Table of Contents Author Index Home

Get Acrobat Reader Get Postscript Version Get PDF Version

Parallel Computing Applications and Environment on the T3D

A. E. Koniges
Leader, Multiprogrammatic and Insititutional Computing Research
Morris A. Jette
Distributed Computing Technology Group Leader
Lawrence Livermore National Laboratory
Livermore, California

koniges@llnl.gov
M&IC
www.llnl.gov

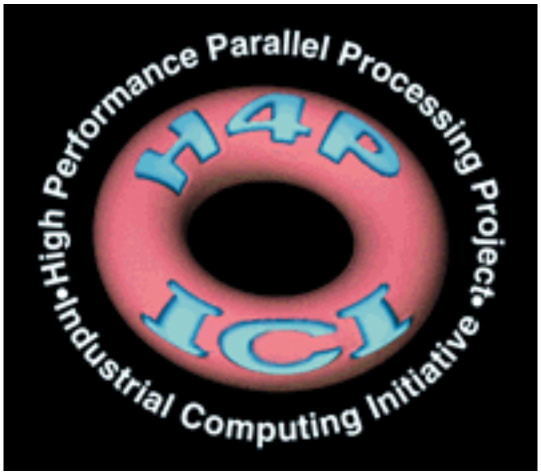
ABSTRACT:

As part of the Parallel Applications Technology Program of Cray Research, a 256 processor T3D was sited at Lawrence Livermore National Laboratory in 1994. Today, that machine has become a workhorse of unclassified supercomputing for the laboratory with utilization rates of 90 - 96% or better. This talk will cover the range of applications on the T3D, performance highlights, and information on how to use the MPP platform as a production computer for both industrial and academic applications.

KEYWORDS:

Applications, production MPP computing, scheduling

Introduction


High Performance Parallel Processing Project
H4P
ICI
Industrial Computing Initiative

The High Performance Parallel Processing Project (H4P) is a package of 9 individual CRADAs (Cooperative Research and Development Agreements), plus hardware (a 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 then 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 (full-time equivalent) 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.

An Environment for Supercomputing Applications

The CRAY T3D is a distributed memory parallel supercomputer with a cache-based chip architecture typical of a distributed memory machines available in the late 1990's. Each node on the T3D has two independent CPUs and each CPU has 64 MBytes of memory. The CPU is a DEC alpha EV4 clocked at 150 Mhz. 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 upcoming years are worth the effort.

Outline

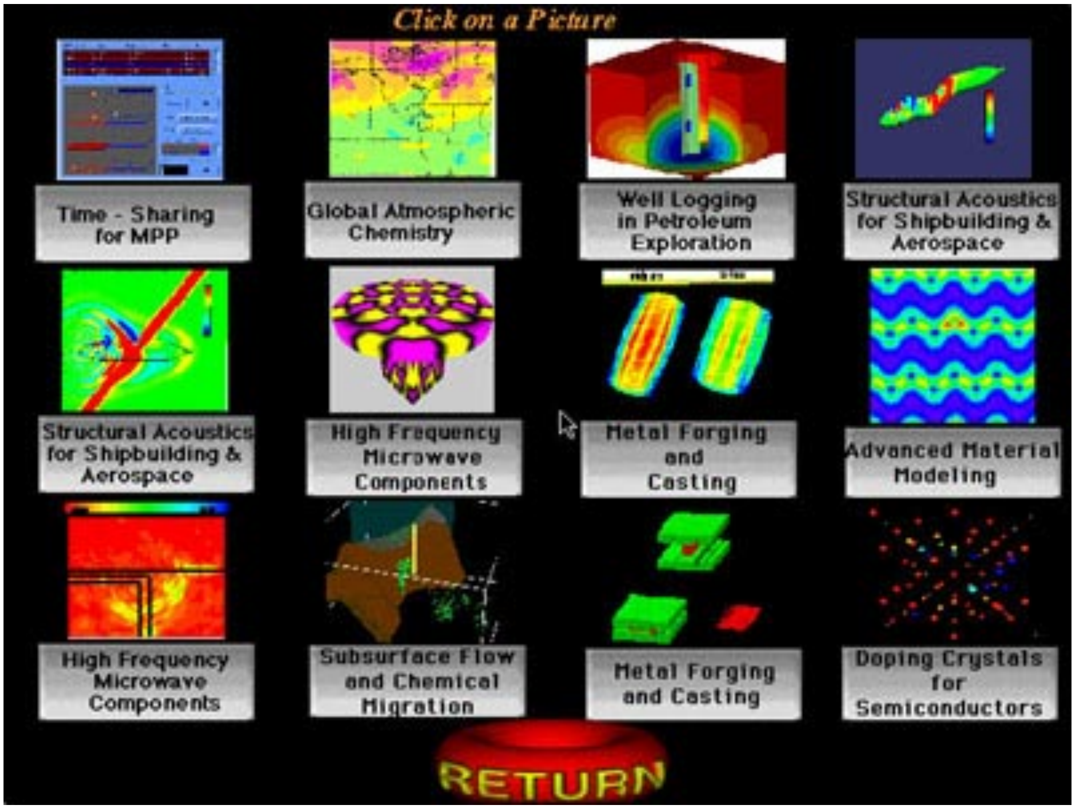
- The Parallel Applications Technology Program at LLNL
- Enabling Technology for the Production System
  - Queues
  - Gang Scheduler
- Sampling of Results: Performance and Scaling
- Retrospective

