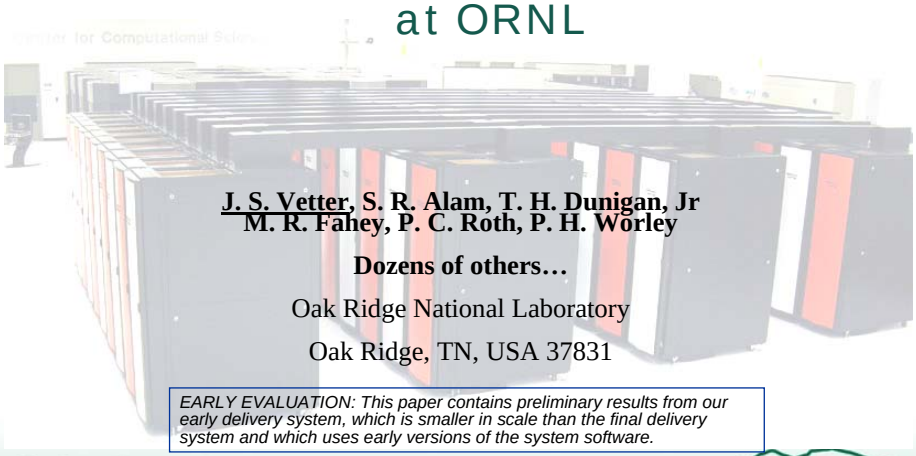
Early Evaluation of the Cray XT3 at ORNL



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.

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

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

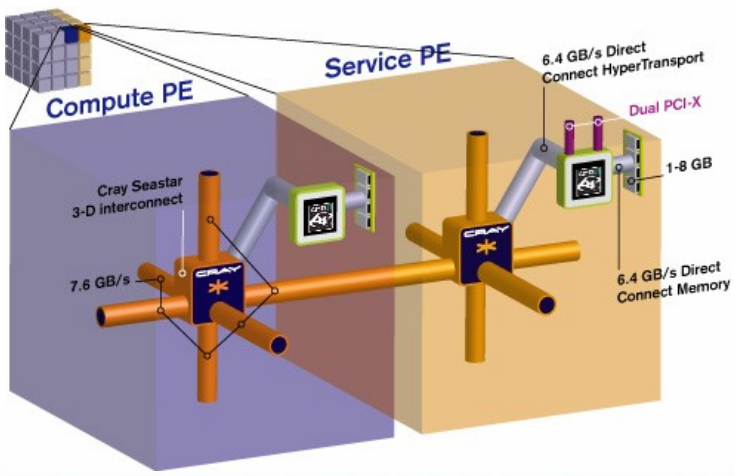
Cray XT3 – Jaguar
Final System Configuration – ETA June

- 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

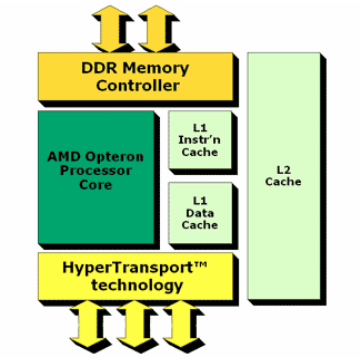
- Cray's third generation MPP
  - Cray T3D, T3E
- Key features build on previous design philosophy
  - Single processor per node
    - Commodity processor: AMD Opteron
  - Customized interconnect
    - SeaStar ASIC
    - 3-D mesh/torus
  - Lightweight operating system – catamount – on compute PEs
  -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

Cray XT3 PE Design



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



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

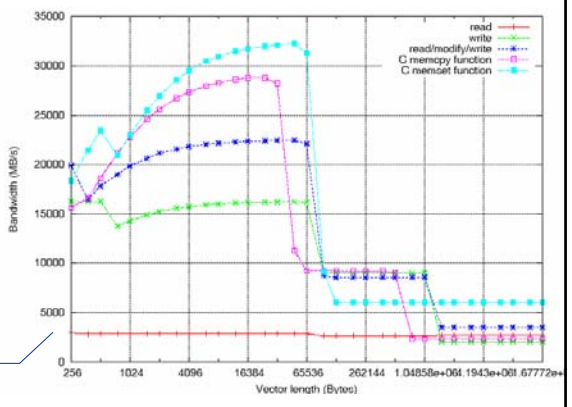
- Cray X1
  - P.A. Agarwal, R.A. Alexander et al., "Cray X1 Evaluation Status Report," ORNL, Oak Ridge, TN, Technical Report ORNL/TM-2004/13, 2004.
  - T.H. Dunigan, Jr., M.R. Fahey et al., "Early Evaluation of the Cray X1," Proc. ACM/IEEE Conference High Performance Networking and Computing (SC03), 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

