real sites, especially those of commercial interest to IT. These colleagues are responsible for developing the subsurface conceptual model, determining appropriate geostatistical parameters, characterizing the contaminant plume, and specifying other simulation parameters. In addition, they help us to establish priorities with respect to new code capabilities, and they provide feedback on the simulator's performance.

Colleagues at Cray Research, Inc. (Cray), are working with us to optimize the performance of the ParFlow simulator on various Cray computers, particularly the CRAY T3D. In addition, they are providing valuable feedback on the usability of the code and the adequacy of our documentation.

Progress for FY95

Capabilities

We developed and implemented a multigrid preconditioned conjugate gradient algorithm for solving the discretized pressure equation. This algorithm uses operator-induced prolongation and restriction, and an algebraic definition of the coarse grid operators, to enable the fast solution of extremely large problems. For example, we are able to solve a problem with more than 8M spatial zones in under thirteen seconds on the 256-processor CRAY T3D. Both the algorithm and its implementation are scalable. A paper describing the

algorithm and its performance has been submitted to a special issue of *Nuclear Science & Engineering.*

We added a simple two-phase capability to the simulator via the implementation of a Godunov routine for solving the saturation equations. Since this capability was added to enable simulation of a moving water table, we were able to ignore capillary pressure in this first version of the code. (Eventually, we will enable true multiphase simulations.) This explicit routine works well in that it accurately tracks the water table, but at the price of a small time step. We are currently investigating various "tricks" that might allow us to increase the time step without sacrificing accuracy.

We began adding those capabilities needed to do realistic plume migration studies at the Ajax (pseudonym) site in Northern California. In particular, we will need to rework the way we handle injection/extraction wells and we will need to develop a conditional simulation capability. This work will be completed by the end of the first quarter of FY96.

Simulations

We are working with colleagues in the Environmental Protection Department to apply the ParFlow simulator to the Lawrence Livermore National Laboratory (LLNL) site. To date, we have run detailed simulations using as many as 8M spatial zones. Several visualizations have been produced, and we are currently planning a new video.

In collaboration with Professor Graham Fogg (University of California at Davis), we ran a proof-of-concept flow calculation using more than 50M spatial zones. Professor Fogg wishes to investigate an indicator simulation technique that requires this resolution, and according to him, ParFlow is the only code capable of the calculation. We are now preparing to run a series of large runs that he needs.

Jeffrey Butera (North Carolina State University) ran a series of validation studies comparing ParFlow to MODFLOW for a variety of simple problems, as well as for the LLNL site.

We ran some preliminary plume migration studies for the Ajax site as part of our work with IT Corporation. Our results impressed several IT engineers, who remarked that they were able to see certain flow phenomena that previous modeling efforts had missed. (This was enabled by our ability to resolve fine-scale heterogeneities.) As a result of these simulations, we are now refining the subsurface conceptual model and boundary conditions.

Performance

We conducted a thorough investigation of the code's parallel performance and found it to be exceptional. In particular, we demonstrated near perfect scalability up to 256 processors (at about 80% scaled efficiency) of our multigrid solver. Other key routines also scaled well, with the exception of the turning bands algorithm. (We are currently implementing an alternative to turning bands which has better parallel performance.) These results are being written up and will be submitted to a refereed journal.

We replaced most of our critical PVM calls with equivalent SHMEM calls on the CRAY T3D, and boosted scaled speedup by 10%.

Cray analyzed our Godunov advection routine (a key computational kernel) and doubled its execution speed.

Plans for FY96 and FY97

We will devise and implement a parallel version of the sequential Gaussian simulation technique that is conditioned to data, especially well data. This will enable us to run realistic plume migration studies for the Ajax site. This work will be completed by the first quarter of FY96.

We will modify our wells capability to enable more realistic simulations. For example, we will allow for both constant head and flux wells, and we will allow for wells that are screened over several zones. This work will be completed by the first quarter of FY96.

Our simple two-phase capability will be fully tested and simulations of a moving water table (for the LLNL site) will be carried out. We will begin work on adding capillary pressure to the modeling equations, thereby enabling true multiphase simulations. We also will consider implicit and semi-implicit alternatives to our explicit Godunov approach.

We will add simple reactive chemistry to the advective Godunov routine, thereby allowing us to use the ParFlow simulator for a larger number of interesting plume migration studies. We also will consider incorporation of a dispersion tensor.

We will update our conceptual model of the LLNL site (to include the current pumping well configuration), run detailed simulations of flow and transport, and compare our results to those obtained by the CFEST code currently being used.

We will work with IT to run detailed simulations of the Ajax site. We plan to make a joint presentation to the management of Ajax in the hopes that they will contract with IT to conduct a proper study.

We will send ParFlow to colleagues at both CRI and IT for alpha user testing. We expect to receive both extremely valuable feedback as to the usability of the code and suggestions for improving the simulator's efficiency and capabilities.

Cray will help us optimize the code for the T3D, T3E, and J90 supercomputers. Cray also will help us with I/O and visualization issues.

We will develop a prototype graphical user interface (GUI) for controlling the simulation process, from problem setup to visualization of the results. This prototype will be given to an independent software vendor for further development as a prelude to commercialization of ParFlow.

Publications (FY95)

S. F. Ashby, W. J. Bosl, R. D. Falgout, S. G. Smith, A. F. B. Tompson, and T. J. Williams, "A numerical simulation of groundwater flow and contaminant transport on the CRAY T3D and C90 supercomputers," in *Proc. 34th Cray User Group Conference, Cray User Group, Inc.*, 1994, pp. 275–282. Conference held in Tours, France, October 10–14, 1994. Also available as LLNL technical report UCRL-JC-118635.

S. F. Ashby and R. D. Falgout, *A parallel multigrid preconditioned conjugate gradient algorithm for groundwater flow simulations*, Tech. Report UCRL-JC-122359, Lawrence Livermore National Laboratory, October 1995. Submitted to *Nuclear Science & Engineering*.

S. F. Ashby, R. D. Falgout, S. G. Smith, and T. W. Fogwell, "Multigrid preconditioned conjugate gradients for the numerical simulation of groundwater flow on the CRAY T3D," in *Proc. ANS International Conference on Mathematics and Computations, Reactor Physics, and Environmental Analyses*, vol. 1, 1995, pp. 405–413. Held in Portland, OR, April 30–May 4, 1995. Refereed proceedings.

S. F. Ashby, R. D. Falgout, S. G. Smith, and A. F. B. Tompson, "The parallel performance of a groundwater flow code on the CRAY T3D," in *Proc. Seventh SIAM Conference on Parallel Processing for Scientific Computing*, Society for Industrial and Applied Mathematics, 1995, pp. 131–136. Conference held in San Francisco, February 15–17, 1995. Also available as LLNL technical report UCRL-JC-118604.

S. F. Ashby, R. D. Falgout, and A. F. B. Tompson, *A scalable approach to modeling groundwater flow on massively parallel computers*, Tech. Report UCRL-JC-122591, Lawrence Livermore National Laboratory, October 1995. To appear in the proceedings of the U.S. EPA Workshop on Next Generation Environmental Models Computational Methods, August 7–9, 1995, Bay City, MI.

Global Atmospheric Chemistry Models

Doug Rotman
Lawrence Livermore National Laboratory

Steve Baughcum
The Boeing Company

Project Objectives

The objective of this project is to improve the understanding of the atmospheric chemical and climatic impacts, particularly on ozone in the troposphere and stratosphere, that may have occurred or will occur from current and future projected fleets of supersonic aircraft. This effort involves a close collaboration with scientists at the Lawrence Livermore National Laboratory (LLNL), The Boeing Company, and Cray Research, Inc. The primary tool to be used in these atmospheric chemistry transport studies will be the LLNL IMPACT three-dimensional chemical-transport model of the troposphere and stratosphere. A significant fraction of this project will be aimed at further development, testing, and validation of this model on the CRAY T3D to meet the needs for aircraft assessment plus use of the model in studies to evaluate aircraft effects on ozone and climate.

General Approach

The Lawrence Livermore National Laboratory has had a long history in the assessment of atmospheric impacts of supersonic aircraft. Much of this work in the past has been accomplished through the use of a two-dimensional model, a so-called zonal averaged model. While adequate for initial studies, the amount and level of uncertainties related to the zonal averaged fields required that complete global three-dimensional models be used for analysis and assessments. However, these three-dimensional models require large amounts of computer time, memory, and storage. Current vector supercomputers were not adequate for the level of computing needed for aircraft assessments. A number of years ago, LLNL began the development of a three-dimensional chemical transport model designed specifically for parallel machines, but capable of running on vector supercomputers and workstations. Under this project, that model, named IMPACT, is being advanced and implemented onto the CRAY T3D to carry out simulations that can further our understanding of the atmospheric impacts of supersonic aircraft.

The impact of aircraft is most clearly seen in the ozone depletion caused by the emission of oxides of nitrogen, i.e., NO_x emissions. The added emission of nitrogen oxides furthers the catalytic cycle of nitrogen in the stratosphere which acts to destroy ozone. Key to this depletion is the altitude of the aircraft flight patterns and hence, its emission. The proposed supersonic aircraft cruise at an altitude of about 18–22 km, nearly at the band of maximum ozone levels. It is important to understand this ozone depletion in terms of a model that includes both the troposphere and stratosphere since the emissions are just above the dividing line of these two distinct and different regions of the atmosphere. The exchange of mass between these regions of the atmosphere is highly unknown and important to the analysis of emission deposition and chemical effects. Much of our scientific work in this project is aimed at understanding this mass exchange, the chemical processes caused by aircraft emissions that affect the ozone distribution, and trying to assess any climatic effects from aircraft.

Our general approach is to make use of the computing power of the CRAY T3D by implementing our parallelized three-dimensional chemical transport model and adding the needed physics to further our understanding and analysis of the impacts of supersonic aircraft on the atmosphere.