High Performance Parallel Processing Project (H4P), also known as one of 5 Parallel Applications Technology Program or PATP sites by CRI.

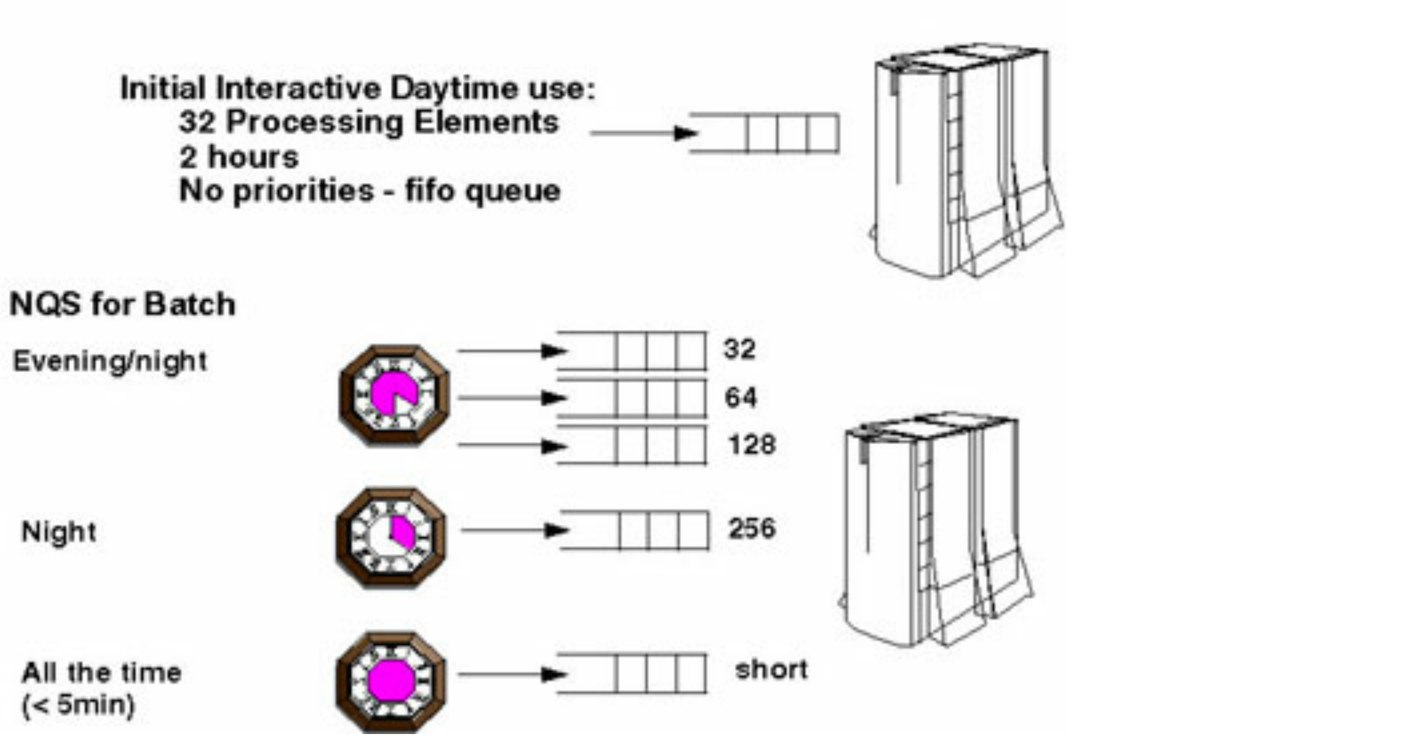
- LLNL's piece of the Largest DOE CRADA Package (~40 million over 3 years), announced by O'Leary as the"SuperCRADA"

