

# Early Evaluation of the Cray XT3 at ORNL

Center for Computational Science

**J. S. Vetter, S. R. Alam, T. H. Dunigan, Jr  
M. R. Fahey, P. C. Roth, P. H. Worley**

**Dozens of others...**

Oak Ridge National Laboratory  
Oak Ridge, TN, USA 37831

*EARLY EVALUATION: This paper contains preliminary results from our early delivery system, which is smaller in scale than the final delivery system and which uses early versions of the system software.*

OAK RIDGE NATIONAL LABORATORY  
U. S. DEPARTMENT OF ENERGY

UT-BATTELLE

## Highlights

- ORNL is installing a 25 TF, 5,200 processor Cray XT3
  - We currently have a 3,800 processor system
  - We are using CRMS 0406
- ***This is a snapshot!***
- We are evaluating its performance using
  - Microbenchmarks
  - Kernels
  - Applications
- The system is running applications at scale
- This summer, we expect to install the final system, scale up some applications, and continue evaluating its performance

OAK RIDGE NATIONAL LABORATORY  
U. S. DEPARTMENT OF ENERGY

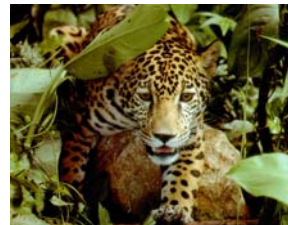
UT-BATTELLE

## Acknowledgments

- Cray
  - Supercomputing Center of Excellence
  - Jeff Beckleheimer, John Levesque, Luiz DeRose, Nathan Wichmann, and Jim Schwarzmeier
- Sandia National Lab
  - Ongoing collaboration
- Pittsburgh Supercomputing Center
  - Exchanged early access accounts to promote cross testing and experiences
- ORNL staff
  - Don Maxwell
- This research was sponsored by the Office of Mathematical, Information, and Computational Sciences, Office of Science, U.S. Department of Energy under Contract No. DE-AC05-00OR22725 with UT-Batelle, LLC. Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

## Cray XT3 – Jaguar Current System Configuration

- 40 cabinets
- 3,784 compute processors
- 46 service and I/O processors
- 2 GB memory per processor
- Topology (X,Y,Z): 10 torus, 16 mesh, 24 torus
- CRMS software stack



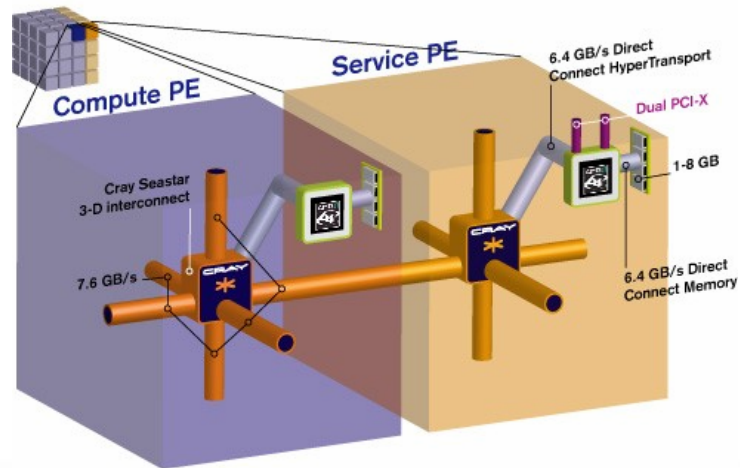
## Cray XT3 – Jaguar Final System Configuration – ETA June

- 56 cabinets
- 5,212 compute processors (25 TF)
- 82 service and I/O processors
- 2 GB memory per processor
- 10.7 TB aggregate memory
- 120 TB disk space in Lustre file system
- Topology (X,Y,Z): 14 torus, 16 mesh/torus, 24 torus

## Cray XT3 System Overview

- Cray's third generation MPP
  - Cray T3D, T3E
- Key features build on previous design philosophy
  - Single processor per node
    - Commodity processor: AMD Opteron
  - Customized interconnect
    - SeaStar ASIC
    - 3-D mesh/torus
  - Lightweight operating system – catamount – on compute PEs
  - Linux on service and IO PEs

## Cray XT3 PE Design



OAK RIDGE NATIONAL LABORATORY  
U. S. DEPARTMENT OF ENERGY

Image courtesy of Cray.