Role of the Industrial Partner

The Boeing Company has long been in the business of aircraft design and building. To further their already leading role in the aircraft business, they are beginning to evaluate the feasibility of supersonic aircraft capable of carrying hundreds of people at Mach 2 to 2.4. As part of their effort, Boeing has worked with engine designers and marketing people to detail the possible engine designs, performance, and emissions, along with possible routes and markets so that such aircraft will make good business sense. This information is imperative to the analysis of environmental consequences of the aircraft.

To accurately analyze the effects of emissions, one obviously must know in great detail exactly what emissions are coming out of the aircraft engine at cruise altitude (where they spend most of their time). Boeing is providing this information for the model runs. In addition, marketing analysis is showing Boeing what routes and times are the best for usage of these types of aircraft. Boeing is also sharing this information. In short, they are providing the planned supersonic aircraft route scenarios and emission data for those scenarios, and providing the basic emission data for the model runs of IMPACT.

In addition, the primary contact at Boeing, Steve Baughcum, is a top atmospheric chemist, and hence, he is also providing input and suggestions as to the chemical parameters used on the chemistry modules of IMPACT. These parameters include the chemical mechanism itself, reaction rates, and photolysis coefficients. Steve will also play an important role in the analysis of the model output. Steve has been involved with NASA and their assessment of aircraft impacts for quite some time and brings a large level of experience and knowledge to the analysis.

Progress FY95

Our project was started in May 1995. Given our history and experience on parallel computers, the initial implementation of the IMPACT model onto the CRAY T3D was accomplished very quickly. This model incorporates the following primary physics operators:

- The solution to the set of ordinary differential equations representing the chemistry operator is solved using LSODE, a solution technique developed by Alan Hindmarsh at LLNL to solve a set of stiff ODEs.
- Photolysis coefficients are obtained through a two-stream temperature dependent delta Eddington scheme.
- Advection is accomplished though a van Leer scheme.

Using the SHMEM libraries, we have determined the scaling and parallel performance to be excellent. Table 1 lists the performance of the code on the CRAY T3D as compared to the CRAY C90 for a test case of N_2O chemistry for 50 time steps.

The scalability of the machine also appears to be very good. Moving from using 4 processors of the T3D to 64 processors, we were able to get 86% of that possible increase in speed.

The transport capability of IMPACT was immediately put to test using a tracer test problem, representative of a longer lived species in the stratosphere with a loss in the troposphere. A yearlong simulation was carried out using an emission scenario provided by Boeing. Results of the yearlong run were analyzed and while the results were very good, there appeared to be a relatively large amount of vertical diffusion in the advection scheme based on similar studies using other models. Hence, we are currently implementing a new advection scheme in all directions that makes use of the PPM (Piecewise Parabolic Method) in all directions (in addition, there are options for the van Leer as well). In addition, this scheme gives us the capability to use a time step 5 times larger through a semi-Lagrangian approach in the zonal direction near the poles. This region is a troublesome location because of high wind velocities and small grids, hence providing strict and small allowed time steps. The

Table 1. Performance of IMPACT on the CRAY T3D compared to CRAY C90.

Machine	No. of processors	Time (secs)	Cray Equiv.
CRAY C90	1	58	1.0
CRAY T3D	4	63	0.9
CRAY T3D	8	34	1.7
CRAY T3D	16	17	3.4
CRAY T3D	32	9	6.4
CRAY T3D	64	4.6	12.6

semi-Lagrangian gets around this by providing a method that is not Courant limited. The algorithm is implemented and running on the C90, and will be parallelized soon. The original soot run was used as a sample calculation for a National Energy Research Supercomputer Center movie shown at Supercomputing '95.

In addition, other needed physics processes were included in the model; for example, wet convection, the dry deposition of chemical species on the Earth's surface and, as we increase the numbers of species in the chemical mechanism, ground emissions (and altitude emissions of NO_x from lightning) are done or are currently being added.

All deliverables and milestones for the first year have been accomplished, except one. The missing milestone is the linkage of the chemical model to the general circulation model (GCM) output and linking to the general circulation model for concurrent and interactive running. The first part of this will be done very soon. Under different funding, we have already linked the model to other GCM output; hence, the addition of another GCM-based output should move smoothly. However, the interactive linking to a GCM may be pushed back for some time until this new advection scheme is implemented and working in parallel and complete physics and chemistry is implemented.

Plans for FY96–97

Short-term work will include the parallelization and completion of the new advection scheme and the implementation of a new chemistry solution technique. Even with the power of the T3D, the LSODE algorithm has proved to be too slow for realistic calculations. Hence, we are already beginning the implementation of a new scheme, equally accurate to LSODE, but up to 100 times faster (as measured on the CRAY C90). Wet scavenging and vertical diffusion algorithms should be completed within the next two months, thus providing a model with complete enough physics to perform stratospheric calculations. The main objective for this next year will then be to perform calculations establishing the background atmosphere and then adding the aircraft emissions to understand the chemical perturbations caused by the emissions. Near the end of the second year and throughout the third year, interactive climate runs will be carried out and further sensitivity analysis from aircraft emissions will be accomplished.

Computer Simulation of Nuclear Well Logging Devices

James Ferguson, Milo R. Dorr, Peter Brown, John Rogers
Lawrence Livermore National Laboratory

Project Objectives

The goal of this project is to develop a simulator for use in the design of nuclear logging tools. The simulator is being designed for execution on distributed memory platforms ranging from workstation clusters to massively parallel computers. The simulator will use deterministic algorithms similar to those used in some Lawrence Livermore National Laboratory (LLNL) Defense Programs applications to provide a complement to the Monte Carlo approach used at Halliburton Corp. and other oil and gas service companies. The experimental validation of the numerical methods and parallelization strategies used in the creation of this simulator will be of substantial dual benefit to Defense Programs and the Accelerated Strategic Computing Initiative (ASCI) Program.

General Approach

In this project, deterministic methods are being applied to solve the neutron transport equation. Specifically, a finite element method on a three-dimensional spatial grid is used to discretize the spatial dependence of the neutron flux. A discrete ordinate (S_N) algorithm is employed to model the directional variation of the flux, together with a standard multigroup treatment of the energy dependence. The resulting system is solved using an iterative method that alternates the evaluation of scattering sources with the inversion of the first-order differential terms via "transport sweeps." A parallel implementation strategy is pursued that maximizes the exploitation of concurrency with respect to all variables, and portability across distributed memory platforms.

Role of Industrial Partner

LLNL has principal responsibility for algorithm and code development. Halliburton Energy Services is responsible for providing tool specifications, data, and validation of the simulator with other models and experimental measurements.

Progress FY95

Port of Research Code to the CRAY T3D

The project was begun by porting the research code, **`Ardra`**, to the CRAY T3D. To insulate the code from vendor-specific implementations of message passing, **`Ardra`** employs an application-specific communication layer. This approach had previously enabled the code to be ported to a number of distributed memory platforms, including Sparcstation clusters, the Intel Paragon, the nCUBE/2, and the Meiko CS-2. The initial T3D port therefore involved implementing this layer atop the Cray Research Inc. (CRI) PVM (Parallel Virtual Machine) and SHMEM libraries. This work was performed by Eric Salo at the National Energy Research Supercomputer Center.

The communication requirements of the algorithm implemented in **`Ardra`** involve both point-to-point and collective operations. Initially, both types were implemented in PVM by Salo. The overhead of collective operations implemented in PVM became excessive when more than 16 PEs were used, so these operations were re-implemented using SHMEM functions. The performance of the point-to-point operations implemented in PVM were comparable to the use of SHMEM gets and puts, so the PVM version was retained.

Initial timings of the T3D **`Ardra`** port revealed that communication costs were very small, and that more attention should be paid to the single node performance. Salo modified the computationally intensive parts of the code to make it easier for the C compiler to optimize them. This work resulted in single-node speeds as high as 30 Mflop/s, with 15–20 Mflop/s being more typical. After convincing himself that no further performance gains would likely be achieved by this route, Eric rewrote the main kernel of the code in Alpha assembly language,

and made other modifications to utilize the cache more efficiently. This resulted in single-node speeds as high as 87 Mflop/s. This high level of performance came at the cost of wasting a fairly large amount of memory in order to store data in a cache-friendly manner, which involved some replication.

Scalability Tests

Ardra allows full phase space decomposition, i.e., it is simultaneously parallelized with respect to neutron position, direction and energy. Using a suite of problems designed to test scalability with respect to each of these variables, the T3D **Ardra** port was tested on the 128-PE system at LLNL and a 1024-PE system at Cray Research Inc.'s Eagan facility. The tests demonstrated very good scaled speedups, indicating that no further work was required on the communication layer. Using Salo's Alpha assembly version of the code, a maximum performance rate of 74 Gflop/s was achieved on the 1024-PE machine. The results of these tests were documented in a paper[1] presented at a conference sponsored by the American Nuclear Society.

Addition of Anisotropic Scattering

To model the anisotropic scattering that occurs in subsurface materials such as limestone, **Ardra** was generalized to allow the specification of arbitrary-order scattering operators. The approach taken was the standard one of expanding the scattering kernel and angular flux in Legendre and spherical harmonics respectively, then applying an addition formula to obtain a scattering representation in terms of spherical harmonic moments of the angular flux and differential cross sections indexed by the scattering order. The need to save the higher-order flux moments resulted in significantly increased memory requirements, so further changes were made to make the code as memory-efficient as possible.