- A collaborative research project aimed at putting real applications on MPP'S
- Industry Partners at LLNL- Alcoa - Boeing - Arete Associates - Cray Research (CRI)- AT&T - Xerox-Halliburton - Hughes- IT Corp.
- 256PE CRI T3D at LLNL

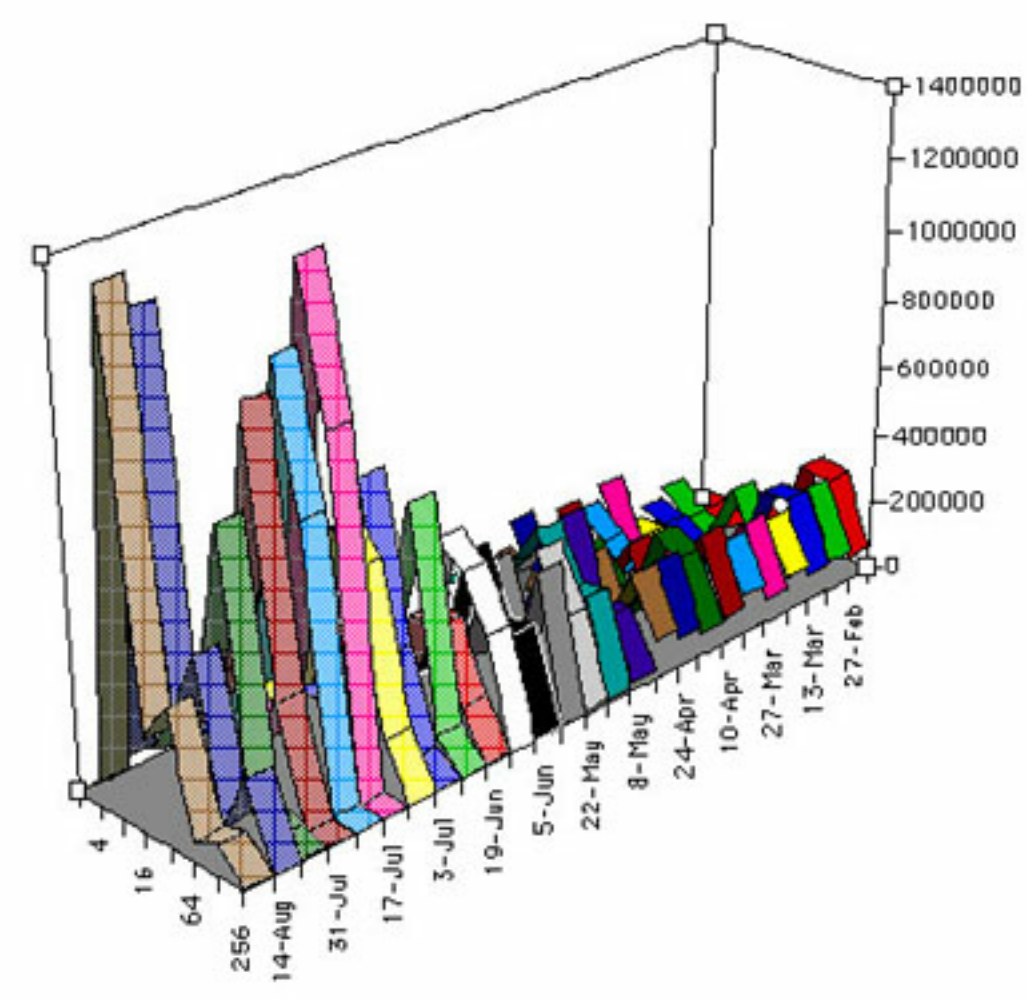
Collage (from the project CD-ROM) shows the range of applications.


Click on a Picture
Time - Sharing for MPP
Global Atmospheric Chemistry
Well Logging in Petroleum Exploration
Structural Acoustics for Shipbuilding & Aerospace
Structural Acoustics for Shipbuilding & Aerospace
High Frequency Microwave Components
Metal Forging and Casting
Advanced Material Modeling
High Frequency Microwave Components
Subsurface Flow and Chemical Migration
Metal Forging and Casting
Doping Crystals for Semiconductors
RETURN

Over the 3 year period, usage substantially increased both with new user base and improved software. With a small number of users and jobs, scheduling was relatively easy as shown in the following diagram.



However, very shortly the machine started to require more aggressive scheduling techniques. The following chart shows an initial surge in the machine usage.