UT-BATTELLE

7

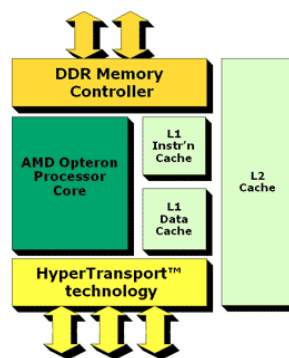
## AMD Opteron

### ➤ AMD Opteron Model 150

- Processor core / 2.4 Ghz
  - three integer units
  - one floating-point unit which is capable of two floating-point operations per cycle
  - 4.8 GFLOPS
- Integrated memory controller
- Three 16b 800 Mhz HyperTransport (HT) links
- L1 cache: 64KB I and D caches
- L2 cache: 1MB Unified

### ➤ Model 150 has three HT links but none support coherent HT (for SMPs)

- Lower latency to main memory than SMP capable processors



OAK RIDGE NATIONAL LABORATORY  
U. S. DEPARTMENT OF ENERGY

Image courtesy of AMD.

UT-BATTELLE

8

## Cray SeaStar Interconnect ASIC

- Routing and communications ASIC
- Connects to the Opteron via 6.4 GBps HT link
- Connects to six neighbors via 7.6 GBps links
- Topologies include torus, mesh
- Contains
  - PowerPC 440 chip, DMA engine, service port, router
- Notice
  - No PCI bus in transfer path
  - Interconnect Link BW is greater than Opteron Link BW
  - Carries all message traffic in addition to IO traffic

## Software

- Operating systems
  - Catamount
    - Lightweight kernel w/ limited functionality to improve reliability, performance, etc.
  - Linux
- Portals Communication Library
- Scalable application launch using Yod
- Programming environments
  - Apprentice, PAT, PAPI, mpiP
  - Totalview
- Filesystems
  - Scratch space through Yod
  - Lustre
- Math libraries
  - ACML 2.5, Goto library
- Details
  - CRMS 0406
  - 0413AA firmware
  - PIC 0x12

## Evaluations

### ➤ Goals

- Determine the most effective approaches for using the each system
- Evaluate benchmark and application performance, both in absolute terms and in comparison with other systems
- Predict scalability, both in terms of problem size and in number of processors

### ➤ We employ a hierarchical, staged, and open approach

- Hierarchical
  - Microbenchmarks
  - Kernels
  - Applications
- Interact with many others to get the best results
- Share those results



## Recent and Ongoing Evaluations

### ➤ Cray X1

- P.A. Agarwal, R.A. Alexander et al., "Cray X1 Evaluation Status Report," ORNL, Oak Ridge, TN, Technical Report ORNL/TM-2004/13, 2004.
- T.H. Dunigan, Jr., M.R. Fahey et al., "Early Evaluation of the Cray X1," Proc. ACM/IEEE Conference High Performance Networking and Computing (SC03), 2003.
- T.H. Dunigan, Jr., J.S. Vetter et al., "Performance Evaluation of the Cray X1 Distributed Shared Memory Architecture," IEEE Micro, 25(1):30-40, 2005.

### ➤ SGI Altix

- T.H. Dunigan, Jr., J.S. Vetter, and P.H. Worley, "Performance Evaluation of the SGI Altix 3700," Proc. International Conf. Parallel Processing (ICPP), 2005.

### ➤ Cray XD1

- M.R. Fahey, S.R. Alam et al., "Early Evaluation of the Cray XD1," Proc. Cray User Group Meeting, 2005, pp. 12.

### ➤ SRC

- M.C. Smith, J.S. Vetter, and X. Liang, "Accelerating Scientific Applications with the SRC-6 Reconfigurable Computer: Methodologies and Analysis," Proc. Reconfigurable Architectures Workshop (RAW), 2005.

### ➤ Underway

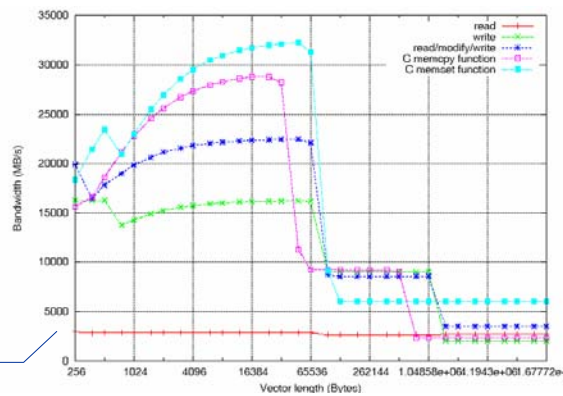
- XD1 FPGAs
- ClearSpeed
- EnLight
- Multicore processors
- IBM BlueGene/L
- IBM Cell