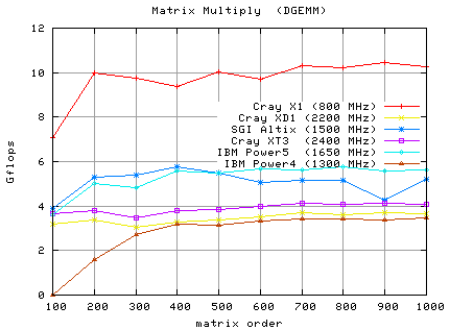
Memory Performance

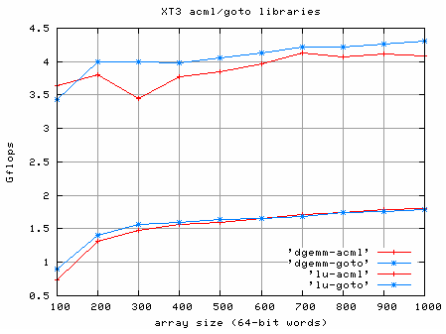
Platform	Measured Latency to Main Memory (ns)
Cray XT3 / Opteron 150 / 2.4	51.41
Cray XD1 / Opteron 248 / 2.2	86.51
IBM p690 / POWER4 / 1.3	90.57
Intel Xeon / 3.0	140.57