CPU sec as a function of the number of processors from Feb 95 to Aug 95. This is before implementation of the Gang Scheduler. Notice the large number of small processor jobs due both to people learning MPP techniques and scheduling limitations.

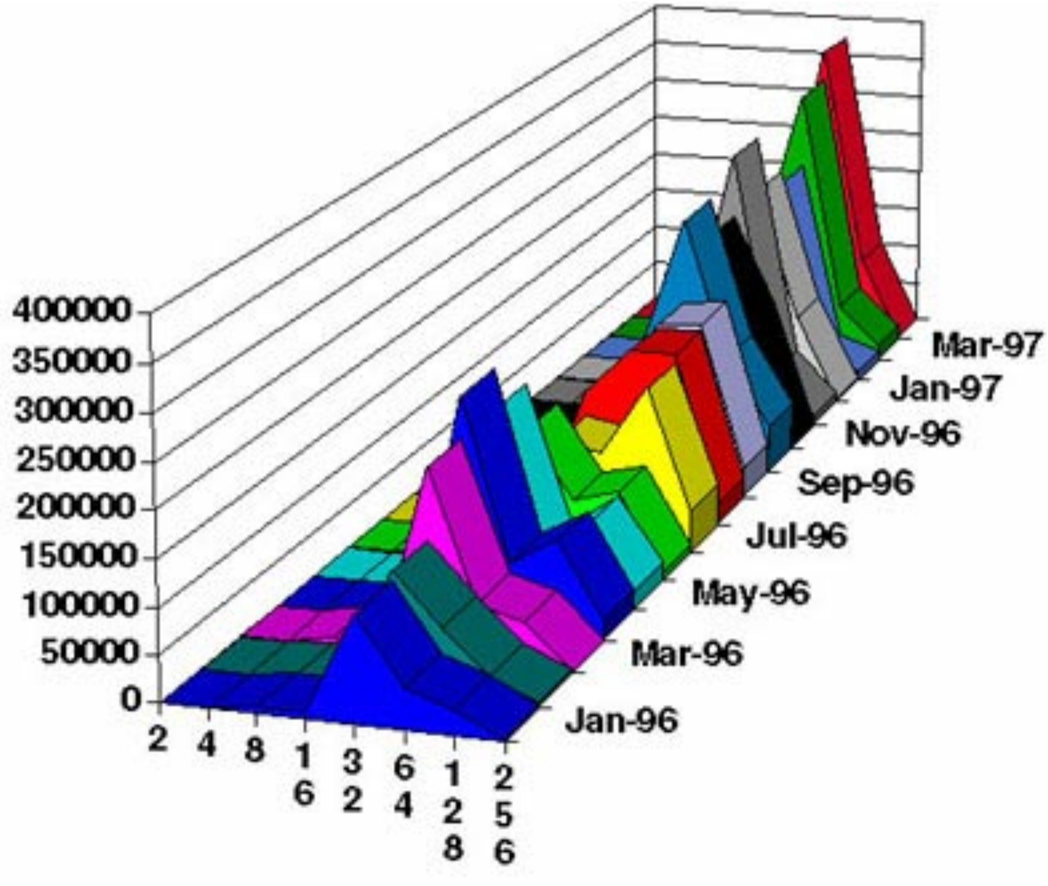
The growth in user numbers required a large variety of queues with some dynamics in order to leave an adequate number of processors available for interactive computing. The following tableshows the initial NQS processor limits at various times of the day.

Queue Name	User Limit	Time Limit	Aggregate PE Limit	Time of Day
128-s	1	15min	64 128	0800-1800 1800-0800
32	1	4hr	128	
64	1	4hr	192	
128	1	4hr	256	
256-s	1	15min	256	
64-L	2	19hr	96	0400-0800
256	1	4hr	256 96 192	0000-0400 0400-1800 1800-0000

Gang Scheduler was developed jointly by LLNL and CRI to deal with increased load

- Tool for providing time- and space- sharing
- Uses roll-in/roll-out for coarse-grained time sharing
- Repacks jobs to achieve high CPU utilization
- Five classes of jobs with different scheduling characteristics
    - interactive: excellent response and throughput during working hours
    - debug: rapid response during working hours, cannot be preempted
    - production: excellent throughput outside of working hours
    - benchmark: cannot be preempted
    - standy: low priority

The following graph shows the improvement in system usage with installation of the gang scheduler.



CPU seconds as a function of number of processors. Note the trend towards larger jobs and better throughput with gang scheduler.