Two Treatments of Point Sources

Discrete ordinates algorithms such as are used in **Ardra** are well known to have difficulty with isolated sources when scattering ratios are small. A standard remedy is to implement a "first-scatter" source, in which an analytic solution of the transport equation with scattering omitted is used to convert a point source, such as exists in models of porosity tools, to a distributed source. This approach was implemented in **Ardra**, but some oscillations in the scalar flux were still observed. A second method was therefore investigated that generalizes the "fictitious source" method of Miller and Reed.[2] The new harmonic projection algorithm[3] allows spherical harmonic (a.k.a., P_N) solutions to be obtained from a modified discrete ordinates calculation. Used in conjunction with the first-scatter source, this approach has eliminated the ray effects problem. This benefit comes at a cost of the additional memory required to store the additional basis moments needed to obtain a nonsingular transformation between the spherical harmonic and discrete ordinate bases.

Addition of Pure Downscatter Capability

Because **Ardra** was designed to be a general purpose transport code, the initial implementation of scattering assumed an arbitrary coupling of energy groups. The resulting fully-coupled multigroup system was solved iteratively. For the nuclear well logging problems, this approach is inefficient, since it does not permit exploitation of the fact that the scattering is purely down-scattering. The explicit assumption of down-scattering results in a block lower triangular multigroup system that can be solved sequentially in a group-by-group manner in the direction of decreasing energies. An option was added to **Ardra** to allow the multigroup system to be solved by block forward elimination, while maintaining scalability with respect to the remaining phase space variables.

Acquisition and Integration of a Cross-Section Data Base

A database of cross sections was obtained from the Radiation Shielding Information Center (RSIC) at Oak Ridge. The library contains microscopic cross sections for all atomic species comprising the materials of interest for well logging problems, and provides resolution of near-thermal energies superior to the cross-section data available locally. A conversion utility was created to convert the RSIC format to the Silo format used in B Division. **Ardra** was then modified to read Silo-formatted files. The acquisition of this library from RSIC simplifies the transfer of this data to Halliburton, since they will be able to obtain the same data directly from RSIC for a nominal fee.

Comparison with Monte Carlo Calculation of a Benchmark Problem

As an initial test of the simulator, the Halliburton and LLNL investigators agreed upon a benchmark problem for which results obtained by two Monte Carlo codes, MCNP and McDNL, are reported in Ref. 4. The benchmark problem specifies a compensated dual-detector, porosity tool. For a limestone formation with 20% porosity, **Ardra** obtained a near/far counting ratio (i.e., the rate at which neutrons are detected at the near detector divided by the rate arriving at the far detector) of 3.93, versus 3.47 for MCNP and 3.61 for McDNL. This calculation employed P_1 scattering, 47 energy groups, 48 neutron directions and an $80 \times 80 \times 220$ spatial grid. Further tests are necessary to determine the reason for the slightly higher ratio obtained by **Ardra**, but nevertheless this is a promising first result.

Plans for FY96 and FY97

Continuation of Ardra Development

The development of **Ardra** will continue, with emphasis on getting solid results for the test problem discussed above and for related benchmark problems provided to us by Halliburton. The follow-on benchmarks will involve problems with epithermal detectors, which require better resolution of the detector features, and will also involve engineering representations of the tools, as opposed to the idealizations used in the first benchmark problems.

At the end of FY96 we should be doing tests on full-scale engineering models of tools placed in real logging environments. At that point we can begin carrying out calculations to compare with experimental data from the Halliburton test pits and from field data, rather than computational benchmarks. FY97 will be occupied with improving and validating the **Ardra** results for these real-world applications, and with providing the input-output tools needed to make the code user friendly.

Algorithm Development

We have been examining extensions and generalizations of the algorithms in **Ardra** that have the potential to improve the performance of **Ardra** on the class of problems in the FY95 studies, and also to carry over into other classes of problems (the current benchmark represents one of about half a dozen types of tools currently used in the logging industry). Some of these changes include:

- Block zoning, permitting use of fine zoning in regions with lots of detail, and coarse zoning in relatively homogeneous regions.
- Higher-order elements: quadratic instead of linear.
- Incorporation of density gradients and material boundaries directly in the integrals of the finite elements. Our experience indicates that this feature can reduce errors by a factor of 2 for a given zoning resolution.

In some of the other types of tools, the physics is quite different than that treated in **Ardra**. For these cases we will study the development of modifications or different versions of **Ardra**, or the algorithms within, for more efficient handling of these other classes of tools. These would include gamma-ray tools, fast neutron tools, and tools based on the observation of secondary neutrons produced by fast or slow neutrons interacting with the geological medium.

References

1. M. R. Dorr and E. M. Salo, "Performance of a Neutron Transport Code with Full Phase Space Decomposition on the Cray Research T3D," in the *Proceedings of the International Conference on Mathematics, and Computations, Reactor Physics, and Environmental Analyses*, held in Portland, OR, April 30–May 4, 1995.
2. W. F. Miller and W. H. Reed, "Ray-Effect Mitigation Methods for Two-Dimensional Neutron Transport Theory," *Nuclear Science and Engineering*, **62** (1977), pp. 391–411.
3. P. N. Brown and M. R. Dorr, *Spherical Harmonic Solutions of Neutron Transport Systems via Discrete Ordinates*, UCRL-JC-119761, Lawrence Livermore National Laboratory, January 1995.
4. R. C. Little and M. Mickael and K. Verghese and R. P. Gardner, "Benchmark Neutron Porosity Log Calculations: A Comparison of MCNP and the Specific Purpose Code McDNL," *IEEE Transactions on Nuclear Science*, Vol. 36, No. 1 (1989), pp 1223–1226.

Large-Scale Fluid/Structure Interaction

G. L. Goudreau and Richard Procassini
Lawrence Livermore National Laboratory

Joseph Sabatini, Terry Bazow, and Frank Fernandez
AETC (Arete Engineering Technologies Corporation)

Project Objectives

The primary objective for this work is to demonstrate the use of the CRAY T3D in the simulation of a grand challenge problem, namely the structural acoustics problem. To do this, Lawrence Livermore National Laboratory (LLNL) will produce a parallel structural acoustics application that can be used by Arete Engineering Technologies Corporation (AETC) and the LLNL structural acoustics group for the analysis of complex geometry acoustic fluid-structure interaction problems.

LLNL will continue with the development of the acoustic fluid-structure simulation code, PING, and will map the application to the CRAY T3D in several phases. LLNL will run the PING test problems identified in the milestones below, and assist AETC in the running of further models they develop and wish to run.

The primary deliverable will be the parallel version of PING for the CRAY T3D. LLNL and AETC will prepare annual reports and a final report at the conclusion of the project. In addition, informal working notes, technical reports, and a journal article will be prepared to describe the research and simulation results. At the conclusion of the Cooperative Research and Development Agreement (CRADA), AETC will have access to PING and acquire a license for its further use.

General Approach

There are three primary phases of the project:

1. Port PING to the CRAY T3D.
2. Optimize the NIKE3D portion of PING on the CRAY T3D and optimize the acoustics imaging modules on the T3D.
3. Migrate the imaging software to the T3D or appropriate platform as required, and evaluate the simulation capabilities for the Intermediate Scale Measurement System (ISMS) experimental design.

LLNL and AETC will collaborate on the development of simulation capabilities for structural acoustics which can be applied to a broad range of problems. LLNL will focus upon the implementation of PING on the CRAY T3D. After establishing a clear understanding of the communication issues on the T3D. LLNL will port the acoustics component of PING to T3D and evaluate a message passing implementation. This effort will be followed by further optimization of the fluid-structure interface treatment, and the assignment of fluid subdomains to processors.

In the second phase, LLNL will focus upon optimization of the fluid-structure interface treatment and the optimization of the NIKE3D portion of PING on the T3D. During this phase, alternative equation solvers for the structural system will be investigated, and further optimization of the acoustics component, and load balancing among processors will be performed.

This final phase will focus on the evaluation of simulation capabilities with ISMS experimental design. LLNL will continue to focus on optimization of PING. LLNL will assist AETC by performing a large-scale ISMS simulation with PING, while AETC will port the imaging software to the T3D as required. This effort by AETC will help to tune and validate PING for further large-scale simulations by AETC and LLNL's own Maritime Systems Group.

Role of Industrial Partner

AETC will define the interface between PING and the imaging software for the rapid exchange of time history data. AETC will define the PING-imaging interface, will develop appropriate code modules for the efficient I/O of large-scale time history datasets, and will perform preliminary imaging computations using PING time history data to validate the I/O modules and the parallel PING results.

AETC will prepare the input deck model and will also provide data for the loading of a scale model geometry consistent with the ISMS facility and perform simulations with PING and their imaging software to demonstrate the use of large-scale scientific simulation for acoustic detection, identification and design. AETC will define the ISMS simulation parameters to be studied (frequencies, incident direction, etc.), and they will continue with the PING-imager validation studies. Final streamlining of the data interface between PING and the imaging software will be made at this point. AETC will aid in the validation and demonstration of the parallel version of PING for simulations which are beyond the current capabilities of conventional acoustic fluid-structure software.

Progress in FY95

The current state of the milestones represents the further implementation of the above general objectives and approach. These milestones are listed as 1–6 in Table 1.

Plans for FY96 and FY97

We have recently started the process of changing the coupling of the fluid and structural modules in PING from a very-tightly-coupled system in which both modules utilized the same data structures, to a scheme in which each module uses its own data structures and "shares" information via the use of interface data structures. This effort will have three important impacts on the code:

1. Reduction in the required storage, since nodal point associated exclusively with fluid elements will no longer add to the size of the stiffness matrix in the structural calculation.
2. Allow for 1D and 2D structural elements (currently available), as well as 3D structural elements, along with 3D fluid elements. Currently all 3D elements are fluid elements. This will allow us to model a larger variety of problems.
3. This moderately coupled version of the code will allow us to easily implement a three-segment partitioning scheme, with the following definition of each partitioning segment:
   —The fluid elements that are adjacent to the fluid-structure interface.
   —The remainder of the fluid elements.
   —The structural elements.
   This multi-segment scheme is necessary to allow us to achieve load-balanced computations over each of the three phases of the calculation: the explicit fluid calculation, the costly interface boundary conditions, and the implicit structural calculation.