# Microbenchmarks

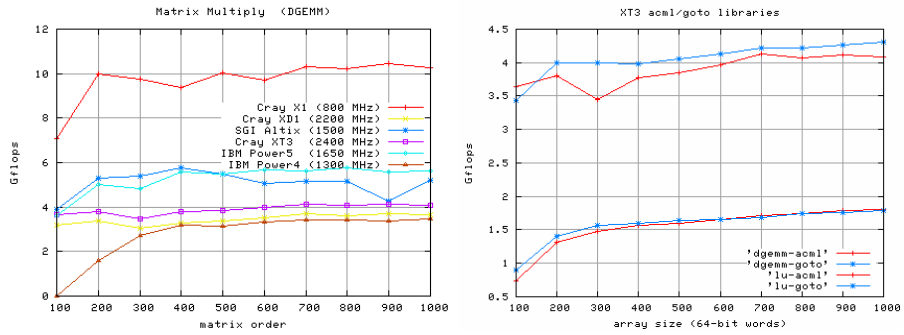
- Microbenchmarks characterize specific components of the architecture
  
- Microbenchmark suite tests
  - arithmetic performance,
  - memory-hierarchy performance,
  - task and thread performance,
  - message-passing performance,
  - system and I/O performance, and
  - parallel I/O

# Memory Performance

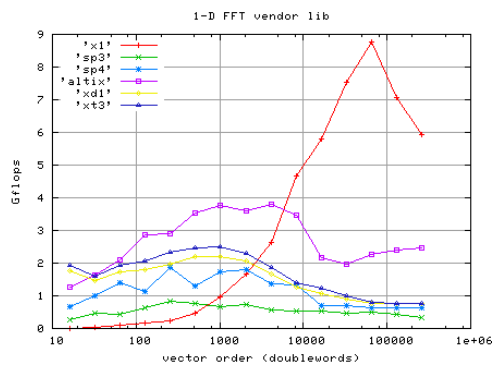
Platform	Measured Latency to Main Memory (ns)
Cray XT3 / Opteron 150 / 2.4	51.41
Cray XD1 / Opteron 248 / 2.2	86.51
IBM p690 / POWER4 / 1.3	90.57
Intel Xeon / 3.0	140.57



# DGEMM Performance

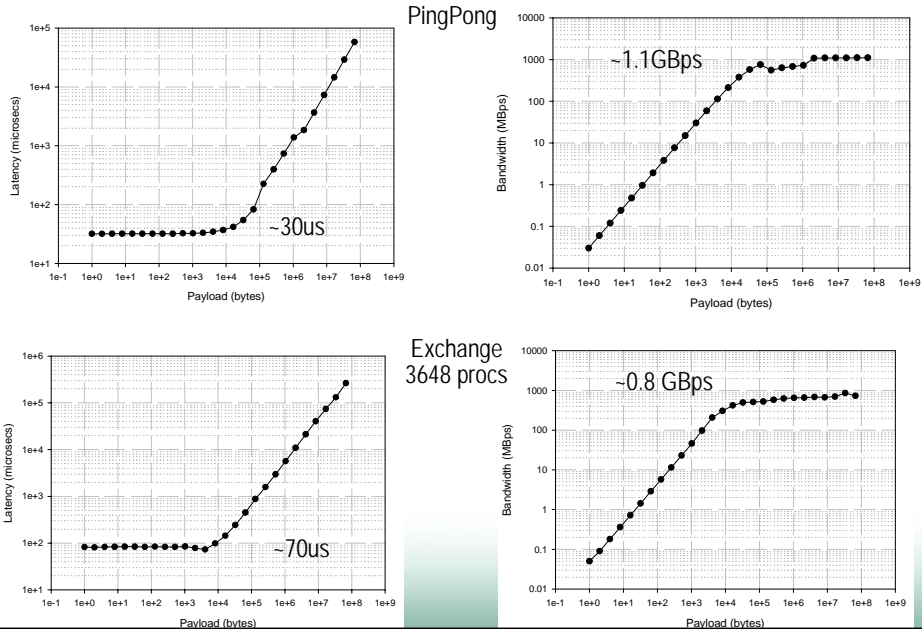


# FFT Performance



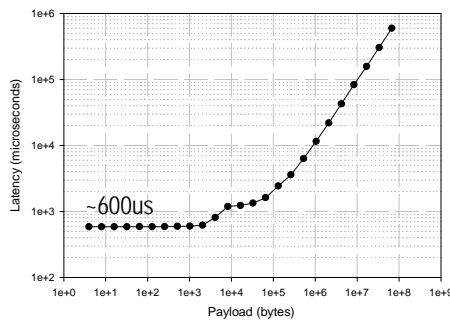


# Message Passing Performance



# Message Passing Performance (2)

➤ AllReduce across 3,648 processors



## Note

- *As mentioned earlier, we are using preliminary versions of the system software for these tests. We expect future versions of the software to improve both the latency and bandwidth of these MPI operations. In fact, other sites are reporting much improved MPI results.*