Gang Scheduler Progress

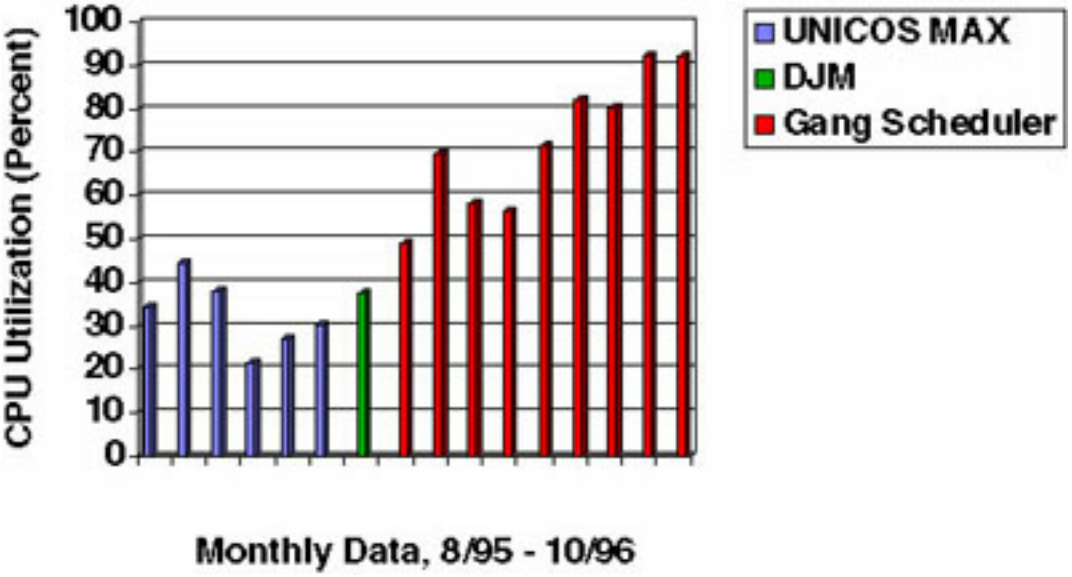
- Roll-in/Roll-out
  - soon saw speeds of 100MB/sec (up from 3MB/sec!)
  - allows re-packing of torus
  - should be able to re-pack the entire machine in less than 6 minutes.
- History
  - First used only in testing mode: (gang scheduler knows about all jobs, but will only "schedule" jobs under its control)
    - i.e. Manages only job's registered to it, other jobs will have resources freed to load them and will not be rolled out
  - Much work on improved scheduling algorithms
  - Needed integration into resource allocation system so jobs are moved to/from standby class to control resource use
- Weekly CPU utilization rates over 95 percent
- Interactive workload slowdown of 18 percent (This means the interactive workload completes in a

wall-clock time that is 118% of CPU time on a saturated machine, on par with SMP architectures)

Details of Gang Scheduler Development.

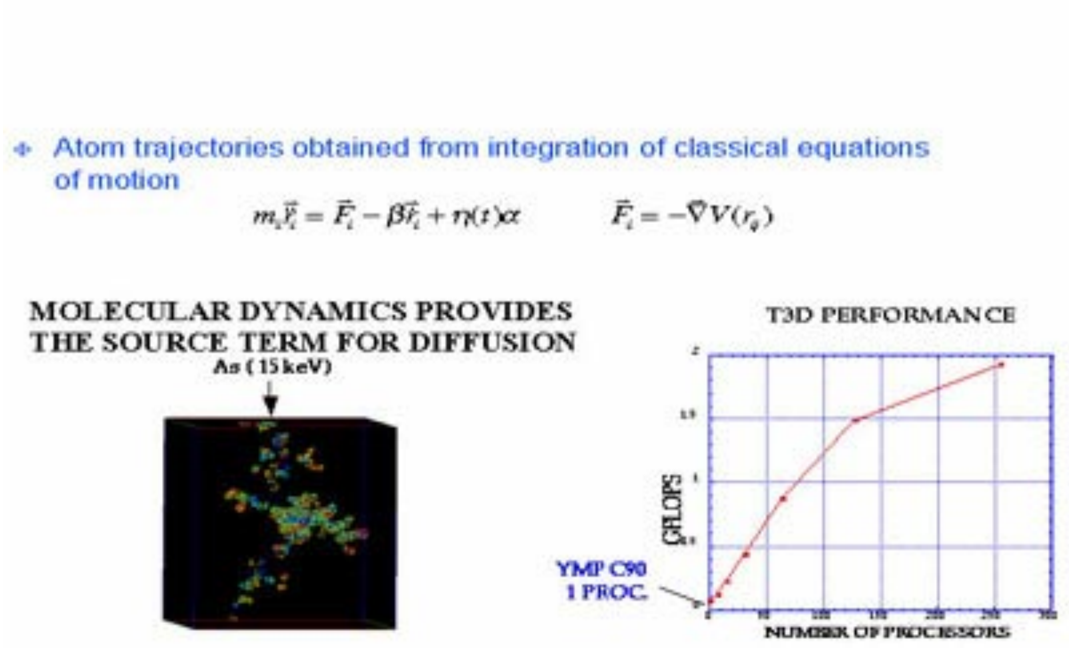
Gang Scheduling offers better throughput, allows users to run large memory jobs