Milestones for FY96 and FY97 are listed as items 7–21 in Table 1.

Presentations

1. R. J. Procassini, A. J. DeGroot, and J. Maltby, "An Analysis of Graph Partitioning Methods as Applied to the Decomposition of Three Dimensional Unstructured Finite Element Meshes," Supercomputing '94, November 14–18, 1994, Washington, D.C.
2. R. J. Procassini, "Parallel Implementation of a Structural Acoustics Code for Use on Massively Parallel Processors," Supercomputing '95, December 4–8, 1995, San Diego, CA.
3. R. J. Procassini, "Parallel Implementation of a Structural Acoustics Code on the Cray T3D Massively Parallel Processor," Cray PAPT Users Workshop, August 28–29, 1995, Pasadena, CA.
4. Mark Christon and G. L. Goudreau, "PING, the LLNL Structural Acoustics Code", ONR Annual Research Review, January 15, 1995, Boca Raton, FL.
5. G. L. Goudreau and James Foch, "Benchmark Studies with the LLNL Structural Acoustic's Code PING," and "Status of H4P CRADA T3D parallel port, and Large Problem Scaling Laws" (to be presented at ONR Annual Research Review, Feb. 12-16, Orlando, FL).
6. Terry Donich, multiple presentations to Navy and DOD Program officers, furthering the LLNL Maritime Systems Program, seeking funding for further development of PING.

Table 1. Milestones for the Fluid/Structure Interaction CRADA.
(NOTE: **Boldface** milestone denotes significant deliverable to partner)

Milestone	Scheduled for end of	Actual
Fiscal Year 1995:		
1. Evaluation of Interprocessor Communications	11/94	12/94
2. Port of Driven Boundary/Rigid Scattering	03/95	04/95
The parallel deliverable of the rigid scattering module of PING was met at the seven-month point.		
3. Optimize Wave Solver Communication	06/95	08/95
During the fourth quarter of this CRADA, we have focused upon the optimization of the message-passing-based communications within the parallel version of the PING code on the CRAY T3D MPP. We have tested the performance of both the Cray PVM (Parallel Virtual Machine) and ANL/MSU MPI (MPICH) message passing libraries for both blocking and non-blocking forms of data communication. Contrary to earlier measurements by Paul Marcelin of the National Energy Research Supercomputer Center, our tests indicate that PVM outperforms MPICH for both blocking and non-blocking communications, even though MPICH is implemented using only low-cost SHMEM "puts," while PVM uses both SHMEM "puts" and the more costly SHMEM "gets."		
4. Deliverable to AETC of deformable structure data set (nonparallel)	10/95	11/95
At the nominal time for the first prototype of parallel PING, LLNL made a new run on nonparallel PING of our best data case, the "MIT/NRL 4B" model, currently being studied in greater depth by the LLNL Maritime Systems Group under ONR funding. A full man week of our most experienced acoustician, Jim Foch, produced data by November 13 on the Cray version of PING.		
5. Provide AETC with T3D user account to move Beamformer code to T3D, and mesh generator manual	12/95	12/95
At the LLNL meeting November 30, user accounts were established for Frank Fernandez, Joseph Sabatini, and Terry Bazow of AETC. Bazow's password was implemented the following week, allowing him to begin the implementation of AETC Beamformer code onto the T3D. A user's manual for the LLNL INGRID mesh generator was sent to Bazow, to allow advanced preparation for his visit to LLNL in January or February.		

(Continued)

Table 1. (Continued)
(NOTE: **Boldface** milestone denotes significant deliverable to partner)

Milestone	Scheduled for end of	Actual
6. Prototype with AETC scalable mesh for simulation of acoustic response on non-axisymmetric structures	12/95	12/95
We are in the process of modeling large-scale (on the order of 100,000 element) scattering problems of the "4B" test object in serial on the CRAY J90 parallel vector processors at LLNL. The time histories of the perturbation pressures on the outer boundary of our computational mesh will be provided to our industrial partner, AETC, to allow them to validate their far-field projection software. This is intended to be an intermediate deliverable, which allows Arete to proceed with their software development and validation while we continue our parallelization of the structural module within PING.		
Coincident with AETC's evaluation of the 11/13/95 data set as still not sufficiently challenging due to the axisymmetric character of the model, Spike Procassini resumed his development of the generalized cylindrical body model with modest internals, replacing the conical transition to sphere on the ends with simpler, and more interesting, spherical caps. This model now has a completed and much improved fluid island design, can no longer be called "4B," and its four internal "King" frames can be used to mount simple internals representing nonsymmetric ballast to provide the nonaxisymmetric challenge that AETC desires.		
Fiscal Years 96 and 97:		
7. Host AETC technical visit; provide workstation version of INGRID mesh generator, and scalable mesh	02/96	
8. Perform first nonaxisymmetric analysis and provide results to AETC	03/96	
9. Separate the fluid (WAVE) and structural (NIKE3D) input parsing and data structures within PING	03/96	
10. Develop a methodology for three-segment partitioning of the fluid plus structure mesh to allow for load balanced parallel computation	05/96	
11. Conduct mid-year technical meeting with AETC to compare technical insights (at LLNL or San Diego)	07/96	

(Continued)

Table 1. (Continued)
(NOTE: **Boldface** milestone denotes significant deliverable to partner)

Milestone	Scheduled for end of	Actual
12. Implement basic parallel, iterative linear solver, plus pre-conditioning (as needed), to provide prototype parallel solution of the coupled problem	08/96	
13. Test prototype parallel code, and benchmark results against scalar/vector code	08/96	
14. Finalize with AETC scalable mesh for simulation of acoustic response on non-axisymmetric structures	09/96	
15. Perform two simulations with new mesh and on parallel machine; provide results to AETC	09/96	
16. Consult with AETC on their results of their Beamformer code validation procedure using data in (#15)	11/96	
17. Optimize computations and communications within the parallel structural module of PING	12/96	
18. Final optimization of parallel version of PING	04/97	
19. Integrate parallel graphics into T3D function	06/97	
20. Evaluate software with an ISMS Model (tbd)	08/97	
21. Final Meeting/Report and Exit Plan	09/97	

Finite Element Simulation of Fluid Dynamics and Structural Response for Industrial and Defense Applications

Richard Couch, Richard Sharp, Ivan Otero,
Rob Neely, Scott Futral, Evi Dube,
Rose McCallen, James Maltby, Albert Nichols
Lawrence Livermore National Laboratory

Project Objectives

A numerical simulation tool is being developed for application to a broad range of industrial and defense applications involving the dynamic interactions of fluids and structures. This software is intended to be used at a number of facilities, both in a "black box" mode and by expert users who are capable of incorporating their own physical models. Consequently, the issues of portability and extensibility become paramount in the software design process. The technical focus of the project is on the development of enhanced capabilities to simulate phenomena associated with metal casting. The computer science aspect of the project involves the development of a parallel version of the software that runs efficiently on the CRAY T3D.

General Approach

ALE3D is a finite element code that treats fluid and elastic-plastic response on an unstructured grid. A parallel version of ALE3D with enhanced physical models is being developed to meet the project requirements.

The development of 3D simulation tools at Lawrence Livermore National Laboratory (LLNL) in the areas of structural, fluid and thermal analysis has followed the traditional path of first developing capabilities limited to the particular topic. DYNA3D is the culmination of two decades of research in structural analysis. TOPAZ3D is the equivalent tool for use in static geometry thermal transport simulations. CALE, a 2D finite-difference arbitrary-Lagrangian-Eulerian (ALE) code provides the heritage for a fluid mechanics capability. ALE3D has been developed as a means of merging many of the capabilities developed in the individual technology areas. ALE3D was developed from a version of DYNA3D. It utilized the basic Lagrangian finite element techniques developed there, but has not maintained an identical set of algorithms as the two code efforts evolved along different paths. The treatment of solid elements, where fluid dynamics is treated, has been completely rewritten, but the coding and the available models for treating beam and shell elements have been kept consistent with the equivalent DYNA3D models, although only a subset are currently available. Fluid mechanics and ALE techniques from CALE were modified for application to unstructured meshes and incorporated into ALE3D. A version of TOPAZ3D has been incorporated into ALE3D to provide a thermal transport capability. The TOPAZ3D package has been enhanced by the inclusion of a reaction chemistry module. These capabilities are utilized in a split operator mode whereby the operator can be applied at a time interval that is appropriate for thermal effects and need not be consistent with the time step for the dynamics. Thermal and structural analysis techniques are generally developed first in DYNA3D and TOPAZ3D, then migrated to ALE3D as required. In addition, an implicit dynamic solution scheme is being specifically developed for ALE3D.

Role of Industrial Partner

The project has two components. One involves the development of models and simulation software capable of modeling a broad range of phenomena relating to aluminum casting technology. The other consists of implementing that software system on the CRAY T3D platform. Alcoa and LLNL are working together on all aspects of the problem, but each has primary responsibility for certain tasks. Alcoa is providing advice, data and material models related to ingot casting processes. LLNL is responsible for developing the simulation software and porting it to the T3D platform. Alcoa is responsible for providing the benchmark casting data and validating the code performance against these benchmarks.

Progress FY95

The initial serial version of ALE3D was used as a test bed for development of parallelization techniques. A "virtual" domain decomposition was constructed with message passing installed between the various domains. The various code modules could then be parallelized with little impact on continued development in the serial mode. We were able to demonstrate that the techniques being used produced scalable parallel computations. Experience gained in operating in this mode also provided insight into the appropriate structure for a new version of ALE3D that incorporates parallelism at the highest level while not compromising the requirements for ease of incorporation of new physical models and the support of multiple types of hardware platforms.

