Evaluation of the Cray XT3: A Status Report

Cray User Group 2006, Lugano, Switzerland
May 9, 2006

Sadaf R. Alam, Richard F. Barrett, Mark R. Fahey,
O.E. Bronson Messer, Richard T. Mills, Philip C. Roth,
Jeffrey S. Vetter, and Patrick H. Worley

Oak Ridge National Laboratory
Oak Ridge, TN 37831

http://www.csm.ornl.gov/ft
http://www.csm.ornl.gov/evaluation
http://www.nccs.gov

Highlights

- We evaluated the system using micro-benchmarks, kernels, and applications from important DOE areas.

- The Cray XT3 is a well-balanced platform and it is demonstrating strong performance for diverse DOE application workloads

Architecture Evaluation Project at ORNL

- Evaluation goals
  - Determine most effective approaches for using a system
  - Evaluate benchmark and application performance, both absolute and relative to other systems
  - Predict scalability, both processor counts and problem size
  - Focus on DOE apps, esp. Office of Science
- Hierarchical, staged approach
  - Microbenchmarks
  - Kernels
  - Applications from important DOE application areas
- Recent examples:
  - Cray X1(E) and XD1, SGI Altix 3700, IBM p690 w/ HPS

XT3 at ORNL: Jaguar

- 5294 nodes
  - 5212 compute nodes
  - 82 I/O and login nodes
- Each node
  - 2.4 GHz Opteron
  - 2GB RAM
- 14x16x24 topology
  - torus in first and third dimensions when data was collected
- 56 cabinets

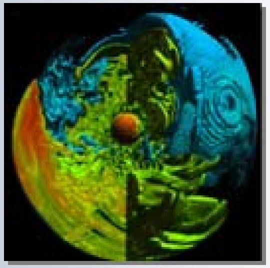
The Cray XT3: Software

- Catamount lightweight kernel on compute nodes
  - *Single process (single thread)*
  - *No demand-paged virtual memory*
  - *POSIX-like system call interface*
  - Linux on login and I/O nodes
- Portals data movement layer
  - *Connectionless, reliable, in-order delivery*
  - *One-sided and two-sided communication models*
  - *MPI is implemented on Portals*
- AMD Core Math Library (ACML)
- Lustre parallel file system