- Concurrently schedules related processes and threads for optimal parallelism and to permit efficient execution of larger problems
- Preempts jobs as needed to optimize interactivity and resource utilization; especially for important for very large jobs
- Relocates jobs for improved processor utilization
- Provide graphical user interface tool for tuning parallel jobs and the operating system
- Gang scheduling across heterogeneous computing environment under development



Samples of the Applications

Molecular Dynamics Runs at 2 GFlops for Microelectronics



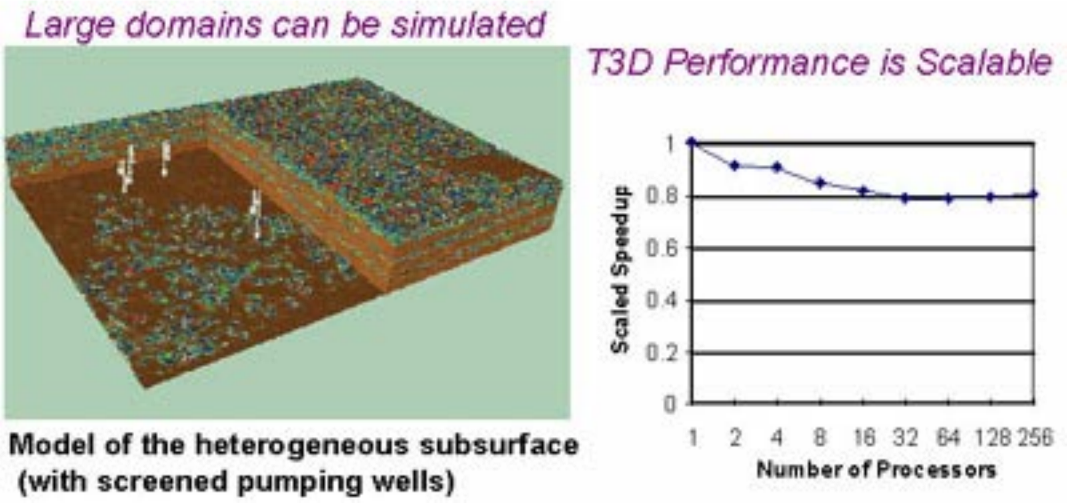
- Atom trajectories obtained from integration of classical equations of motion

$$m_i\ddot{\vec{r}}_i = \vec{F}_i - \beta\dot{\vec{r}}_i + \eta(t)\alpha \qquad \vec{F}_i = -\vec{\nabla}V(r_{ij})$$