The new version of ALE3D is being written in C at the highest levels. Most of the lower level computational models can be incorporated essentially unchanged in their existing FORTRAN versions or written in C. Code modularity has been improved to facilitate development activities by our collaborators. Mesh decomposition, parallel I/O, and graphic rendering continue to be areas in which research is required. A significant component of our strategy is to utilize the concept of domain overloading. There need not be a one-to-one correlation between domains and processors. If one allows for multiple domains on a single processor, one can choose a domain size that optimizes computational efficiency for the processor/memory characteristics of a given platform. Having multiple domains on a single processor will also provide a savings in overall memory usage. The Lagrangian modules have been ported to this new version and performance has been verified.

The serial version of ALE3D continues to be the test bed for new physical and numerical algorithms. The thermal transport logic is based on a version of TOPAZ3D. This coding has been completely integrated with ALE3D. All input now goes through a single generation path and most of the features of TOPAZ have been made available. Significant problems are now being run. Chemical reactions have been integrated into the transport package and represent a source term for the transport. Integration of the reaction package into the hydro modules is ongoing. The implicit solver is to the point where simple problems are being run to test the coding and the numerical algorithms. While simple, these problems include most material models and boundary conditions of the serial version as well as the slide surface logic. ALE3D is being structured so that a transition from an explicit method to an implicit method and back again can be performed within a single calculational sequence. As experience is accumulated, options will be built into ALE3D to detect when it is appropriate to switch from one method to the other and make the transition automatically if that is the analyst's wish.

Table 1 lists milestones scheduled to be met within the first program year and their status are listed in Table 1.

Plans for FY96 and FY97

The project is proceeding at essentially the rate assumed in the original program plan. It is expected that the rate of progress will continue to closely follow the original plan. By the end of FY96 all the explicit code modules will have been ported to the T3D, along with the heat transfer modules. At that time the software will have also been validated against the initial set of benchmark problems.

Table 1. Project milestones for FY95.

Task	Agency	Status
Assembly of initial benchmark problems	Alcoa	Complete
Implement initial solidification model	LLNL/Alcoa	Complete
Implement initial mold filling model	LLNL/Alcoa	Complete
Implement semi-solid model	LLNL/Alcoa	Complete
Complete initial heat transfer modules	LLNL	Complete
Implement explicit Lagrangian coding	LLNL	Complete

Advanced Materials Design for Massively Parallel Environment

Christian Mailhiot and Lin H. Yang
Lawrence Livermore National Laboratory

To port and optimize current LLNL atomistic ab initio materials simulation codes to production MPP environments.

To perform algorithmic and formalism developments on these atomistic ab initio materials simulation codes in order to extend their efficiency and functionality.

John Northrup and Chris G. Van de Walle
Xerox Palo Alto Research Center

To apply these atomistic ab initio materials simulation codes to specific large-scale materials physics problems of significance to DOE and to the Electronic Materials Laboratory at the Xerox Palo Alto Research Center.

Project Objectives

The main goal of this project is to provide U.S. industry with breakthrough capabilities in advanced atomistic materials simulation by using innovative new MPP algorithms, methods, and computers. Moreover, many computational tasks of central importance in materials modeling and simulation have wide applicability, and their optimization would benefit other industrial areas of research outside the confines of *ab initio* electronic structure calculations of materials. Another technical objective is the application of advanced *ab initio* electronic structure methods, at the Xerox Palo Alto Research Center (PARC) and at Lawrence Livermore National Laboratory (LLNL), to specific physics problems of primary importance to research activities at the Electronics Materials Laboratory (EML), including (but not limited to) investigations of the atomic and electronic structure of amorphous silicon and defect energetics in III-V semiconductor materials.

General Approach

Our approach to this project involves three phases:

1. To port and optimize LLNL's materials simulation codes, which include the atomistic *ab initio* and force field materials simulation programs to the production MPP environments.
2. To perform algorithmic and formalism developments on various simulation codes in order to extend their efficiency and functionality.
3. To apply these atomistic *ab initio* materials simulation codes to specific large-scale materials physics problems of significance to DOE and to the Electronic Materials Laboratory at the Xerox Palo Alto Research Center, including (but not limited to):
   - Investigations of the atomic and electronic structure of amorphous silicon (a-Si). Specific issues:
     - — Spatial and energetic distribution of weak bonds.
     - — Band offsets between a-Si and crystalline Si, and amorphous alloys of silicon.
     - — Motion of hydrogen through a realistic amorphous network.
     - — Density of states.

  - — Role of voids in the amorphous structure.
- Investigations of defects in III-V materials. Proposed studies:
  - — Calculations of formation energies of native defects and impurity-defect complexes in GaAs and GaP and their dependence on atomic chemical potentials.
  - — Studies of the convergence properties of formation energies for charged defects with respect to supercell size for cells containing up to several hundreds of atoms.

Role of Industrial Partner

In FY95, Xerox PARC has defined and identified problems such as charge defect formation in GaAs as the first phase of problems to be tested. Xerox PARC has applied LLNL's materials simulation codes developed under the scope of this project.

Progress in FY95

During FY95, we have first optimized *ab initio* total-energy molecular dynamics methods on the CRAY T3D machine. The *ab initio* code provides a self-consistent treatment of electron charge rearrangements simultaneously with a description of ionic motions.

Over the duration of the first project year, substantial progress has been accomplished with respect to the parallelization of our *ab initio* total-energy molecular dynamics code on the CRAY T3D using the Shared Memory (SHMEM) Access Library. In particular, we have implemented the following two features in our parallel code:

1. Band (electronic orbital) partition which allows the minimization of the communications and the optimization of the load balancing.
2. Parallel I/O which speeds up the I/O linearly with the number of processors.

With these efforts, we have obtained high parallel efficiency up to 64 nodes, where the parallel efficiency is around 92%.

The optimization results were presented in the PATP Scientific Conference held at the Jet Propulsion Laboratory, Pasadena, CA, on August 24–25, 1995.

In addition, LLNL has started to test convergence of the applications with respect to supercell size and obtain structures and formation energies for native defects in III-V compounds.

Plans for FY96

In FY96, we plan to optimize and increase the parallel efficiency with nodes up to 256. We anticipate that we will attain parallel efficiency up to 84% for 256 nodes after the parallel linear algebra and 3D Fast Fourier Transform (FFT) routines are incorporated.

Moreover, we plan to apply the *ab initio* simulation code to the study of defects in III-V materials including:

- Calculations of formation energies of native defects and impurity-defect complexes in GaAs and their dependence on atomic chemical potentials.
- Studies of the convergence properties of formation energies for charged defects with respect to supercell size for cells containing up to several hundreds of atoms.

Publication

1. C. G. Van de Walle and L. H. Yang, "Band Discontinuities at Heterojunctions between Crystalline and Amorphous Silicon," *J. Vac. Sci. & Tech.* **B**13, 1635 (1995).

3D Massively Parallel Time-Dependent Computational Electromagnetics

D. J. Steich, N. K. Madsen, J. S. Kallman, S. T. Pennock,
C. C. Shang, B. R. Poole, Grant O. Cook, Jr., William G. Eme
Lawrence Livermore National Laboratory

Project Summary

We are developing parallel, time-dependent electromagnetic codes to analyze electrically large, geometrically complex three-dimensional structures. Current simulation efforts have placed computing requirements in the 60–72 million degrees of freedom regime.[1,2] There have been several advances in the time-domain computational electromagnetics (CEM) simulation technology, notably in nonorthogonal-grid Maxwell solvers and their parallel implementation, as well as improved boundary conditions to better capture the physics at the mesh truncation. The availability of improved capabilities in algorithms, boundary conditions, far-field transformations, post-processing, and grid-manipulation have allowed us to extend application into special regimes. These capabilities have application to a broad class of microwave and electromagnetic (EM) scattering problems of interest to our industrial partners. We have made progress in all these areas.

General Approach

The DSI3D code[3] allows one to accurately simulate electromagnetic scattering, coupling, and far fields. The underlying technique is unique for time-dependent field simulation in that it uses unstructured, non-orthogonal grids composed of a variety of convex, polyhedral element types.

The discrete surface integral (DSI) algorithm is based on the integral form of the Maxwell equations,

$$\int_A \delta \mathbf{B}/\delta t \bullet d\mathbf{A} = \int_\Omega -\mathbf{E} \bullet d\mathbf{l}\,,$$

Ω bounding A

$$\int_{A'} \delta \mathbf{D}/\delta t \bullet d\mathbf{A}' = \int_{\Omega'} \mathbf{H} \bullet d\mathbf{l}'\,,$$

Ω' bounding A',

and two interlocking grids, i.e., dual grids (nodes in one are element centers in the other). In the algorithm, the electric fields are computed along the edges of one grid, and the magnetic fields along the edges of the other (Figure 1). The integral equations are then easily applied at faces of each element (i.e., the line integral along the fields at the face edges yields the normally directed time derivative of the dual field).

Appropriate interpolation, or averaging, provides the face-tangential components. In the limiting case of logically regular grids, the algorithm reduces to the standard finite difference time domain (FDTD) algorithm. Moreover, it closely mimics the physics in that it

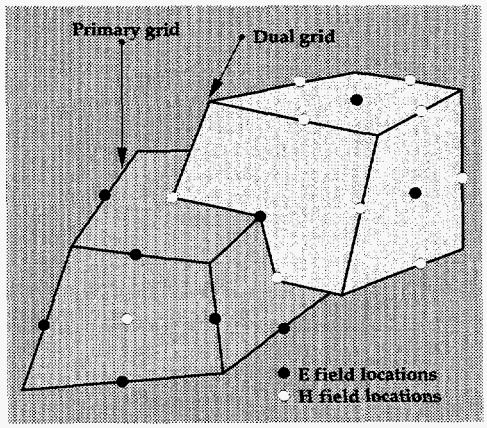

Figure 1. Dual grids used in the DSI algorithm, and the corresponding field locations.

preserves, both locally and globally, field divergence (charge) and is non-dissipative. However, if needed, user-defined amounts of dissipation can be added for highly distorted meshes.