## HPC Challenge Benchmark

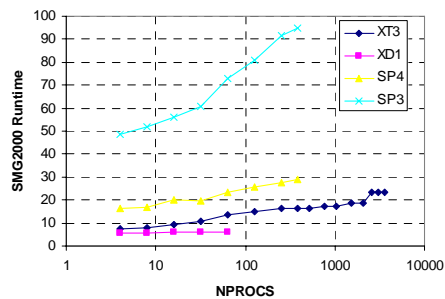
- <http://icl.cs.utk.edu/hpcc/>
- Version 0.8b
- 2,048 processors, 9.8 TFLOPS
  - HPL: 7.4 (9.8), 75%
  - MPI RandomAccess: 0.055 GUPS
- These are unofficial numbers – not recorded at the HPCC website
  - Stay tuned
  - More details from HPCC presentation, and HPCC website later this year.

## Kernels

- SMG2000
- PSTSWM (see paper)

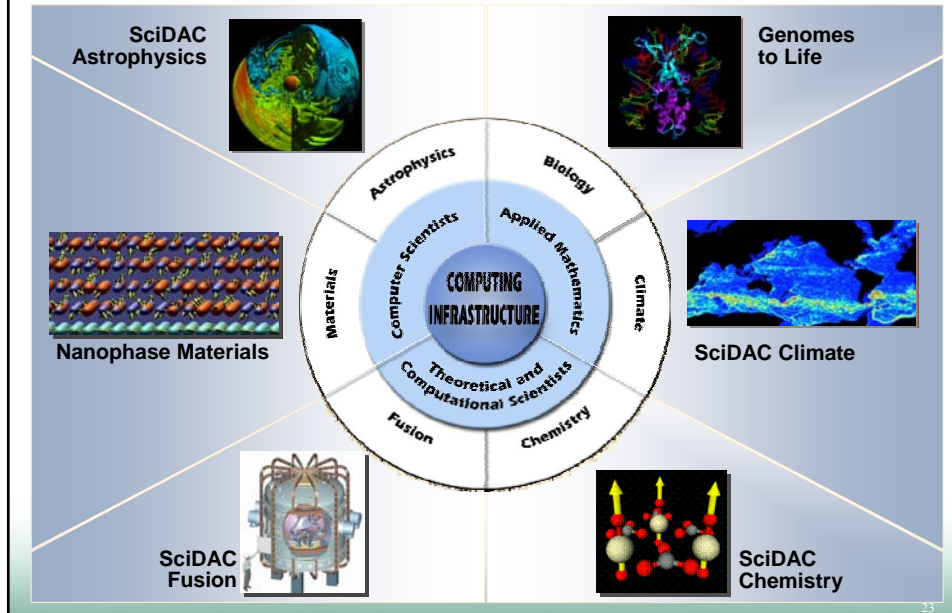
## SMG2000 / Multigrid solver

- SMG2000 is a driver for the linear solver Hypre
  - 7-point Laplacian on structured grid
- Hypre
  - Parallel semicoarsening multigrid solver for the linear systems arising from finite difference, finite volume, or finite element discretizations of the diffusion equation
- Benchmark includes both setup and solve of linear system
  - We only measure the solve phase



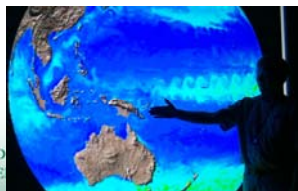
Lower is better.

## ORNL has Major Efforts Focusing on Grand Challenge Scientific Applications



## Climate Modeling

- Community Climate System Model (CCSM) is the primary model for global climate simulation in the USA
  - Community Atmosphere Model (CAM)
  - Community Land Model (CLM)
  - Parallel Ocean Program (POP)
  - Los Alamos Sea Ice Model (CICE)
  - Coupler (CPL)
- Running Intergovernmental Panel on Climate Change (IPCC) experiments



OAK RIDGE  
U. S. DEPARTMENT OF ENERGY

**Science and technology** The Economist December 1, 2014 87

**Also in this section**  
 86 Sorting sperm with optical tweezers  
 89 3D television

**Climate change**  
**A canary in the coal mine**

The Arctic seems to be getting warmer. So what?