MOLECULAR DYNAMICS PROVIDES THE SOURCE TERM FOR DIFFUSION

T3D PERFORMANCE

Ground Water Flow Simulations are now 3D in Heterogeneous Materials



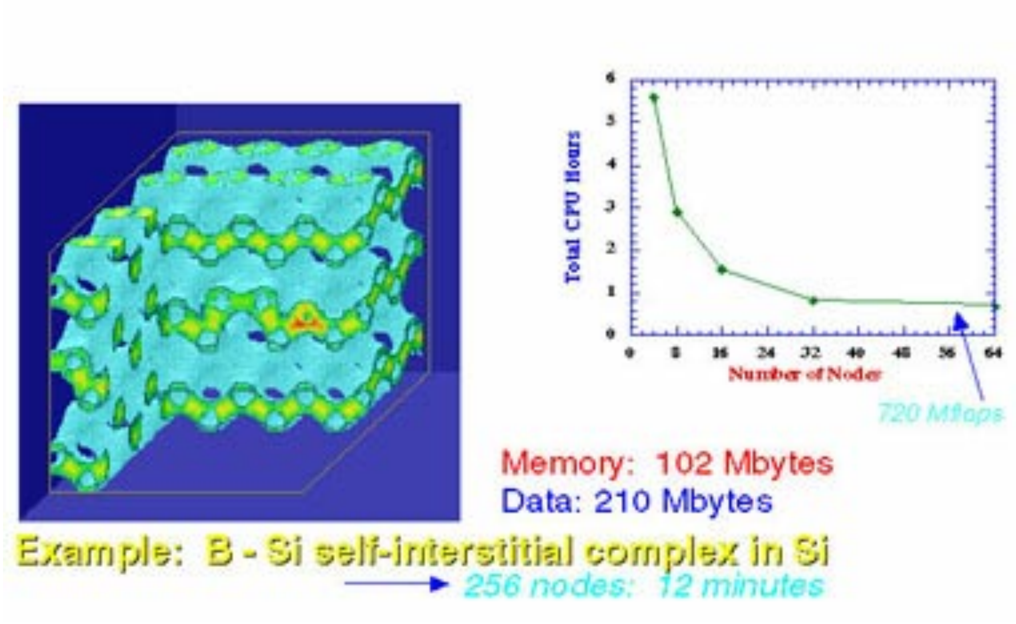
Large domains can be simulated

T3D Performance is Scalable

Model of the heterogeneous subsurface (with screened pumping wells)

Semiconductor Design is fast and T3D allows for large memory implementation

>



720 Mflops

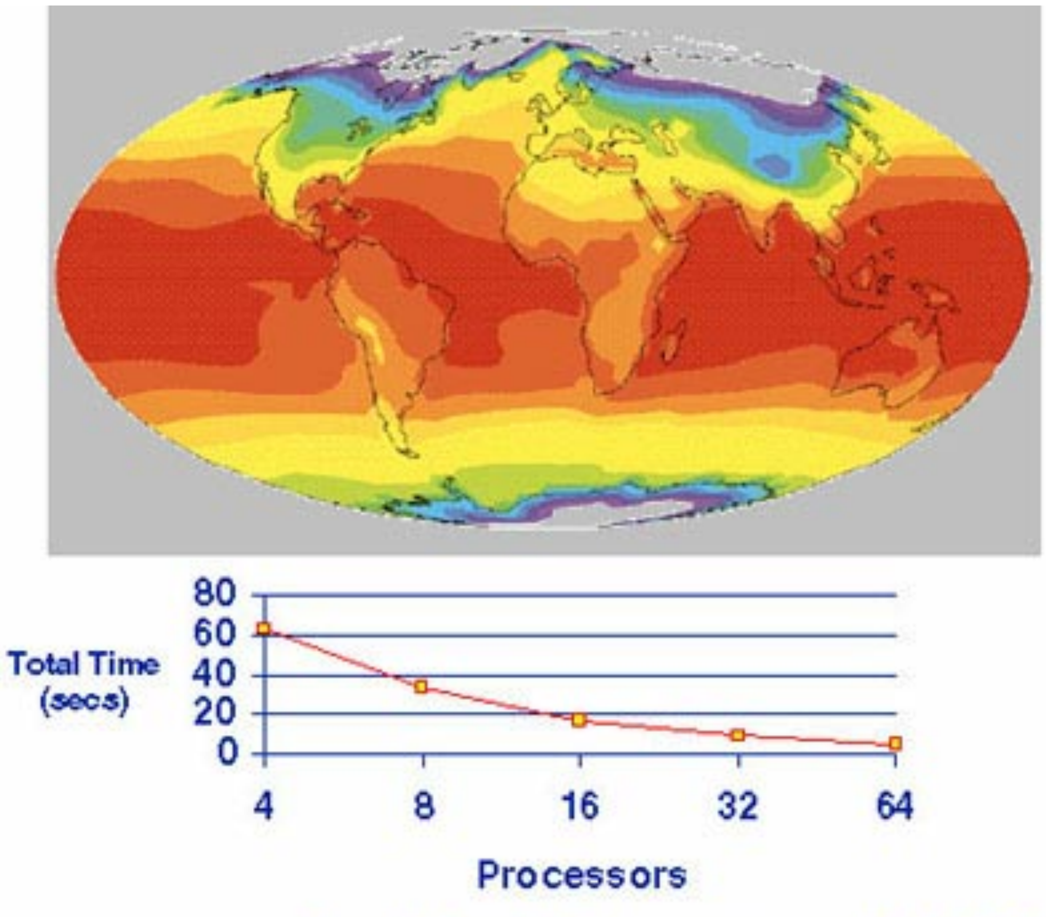
Memory: 102 Mbytes

Data: 210 Mbytes

Example: B - Si self-interstitial complex in Si

→ 256 nodes: 12 minutes

Climate and Chemistry Models forcast impact of Next Generation Aircraft



Scalable Performance on the T3D

Gang Scheduler research is being continued and adapted for the new DEC Cluster at LLNL