The DSI algorithm uses a distributed memory model when running on parallel platforms. First, a given problem is partitioned into equal-sized numbers of cells using a recursive bisection method (RBM). Then, each processor independently computes the fields over the volume of cells that it owns. Next, the surface cell information is communicated (and received) from all its neighboring volumes. Afterwards, the fields are computed for the next time step using the newly updated surface information. The "compute over volume then pass surface information" sequence is repeated thousands of times, letting the fields progress forward in time.

Efforts this year have focused on boundary conditions, far-field transformations, vectorization, post-processing, and grid-manipulation tools. The DSI3D code is already a parallel-processing code. Various design and benchmark activities have involved analysis and design of photonics components, RF-microwave structures, radar cross-section analysis and control, and high-speed interconnect modeling.

Progress in FY95

Progress was made during this year in a number of areas, in order to extend the 3D time-domain CEM code capabilities into special regimes. The parallel implementation on the T3D has been accomplished. The code development strategy is to rely simultaneously on improved computational physics techniques to reduce the number of unknowns for proper solution convergence, as well as parallel computing to increase the envelope of problem size. In Table 1 we estimate the possible maximum enhancement factors using individual time-domain technology improvements. It is important to note that composite attainable speedups are not always multiplicative; specifically, we have been concentrating on items 1, 2, and 4 of Table 1 this year. During the year we have also added the capability to add user-defined amounts of dissipation. This allows the grids to be much more distorted than in the past. Higher accuracy on highly distorted grids is one of the focuses of our present research.

Benchmark Results for Scattering Physics

The DSI3D-RCS (radar cross-section) variant of the main code set has been applied to numerically evaluate radar cross sections on complex objects by solving the time-dependent curl equations. In order to generate each of the angle sweeps, it was necessary to run DSI3D once for each data point on the graph. This is a result of performing backscatter calculations where the incidence pulse comes from a different direction as the angle is changed. A sequence of calculations were performed on a standard class of cross-section geometries. Good agreement with measurements was obtained. Figures 2(a) and 2(b) show computed backscatter at 1.19 GHz for a perfectly electric conducting (PEC) almond target. The RCS backscatter results are within a fraction of a decibel of measured data. In addition, results for a wedge cylinder with plate extension, plate with half cylinder extension, rectangular plate, wedge cylinder with gap, and circular cavity results were obtained. [4]

Table 1. Time domain technology enhancement factor.

Technology	Enhancement Factor	Accuracy	CPU	Memory
Conforming Grids	up to 100x	•	•	•
Serial/Parallel	up to 60x		•	•
Structured/Unstructured	up to 30x		•	•
Improved RBCs	up to 20x	•	•	•
High order schemes	up to 8x	•	•	•
Adv. interface treatments	up to 4x	•	•	•

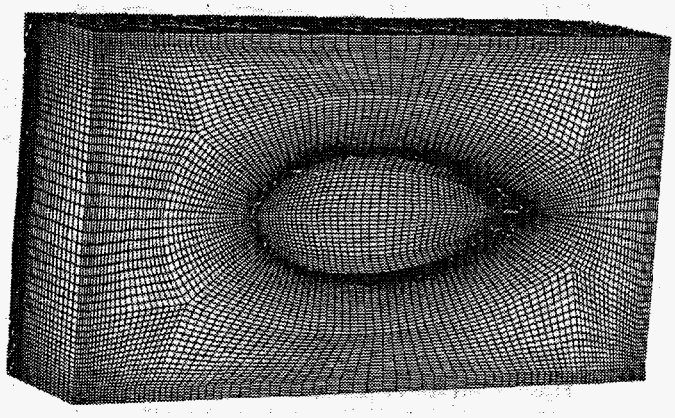

Figure 2(a). View of half of the PEC almond target mesh.

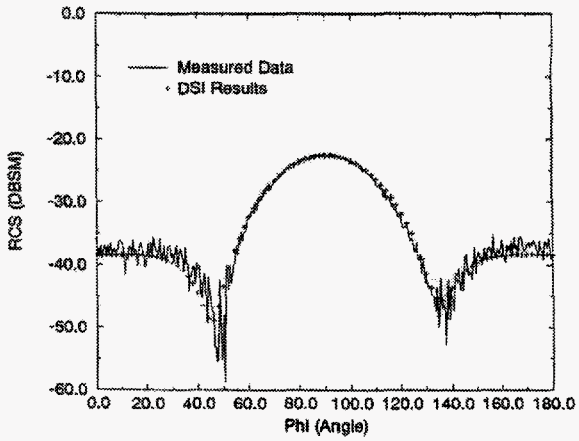

Figure 2(b) VV polarization cross section at 1.19 GHz for the PEC almond target.

Parallel Implementation of Field Solver

Recently, we have obtained parallel results for the CRAY T3D. For our benchmark 184,320 cell waveguide problem, we have concluded that significant speedups can be obtained. Table 2 shows typical speedups we are achieving on the CRAY T3D. Based on our experiences with the time-dependent wave-phenomena physics, a fundamental observation is that the explicit EM computing algorithms will significantly profit from increased memory per processor (e.g., lower surface-to-volume ratios). In addition, we have experienced some apparent bugs in the mpich.1.0.9, mpich.1.0.10, and mpich.1.0.11 versions of the message passing interface (MPI) software on the T3D. The results shown were obtained by running a fixed sized problem and subdividing the problem on more and more processors. These results are not to be confused with test cases where the workload per processor stays the same and the problem size grows with the number of processors. Near maximum theoretical speedups have been obtained when the problem size is allowed to change with number of processors while the workload per processor remained constant.

Table 2. Waveguide problem using a fixed 184,320 cells.

No. Proc.	Total time	Without I/O
8	1.00	1.00
16	1.81	1.83
32	3.73	3.80
64	7.34	7.44
128	10.16	13.26
256	10.61	17.92

Boundary Conditions

The primary boundary conditions can be categorized as either symmetry planes or absorbing boundaries. Symmetry planes require the tangential components of the electric, or magnetic, field to be either even or odd with respect to the plane (i.e., Neumann or Dirichlet boundary conditions), with the other field having the complementary symmetry. Perfect-electric-conductor boundary conditions yield the odd tangential electric fields. Even symmetry for the tangential electric fields, i.e., the Neumann boundary condition, is more difficult, but was implemented this year, using appropriately symmetrized path integrals of the magnetic fields, to time-integrate the tangential electric fields on the symmetry plane.

Differential operator forms of boundary conditions that we have implemented and tested include Mur,[5] Liao,[6] Higdon,[7] and Taylor.[8] Second-order forms performed reasonably well, but are currently restricted to structured, orthogonal grids, which is a big handicap in efficient grid generation for most of our applications. However, we have also successfully implemented first-order versions of these approximations, for unstructured

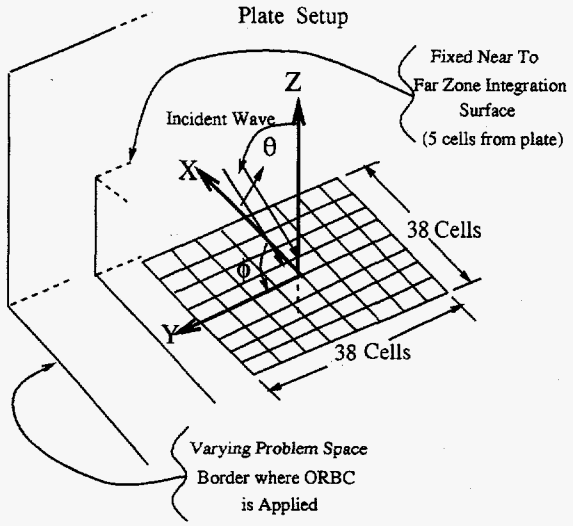

Figure 3. Problem setup for a PEC plate.

grids. In addition, there has been an investigation of other high-order boundary conditions.[8] In Figure 3 we describe, and in Figures 4(a) and 4(b) we examine, a radiation pattern convergence comparing standard second-order Mur with a new third-order Taylor boundary condition. In the calculational problem, we compute a plane wave scattering radiation pattern from a perfectly conducting plate at fixed angle of incidence. The legends show the required number of cells for a given result. From the data, it is clear the Taylor boundary condition with 80K cells is performing better than the Mur boundary condition with over 1.2 million cells.

One of our applications called for modeling two coupled waveguides (Figure 5). The driving waveguide is rectangular and single-moded in the frequency range of interest. The second waveguide is circular and runs parallel to the first, connected by an array of rectangular ports. To accurately model the fields near the ports, and to understand the multi-moded nature of the coupling in the circular guide, a dense grid is required, approaching one million nodes. Our best results to date for this large model showed excellent agreement with other calculations and with experiment for the lower half of the frequency range of interest. However, the higher frequencies require improvements in the absorbing boundary conditions. In addition, the question of resonances between the grid and geometry is under further study.

Far-Field Transformations

To compute the far-field values of the electric and magnetic fields that are required to evaluate radar cross-section values, we have implemented a near- to far-field transformation using the standard approach, as described by Kunz and Luebbers.[9] This technique requires effective electric and magnetic currents on an enclosing surface, and, consequently, requires total fields on that surface and an accurate, absorbing boundary condition outside the surface. The total fields can easily be computed in the DSI algorithm and interpolated to the surface positions. Validation of this transformation focused on ideal dipole solutions and applications to several canonical scatterers (e.g., sphere and metal "business card"), with reasonable to excellent results for low- to high-resolution grids, respectively.