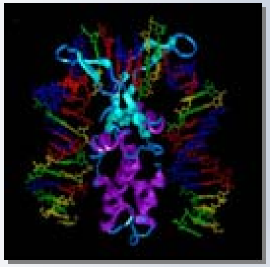
Application Performance

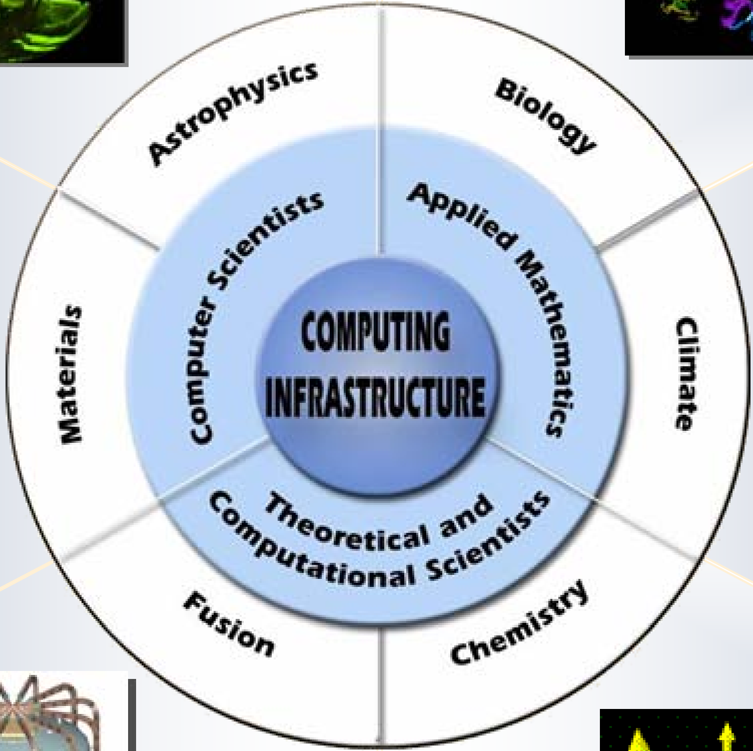
Applications at DOE and ORNL

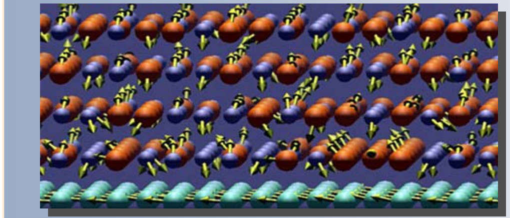
SciDAC Astrophysics



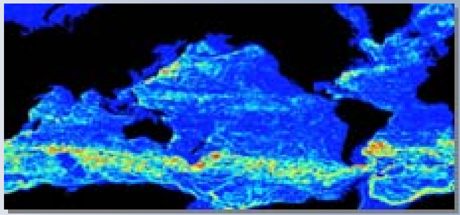
Genomes to Life



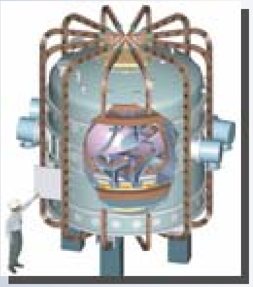




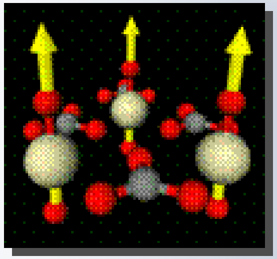
Nanophase Materials



SciDAC Climate



SciDAC Fusion

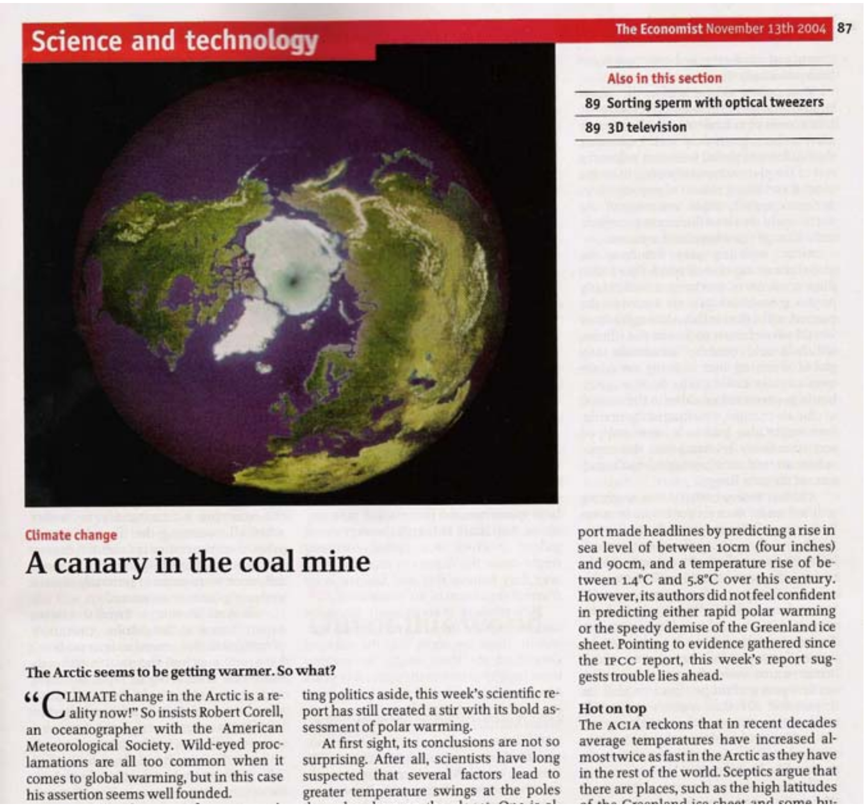


SciDAC Chemistry

Climate Modeling

- Community Climate System Model (CCSM) is the primary model for global climate simulation in the USA

- Running Intergovernmental Panel on Climate Change (IPCC) experiments



Science and technology

The Economist November 13th 2004 87

Also in this section

89 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?

"CLIMATE change in the Arctic is a reality now!" So insists Robert Corell, an oceanographer with the American Meteorological Society. Wild-eyed proclamations are all too common when it comes to global warming, but in this case his assertion seems well founded.

ting politics aside, this week's scientific report has still created a stir with its bold assessment of polar warming.

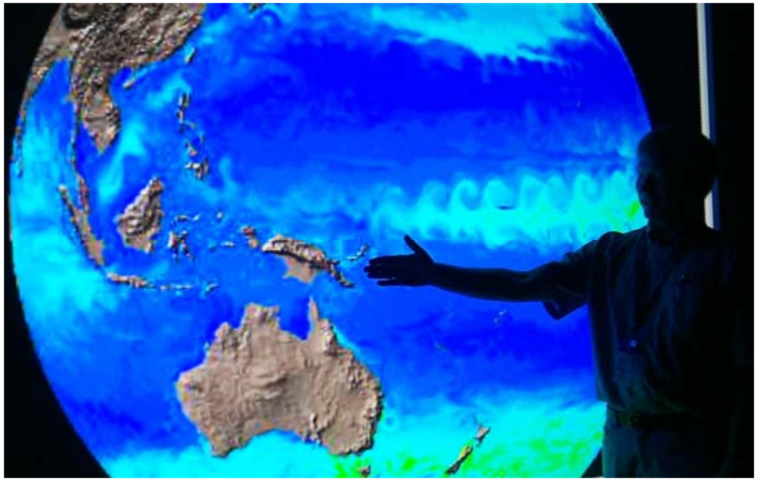
At first sight, its conclusions are not so surprising. After all, scientists have long suspected that several factors lead to greater temperature swings at the poles

port made headlines by predicting a rise in sea level of between 10cm (four inches) and 90cm, and a temperature rise of between 1.4°C and 5.8°C over this century. However, its authors did not feel confident in predicting either rapid polar warming or the speedy demise of the Greenland ice sheet. Pointing to evidence gathered since the IPCC report, this week's report suggests trouble lies ahead.