- Software to monitor the status of computer, processes and threads completed
- Graphical user interface to provide computer and job status information completed
- User API to register jobs completed
- Gang scheduling within a computer and across multiple computers completed (except for operating system integration)
- Operating system integration and testing underway

Summary of PATP Experiences

- Large Memory users are very pleased
  - roughly 10X larger than will fit into the C90 200 Megaword limit; also much better than workstations
- Performance depends on level of effort
  - Very high performance is possible with CAM coding e.g., 45 mflops per node (Pierce) and 80 gflops on 1024 processors (Salo), though this was not used significantly in the actual

applications
- Codes often out-perform vectorized C90 at 20-40 processors
- Very scalable performance
- T3D has become a workhorse machine
  - useful for users learning the MPP platform
  - production as available (roughly 4 times slower than T3E)
  - Platform for porting to latest architectures Weekly utilizations over 95 percent with very good interactivity
- Users have found that the move from the T3D to newer machines such as the T3E is generally very easy. For example, see the paper on Nimrod.x

References

[1] Dror G. Feitelson and Morris A. Jette, "Improved Responsiveness and Utilization with Gang Scheduling",Job Scheduling Strategies for Parallel Processing Workshop IPPS, April 1997.

[2] Morris A. Jette, "Gang Scheduler, Timesharing on a Massively Parallel Supercomputer",SC96, November 1996.

[3] A. E. Koniges and K.R. Lind, "Parallelizing Code for Real Applications on the T3D", Computers in Physics 9, 39 (1995).

Acknowledgments

- Work at LLNL for DOE under Contract W7405-ENG-48.

Table 1: The H4P Project at LLNL consists of 9 projects spanning FY95-97

PI Name/Directorate	Project	Technology	Partner(s)	MPP/Code info
(Defense and Nuclear Technology) **Richard Couch**	Finite Element Fluids and Structures	Metal forming, manufacturing	**Alcoa**: Don Ziegler **CRI:** Steve Bernard	Codes: ALE3D and ALEC Domain decomposition for MPP

(Chemistry and Material Science) **Tomas Diaz de la Rubia,**	Shallow-Junction Device Modeling	Microelectonics/device modeling characteristics at microscopic level.	**AT&T**: George Gilmer **CRI**: Kevin Lind	Molecular dynmaics and Monte Carlo
(Computations) **Moe Jette**	T3D Gang Scheduler	Software for MPP systems	**CRI:** Steve Luzmoor	GUI design, roll-in/roll-out for MPP
(Computations) **Peter Brown**	Nuclear Imaging	Petroleum exploration-nuclear well logging	**Halliburton** Larry Jacobsen	Codes: Ardra, AMTRAN
(Computations) **Steve Ashby**	Environmental Remediation	3-D Subsurface flow in heterogenous materials	**I T Corporation** **CRI:** Kevin Lind	Code: Parflow MPP algorithms for conjugate gradients
(Engineering) **Cliff Shang**	Computational Electromagnetics	3-D dynamic E&M field solver for laser, radar and antenna design for high clock rate micro-electronics.	**Hughes Air Craft:** E. Illoken	Parallel mesh generation, Parallel PDE solvers
(Engineering) **Jerry Goudreau**	Fluid Dynamics, Acoustics, Structures	Acoustical studies of submarines	**Arete Corp:** F.L. Fernandez	PING component of DYNA3D. Parallel I/O, MPI implementation
(Environmental) **Doug Rotman**	Global Atmospheric Chemistry	Effect of new Aircraft on Environment	**Boeing:** Steve Baughcum	Code: IMPACT stiff, coupled ODEs

(Physics) **Christian Mailhiot**	Advanced Materials	Materials design. Semi-conductors, metals, surfaces, thin films.	**Xerox:** J. Northrup C. Van der Walle	Code: PCGPP conjugate gradients

Author Biography

Alice E. Koniges is Leader, Multiprogrammatic and Insititutional Computing Research at Lawrence Livermore National Laboratory
Morris A. Jette is Group Leader of the Distributed Computing Technology Group at Lawrence Livermore National Laboratory.

Lawrence Livermore National Laboratory
For information about this page, please contact Alice Koniges, koniges@llnl.gov.
Disclaimers
Last modified May 23, 1997.
UCRL-MI-127382

Table of Contents | Author Index | CUG Home Page | Home