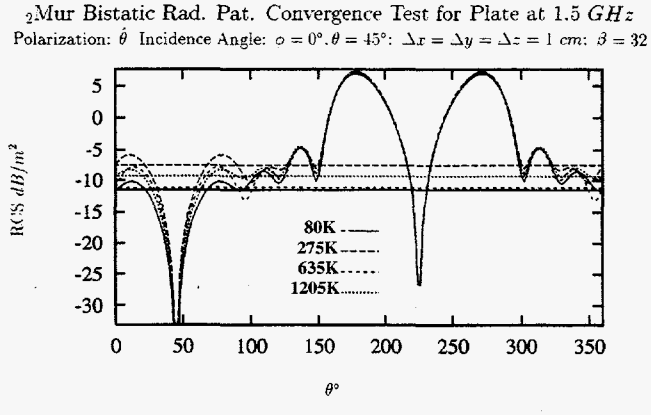

Figure 4(a). Bistatic radiation pattern results using second order Mur boundary conditions varying the required number of elements in the problem.

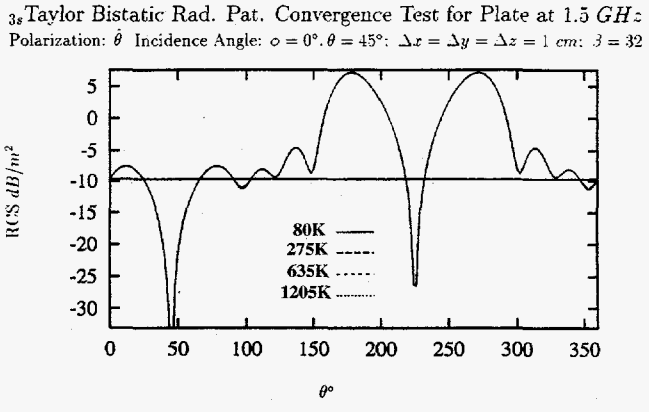

Figure 4(b). Bistatic radiation pattern results using a new third order Taylor boundary condition varying the required number of elements in the problem.

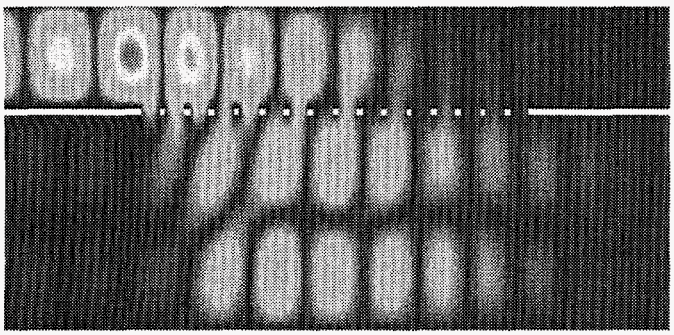

Figure 5. A 2D snapshot in time of the electric field as it propagates in coupled waveguides. The wave originates at the left in the top guide, and is transmitted into the lower guide as it passes over the 16 ports.

Role of Industrial Partner

Hughes is interested in incremental improvements in electrical designs for their product of microwave sources. Hughes' contribution to the effort is in experimental data on the performance of the electron devices. They also provide electrical data on materials and operating parameters to the device. The computer simulation code being augmented under this project will ultimately be used to design new interaction circuit geometries. The parameters provided by Hughes are input to the simulation code and the experimental data provided by Hughes allows us to validate the code. We, in turn, use the simulation code to solve other problems that support the programmatic needs of the laboratory.

Future Work (FY96–97)

Programmatic focus at Livermore is currently aimed at designing advanced accelerator components for advanced radiography and accelerator based production of tritium in addition to other scattering and radiation problems. These are in line with the deliverables of this CRADA (Cooperative Research and Development Agreement). Several of the simulation requirements have increased production computing requirements by an order of magnitude in problem size. In the next year, we will continue our focus on improving the absorbing boundary conditions and on extending material models to include both dispersive and nonlinear materials.

Continued work on an object-oriented DSI code (DSI-Tiger) will emphasize modern software design, hybrid meshes, efficient parallel implementation, and improved far field implementations. In addition, we will examine dynamic load balancing as an alternative or supplement to recursive spectral bisection techniques to achieve even better speedups using MPP machines. We also have begun writing our own memory management scheme for the new code. This will prevent our presently Fortran 77–based code from having to reallocate large blocks of memory based on best guesses of required memory use. The new memory management scheme will allow much more efficient use of memory on the T3D. Having more memory not only will allow us to run large problems but will reduce surface-to-volume ratios in our problems, thereby improving speedup performance. The DSI-TIGER effort is highly leveraged with other projects and is being designed to run hybrid (structured/unstructured) meshes on parallel platforms. Although hybrid meshes allow much more efficient use of computer memory and CPU time, they pose a real challenge from a load balancing point of view. This is another reason for looking into dynamic load balancing caused by the great disparity of required resources between structured and unstructured grids.

Acknowledgments

The authors wish to acknowledge useful discussions with S. T. Brandon, P. P. Weidhaas, R. L. Evans, and K. S. Kunz.

References

1. Poole, B. R., Ng, W. C., Shang, C. C., and Caporaso, G. J., *Theory and Modeling of Wakefields Generated by Relativistic Electron Beams in Long Accelerator Structures*, Tri-Laboratory Engineering Conference on Computational Modeling, November 2, 1995.
2. Molau, N. E., Genin, F. Y., Shang, C. C., and Kozlowski, M. R., *Electro-Mechanical Modeling of*

Nodular Defects in Multilayer Optical Coatings, Tri-Laboratory Engineering Conference on Computational Modeling, November 2, 1995.

3. Madsen, N. K., Steich, D. J., Cook, G. O., and Eme, W. G., *DSI3D-RCS Test Case Manual*, Lawrence Livermore National Laboratory, Livermore, California, UCRL-ID-121836, 1995.
4. Madsen, N. K., *Divergence-Preserving Discrete Surface Integral Methods for Maxwell's Curl Equations Using Non-Orthogonal Unstructured Grids*, Lawrence Livermore National Laboratory, Livermore, California, UCRL-JC-109787 (1992).
5. Mur, G., *IEEE Trans. Electromagn. Compat.* **23**, 377 (1981).
6. Liao, Z., H. L. Wong, B. Yang, and Y. Yuan, *Scientia Sinica (Series A)* **27**, *1063* (1984).
7. Higdon, R. L., *Math. of Comput.* **49**, 65 (July 1987).
8. Steich, D. J., *Local Outer Radiating Boundary Conditions for the Finite-Difference Time-Domain Method Applied to Maxwell's Equations*, Ph.D. dissertation, Pennsylvania State University, 1995.
9. Kunz, K.S. and R.J. Luebbers, *The Finite Difference Time Domain Method for Electromagnetics* (CRC Press), in press (1994).

Modeling of Shallow Junction Processing Technology

Tomas Diaz de la Rubia and M. J. Caturla
Lawrence Livermore National Laboratory

George H. Gilmer
AT&T Bell Laboratories

Objectives

The primary technical objective of this project is to develop a physics-based atomic-scale simulator for ion implantation modeling of semiconductor processing. This simulator consists of two modules, a molecular dynamics simulator and a kinetic Monte Carlo code, that expand all the relevant length and time scales of the process. These modules will be implemented on the CRAY T3D at the Lawrence Livermore National Laboratory (LLNL).

General Approach

There are three phases to this project. During the first two phases, codes will be developed, tested, and implemented on the CRAY T3D. In the third phase, the full simulator (BLAST: Bell-Livermore Atomistic Simulation Tool) will be implemented and tested and the result incorporated into AT&T's process modeling suite of tools.

Role of Industrial Partner

AT&T plays a critical role in all phases of the project. In particular, development of the new kinetic Monte Carlo code and the interface between binary collision and molecular dynamics codes will be done mostly at AT&T. Also, testing of the parallel molecular dynamics code (MDCASK) on the T3D will be carried out at AT&T.

In addition, AT&T suggests relevant and important calculations based on results of their experimental programs in this area. This serves both as motivation for calculations as well as validation of the simulation approach.

Progress in FY95

A new parallel classical molecular dynamics program (MDCASK) that uses the Stillinger–Weber potential for silicon has been implemented and tested on the 256-node CRAY T3D at LLNL. The code runs for silicon crystals up to 1 million atoms in size and performs at a rate of 2 Gflop/s on 256 nodes. This translates into molecular dynamics calculations at a CPU rate of 1 sec/atom/time-step for a crystal with 1 million atoms.

A new kinetic Monte Carlo simulation code (MCDONALDS) has been developed at AT&T. So far only a serial version of the code exists. The code performs Monte Carlo simulations of defect diffusion in silicon. The input defect field is obtained from molecular dynamics simulations of dopant implantation in silicon in the MDCASK code that are carried out on the T3D. The defects are then allowed to diffuse over time scales of seconds in silicon boxes 0.5 $\mu m \times 0.5\ \mu m \times 5\ \mu m$ in size. The hybrid MDCASK/MCDONALDS simulation code has been used at both AT&T and LLNL to provide information on the role of bulk and surface vacancy-interstitial recombination on the mechanisms of transient enhanced diffusion during high temperature annealing of implanted silicon.

In addition, a new 3D lattice Monte Carlo code has been developed by AT&T and LLNL to simulate coupled defect-dopant diffusion mechanisms in silicon. This code is thus far only available in a serial version.

Plans for FY96 and FY97

We plan to implement the kinetic and lattice Monte Carlo codes on the T3D during FY96 and FY97. In addition, we will continue to apply the MDCASK code together with the Monte Carlo simulators to the problem of transient diffusion in silicon.