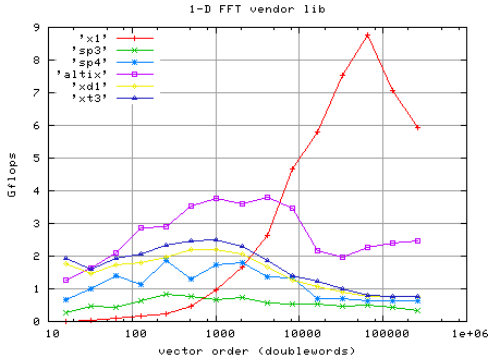
Compiler Issue uncovered

DGEMM Performance



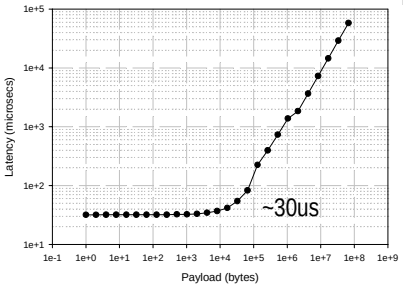


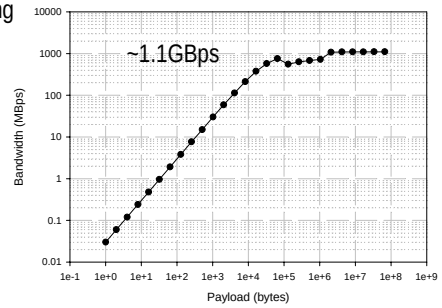
FFT Performance



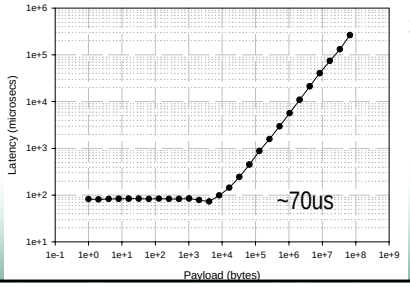
Message Passing Performance

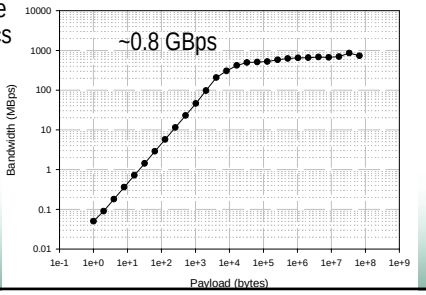
PingPong





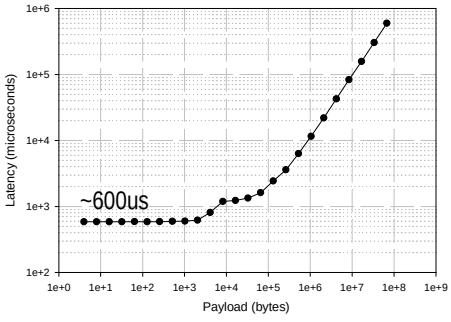
Exchange
3648 procs





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

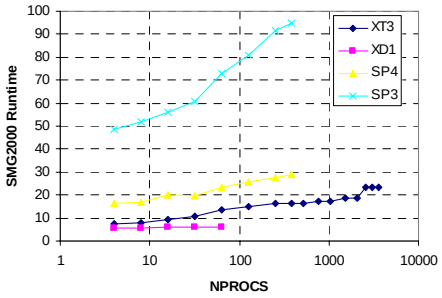
- http://icl.cs.utk.edu/hpcc/
- Version 0.8b
- 2,048 processors, 9.8 TFLOPS
  - HPL: 7.4 (9.8), 75%
  - MPI RandomAccess: 0.055 GUPS
- These are unofficial numbers – not recorded at the HPCC website
  - Stay tuned
  - More details from HPCC presentation, and HPCC website later this year.

Kernels

- SMG2000
- PSTSWM (see paper)

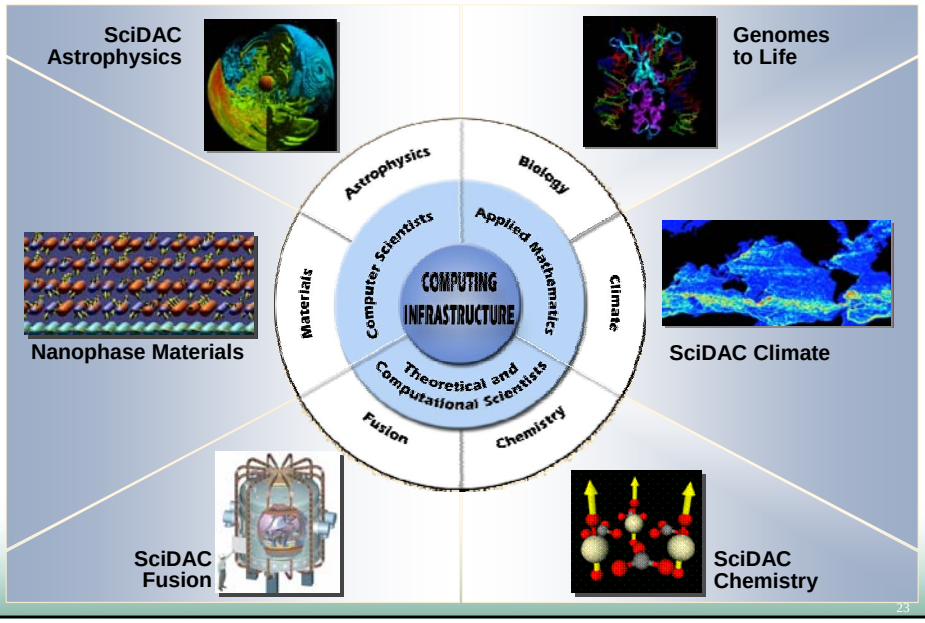
SMG2000 / Multigrid solver

- 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



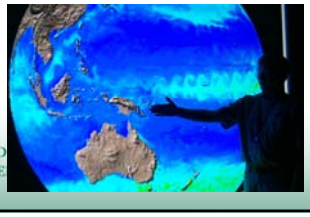
Lower is better.