**44** CLIMATE change in the Arctic is a reality now? It seems Robert Corell, an oceanographer with the American Meteorological Society, will send predictions on all too common when it comes to global warming, but in this case his optimism seems well founded.

Dr Corell heads a team of some 200 scientists who have spent the past four years investigating the matter in a process known as the Arctic Climate Impact Assessment (ACIA). The group, drawn from 18 countries with territories inside the Arctic Circle, has just issued a report called "Impacts of a Warming Arctic", a lengthy summary of the principal scientific findings. A second report, which will strictly not recommend policies, is due out in a few weeks. A third, for better or worse depending on the scientific findings, will not come out for some months yet.

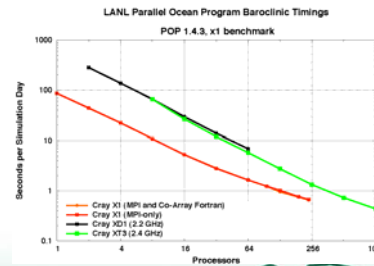
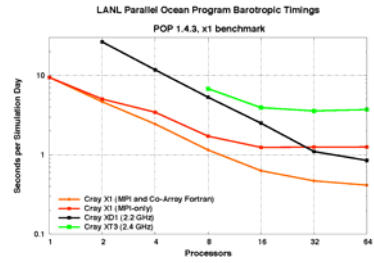
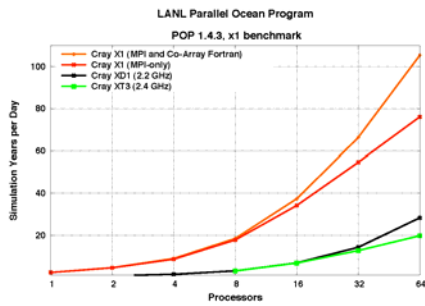
Already, though, the ACIA has made a splash. One reason is the inevitable warning over policy recommendations. News reports have suggested that the Bush administration has tried to ignore part of support in the second, as yet unpublished, report for the use of Kyoto protocol for other mandatory policies for the control of greenhouse emissions. But even so, your made headlines by predicting a rise in sea level of between seven and 10 centimetres, and a temperature rise of between 1.4°C and 5.8°C over this century. However, its authors did not find confidence in predicting other rapid polar warming or the speedy demise of the Greenland ice sheet. Putting its evidence gathered since the 1970s report, this week's report says you should be alarmed.

**Hot on top**  
 The ACIA reduces that in recent decades average temperatures have increased almost twice as fast in the Arctic as they have in most of the world. Scientists agree that there are plenty, such as the high latitudes of the Greenland ice sheet and some bays of sea, where temperatures seem to rise more rapidly. On the other hand, there are also places, such as parts of Alaska, where they have risen far faster than average. Robert Bell, a geophysicist at California University who was not involved in the report's compilation, believes that each of the international, interdisciplinary components of this week's report, as far as progress of the world's weather, can be partly and you can get local on how you look at the whole picture.

And there is other evidence of warming to bolster the ACIA's view. For example, the report documents the widespread melting of glaciers and of sea ice, a trend already making life miserable for the polar bears and seals that depend on that ice. It also notes a thawing of the more southern. The most worrying finding, however, is evidence that indicates that the

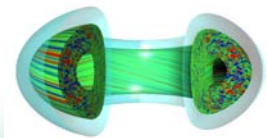
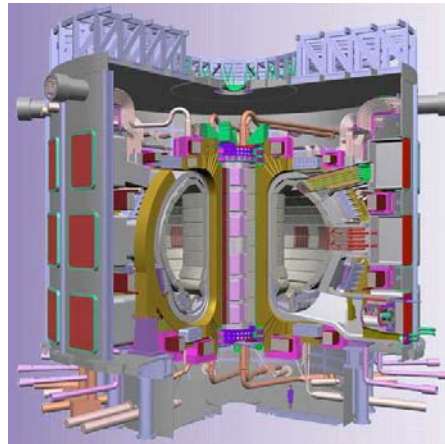
UT-BATTELLE