Plans are also being developed to apply our suite of atomistic simulations codes to the problem of sputter deposition into aluminum lines into recessed features such as vias and trenches for aluminum interconnect fabrication. We will also investigate the mechanisms of void formation and aluminum grain boundary diffusion in Al–Cu interconnect lines.

Selected References and Publications in FY95

T. Diaz de la Rubia and G. H. Gilmer, *Phys. Rev. Lett.* **74**, 2507 (1995).

M. J. Caturla, T. Diaz de la Rubia and G. H. Gilmer, *J. Appl. Phys.* **77**, 3121 (1995).

G. H. Gilmer, T. Diaz de la Rubia, M. Jaraiz, and D. Stock, *Nucl. Instr. Meth. B* **102**, 247 (1995).

M. J. Caturla, T. Diaz de la Rubia, and G. H. Gilmer, *Nucl. Instr. Meth. B*, **106**, 1 (1995).

The Gang Scheduler

Bruce Griffing, Moe Jette, Dave Storch, Emily Yim
Lawrence Livermore National Laboratory

Overview

Most scalable parallel systems available today support the notion of space sharing of system resources among contending jobs, but do not support the notion of timesharing the entire parallel machine. Cray Research Inc.'s (CRI) CRAY T3D is no exception, since its default mode of operation is the allocation of job partitions from the pool of available processors. These job partitions are held until the job relinquishes them, effectively locking out any other use for those processors. With such a processor allocation mechanism, the computational requirements of long-running production jobs directly conflict with those of interactive development jobs. Only a preemptive processor scheduler could satisfy the requirements of all clients.

The Gang Scheduler is a preemptive processor scheduler. It schedules processors and barrier circuits for all jobs. The Gang Scheduler allocates processors based upon the classes of jobs available for execution in such a fashion as to satisfy the diverse computational requirements of our clients, even when these computational requirements are incompatible. For example, the Gang Scheduler can simultaneously provide timely response for interactive computing and long run times for production computing.

Role of the Industrial Partner

The foundation of the Gang Scheduler is the ability to stop a running job and move its state to disk, then later to return the job's state into processors and restart it. Cray Research, Inc., developed this capability and integrated it into their UNICOS MAX operating system. The software developed by Cray Research even permits the job to be restored to different processors, permitting much greater flexibility in scheduling processors, and much greater efficiency than was possible in earlier systems using the Gang Scheduler.

Progress

The first phase of the project was porting all of the original code from a BBN TC2000 to a CRAY T3D, which was completed in October 1995. Several problems were identified in early versions of Cray Research's software to save and restore state of T3D jobs (rollin/rollout). Cray Research has resolved most of the known problems in the rollin/rollout logic. Substantial differences in hardware configuration required new Gang Scheduler logic for assignment of processors and barrier wires. This code was completed and integrated in October. The original Gang Scheduler was designed to support a near instantaneous context switch. Since significant time is required for a context switch on the CRAY T3D due to its memory management scheme, many synchronization difficulties were encountered in the Gang Scheduler code. Several new job states and substantial additional logic was added to the Gang Scheduler to address these synchronization difficulties in the months of October and November 1995. We have designed and written new global processor and barrier wire management logic, patterned after Cray memory management and described below. We have begun documentation development, which is available on the World Wide Web at URL **http://www.nersc.gov/doc/gang**.

Plans

We plan to complete the integration of the new global processor and barrier wire management logic in early December 1995. Beta testing of the Gang Scheduler is planned to start in mid December. Future work will include controls for resource use by user and group, optimization of the scheduler, and additional documentation of this work.

Processor Scheduling

We have found that the timesharing of processors on a CRAY T3D is not a very different problem than timesharing real memory of large computers. The functionality of the Gang Scheduler and its parameters are patterned after memory schedulers used on Cray's vector computers, although several additions were made specifically for this environment.

We have identified four distinct job classes:

1. Interactive class jobs receive the most responsive service.
2. Production class jobs receive better throughput, but less responsive service (equivalent to "batch").
3. Benchmark class jobs are not preempted, but may experience the least responsive service.
4. Standby class jobs are allocated processors that would otherwise be unused.

Each job class is assigned a relative priority. The highest priority job class is given preferential treatment for scheduling. Each job class can only preempt jobs of an equal or lower job class. Within a job class, jobs are scheduled based upon their time waiting to be assigned processors.

We have identified four job class dependent scheduling parameters:

1. Wait time: The maximum wait time desired.
2. Do-not-disturb time: Multiplied by the number of processors to specify the minimum processor allocation time before preemption.
3. Processor limit: The maximum number of processors which can be allocated to jobs of this class.
4. Priority: Job classes are prioritized for service.

The wait time is designed to insure timely responsiveness, especially for interactive jobs. After a job has waited to be loaded for the maximum wait time, a block of processors will be reserved for it. Jobs occupying the reserved processors will be preempted as soon as their do-not-disturb time has been exhausted. When the last of these processors has been made available, the long waiting job will be allocated the reserved processors. Presently, a job can only preempt jobs of an equal or lower job class. We plan to experiment with this algorithm and consider the preemption of higher job classes when a job's maximum wait time has been reached. If the maximum wait times are set to zero, jobs will be loaded as soon as possible.

The desire for timely responsiveness needs to be balanced against the cost of moving jobs from their assigned processors onto disk. In order to prevent job thrashing, a job is assigned processors for a minimum of its do-not-disturb time before preemption. The do-not-disturb time parameter for each job class is multiplied by the number of processors assigned to arrive at the minimum time before preemption. The do-not-disturb time parameter should be set to a value substantially larger than the time required to move a job's state from memory to disk and back to memory. These parameters can result in conflicting job scheduling rules if not set appropriately or if the workload is irregular. Performance will degrade under such circumstances, but "reasonable" behavior can be expected with the highest priority job classes receiving the best performance.

We have also identified three job class independent scheduling parameters:

1. Large job size: The minimum number of processors requested by a job for it to be considered "large."
2. Large processor limit: The maximum number of processors which can be allocated to "large" jobs.
3. Job processor limit: The maximum number of processors which can be allocated to any single job.

These job scheduling parameters may be altered in real time to provide different performance characteristics at different times of the day. For example, it may be desirable to provide a lower level of interactivity at night. The reduced level of job swapping (and reduced time for such idled processors) would result in improved system throughput. It might also be desirable only to execute benchmark class jobs only at night or on weekends.

Administration Tool

A system administration tool exists which can alter scheduling parameters, job class, and job base node. Administrators can also explicitly suspend or resume scheduling of specific jobs.

Job Execution

The execution of a user's job is passed through a Gang Scheduler interface. This interface is built upon the CRI default interface and is upwardly compatible with it. The interface registers the job with the Gang Scheduler and waits for an assignment of processors and barrier circuit before continuing. This typically takes a matter of seconds for small numbers of processors, and possibly much longer for large numbers of processors. The only additional argument to the Gang Scheduler interface is the job class, which is optional. By default, interactive jobs are assigned to the interactive job class and batch jobs are assigned to the production job class.

Job Monitor—Gangster

We provide users with an interactive tool, Gangster, for observing the state of the system and controlling some aspects of their jobs. Gangster communicates with the Gang Scheduler to determine the state of the machine's processors and individual jobs. Gangster's three-dimensional node map displays the status of each node (each node consists of two processing elements on the T3D). Gangster's job summary reports the state of each job, including jobs moving between processors and disk (see the sample display below). Users can use Gangster to change the class of their own jobs, which affects accounting. Users can also explicitly suspend or resume scheduling of their jobs.

Sample Gangster Display

Figure 1 is a sample Gangster display, which identifies jobs in the system and assigned processors. The node map is on the left. A dot or letter denotes each node (two processing elements on the T3D): a dot indicates the node is not in use, a letter designates the job currently occupying that node. On the right is a summary of all jobs. The ST field reports the job's state: R = running; o = being moved out to disk; I = scheduled to be moved into processors; W = waiting for execution.

Node number 000 is in the upper left corner of the lowest plane. The X axis extends down and left within a plane. The Y axis extends up, with one Y value in each plane. The Z axis extends to the right. This orientation was selected for ease of display.

```
            gangster - 12445
   b  b  a  a  .  d  h  h  CLAS JOB-USER      PID  COMMAND  #PE BASE  W ST H:MM
  b  b  a  a  .  d  h  h   Int  b - hanchen  72566 sigma700  32 020   0 R 01:27
 b  b  a  a  e  f  h  h    Int  d - hanchen  80820 B600       8 520   1 R 00:09
b  b  a  a  e  f  h  h     Int  e - hanchen  80846 B500       8 424   0 R 00:09
                          Int  f - hanchen  80874 B550       8 524   1 R 00:09
   b  b  a  a  g  d  h  h  Int  g - shesty   81317 icf3d      2 420   0 R 00:02
  b  b  a  a  .  d  h  h   Int  j - jette    82459 delta64   64 000   0 W 00:00
 b  b  a  a  e  f  h  h
b  b  a  a  e  f  h  h     Prod a - hanchen  66242 relax     32 220   0 R 02:26
                          Prod c - gpt      76381 kiten     64 400   0 R 00:51
   i  i  i  i  c  c  c  c  Prod h - dan      81350 camille   32 620   1 R 00:01
  i  i  i  i  c  c  c  c   Prod i - vickie   81393 kiten     64 000   0 o 02:12
 i  i  i  i  c  c  c  c    Prod k - dan      82545 camille  128         N 00:00
i  i  i  i  c  c  c  c

   i  i  i  i  c  c  c  c
  i  i  i  i  c  c  c  c
 i  i  i  i  c  c  c  c
i  i  i  i  c  c  c  c

gangster: ?
 ? help quit refresh kill base int prod nosched swapout swapin !
```

Figure 1. Sample Gangster display.