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



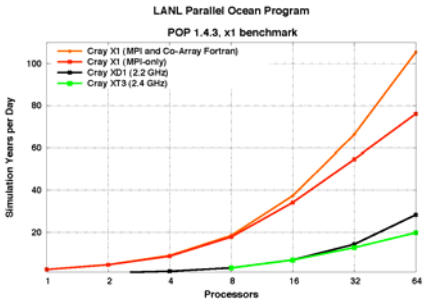
Science and technology

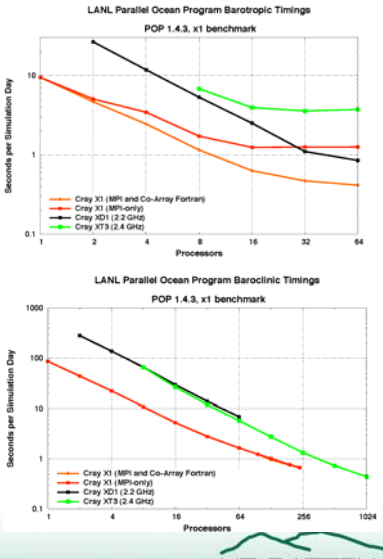
A canary in the coal mine

OAK RIDGE NATIONAL LABORATORY
U. S. DEPARTMENT OF ENERGY

UT-BATTELLE

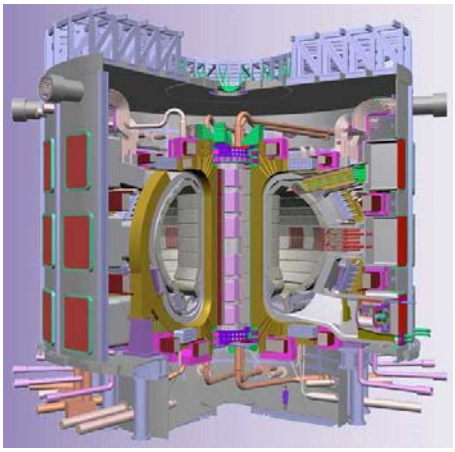
Climate / Parallel Ocean Program / POP

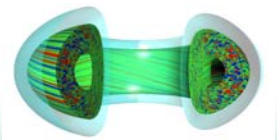




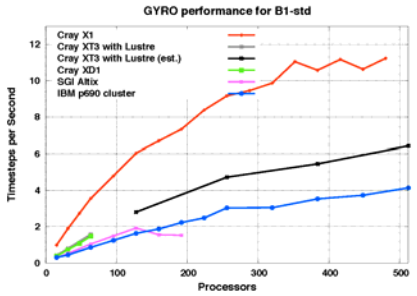
Fusion

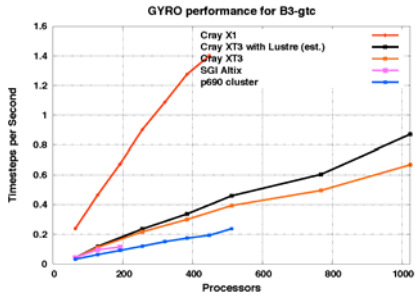
- Advances in understanding tokamak plasma behavior are necessary for the design of large scale reactor devices (like ITER)
- Multiple applications used to simulate various phenomena w/ different algorithms
  - GYRO
  - NIMROD
  - AORSA3D
  - GTC





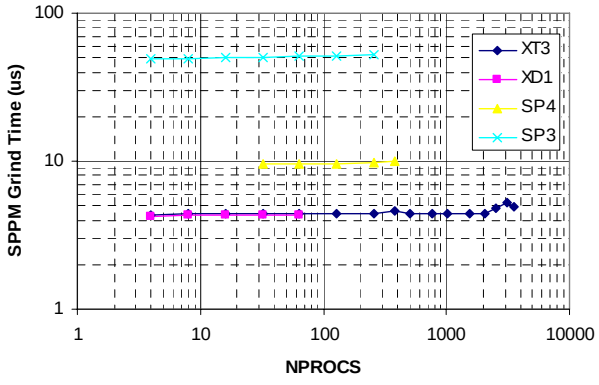
Fusion / GYRO





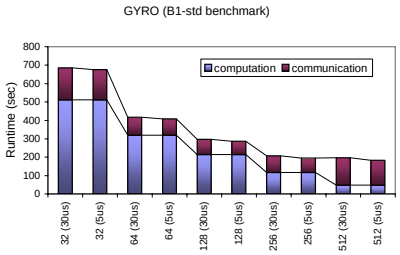
sPPM

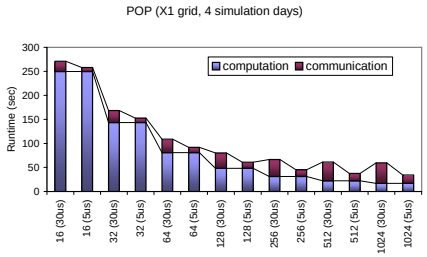
- 3-D Gas Dynamics Problem on uniform Cartesian mesh, using a simplified version of the Piecewise Parabolic Method



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
```