# Climate / Parallel Ocean Program / POP

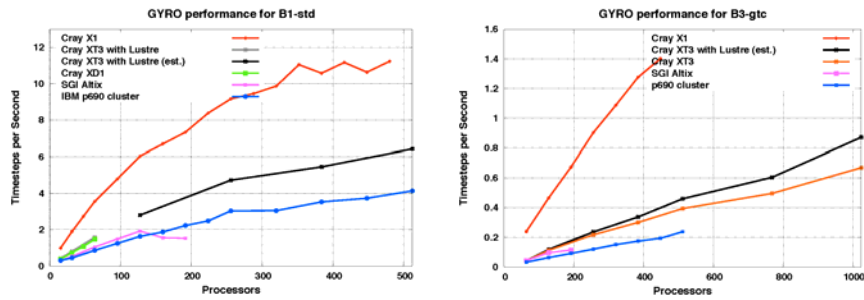


# Fusion

- Advances in understanding tokamak plasma behavior are necessary for the design of large scale reactor devices (like ITER)
- Multiple applications used to simulate various phenomena w/ different algorithms
  - GYRO
  - NIMROD
  - AORSA3D
  - GTC

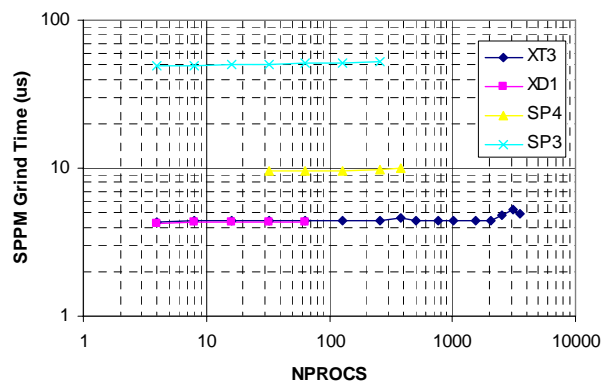


# Fusion / GYRO



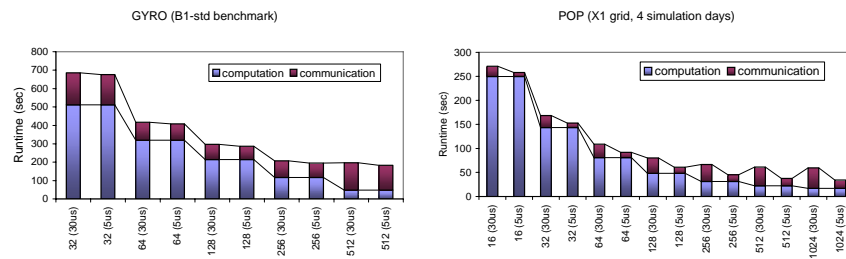
# SPPM

➤ 3-D Gas Dynamics Problem on uniform Cartesian mesh, using a simplified version of the Piecewise Parabolic Method



# Modeling Performance Sensitivities

- Estimate change in performance due to lower latencies



## Next steps

- Eager to work on full system
  - 25 TF, 5,200 processor Cray XT3 this summer
- Preliminary evaluation results
  - Need to continue performance optimizations of system software and applications
  - GYRO and POP are running well – Strong scaling
  - SMG and sPPM are running across the system well – Weak scaling
  - Installing parallel filesystem
- Actively porting many other applications

## Other related presentations

- Comparative Analysis of Interprocess Communication on the X1, XD1, and XT3, Worley
- Cray and HPCC: Benchmark Developments and Results from the Past Year, Wichmann
- Porting and Performance of the Community Climate System Model (CCSM3) on the Cray X1, Carr
- Optimization of the PETSc Toolkit and Application Codes on the Cray X1, Mills
- GYRO Performance on a Variety of MPP Systems, Fahey
- Early Evaluation of the Cray XD1, Fahey
- Towards Petacomputing in Nanotechnology, Wang
- System Integration Experience Across the Cray Product Line, Bland
- High-Speed Networking with Cray Supercomputers at ORNL, Carter

## The Details

- We are running
  - CRMS 0406
  - 0413AA firmware
  - PIC 0x12

```
$ module list
Currently Loaded Modulefiles:
 1) acml/2.5                6) xt-mpt/1.0             11) xt-boot/1.0
 2) pgi/5.2.4              7) xt-service/1.0        12) xt-pbs/1.4
 3) totalview/6.8.0-0     8) xt-libc/1.0           13) xt-crms/1.0
 4) xt-pe/1.0              9) xt-os/1.0             14) PrgEnv/1.0
 5) xt-libsci/1.0         10) xt-catamount/1.15
```