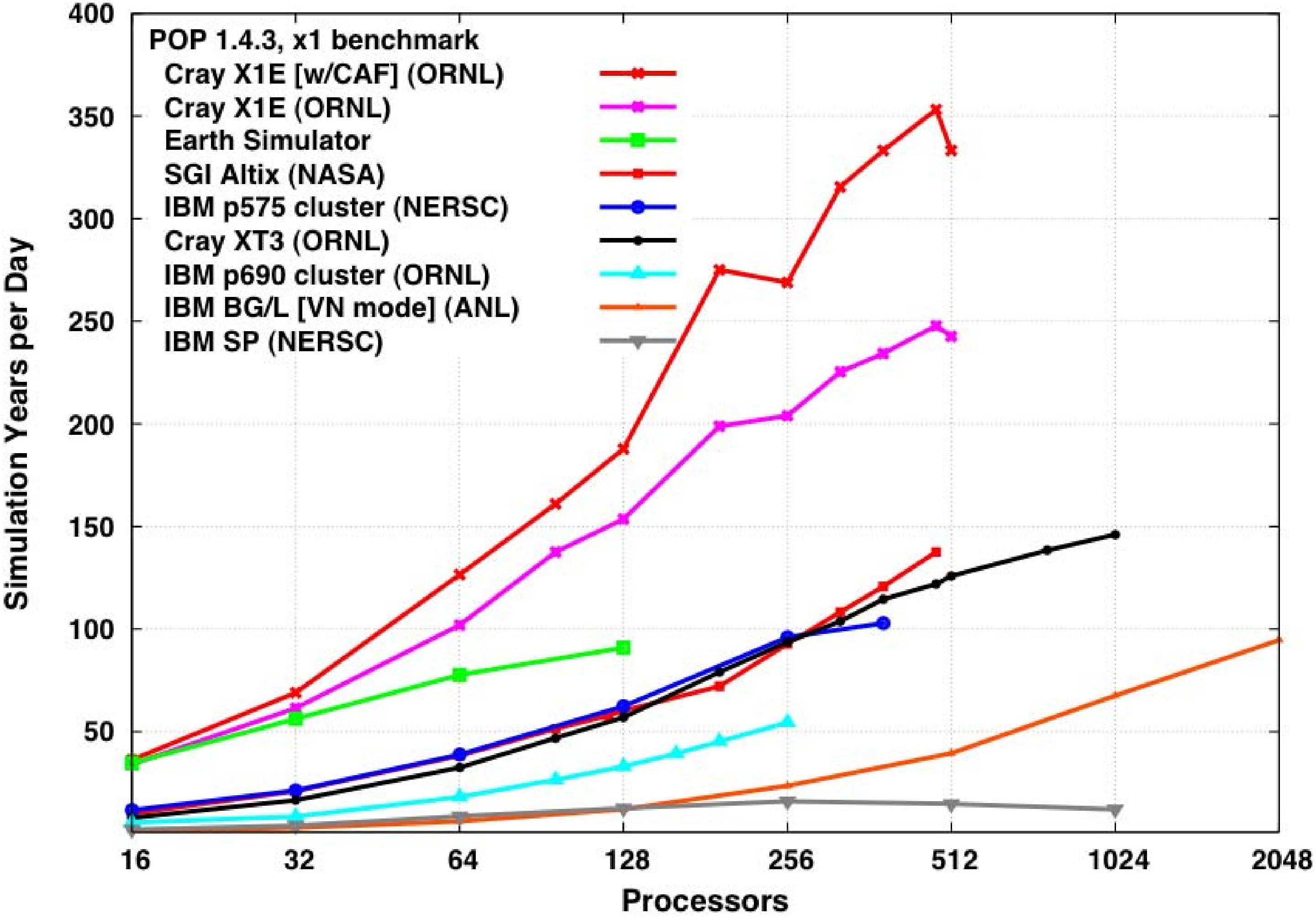
Hot on top

The ACIA reckons that in recent decades average temperatures have increased almost twice as fast in the Arctic as they have in the rest of the world. Sceptics argue that there are places, such as the high latitudes

Economist, 13 Nov 2004

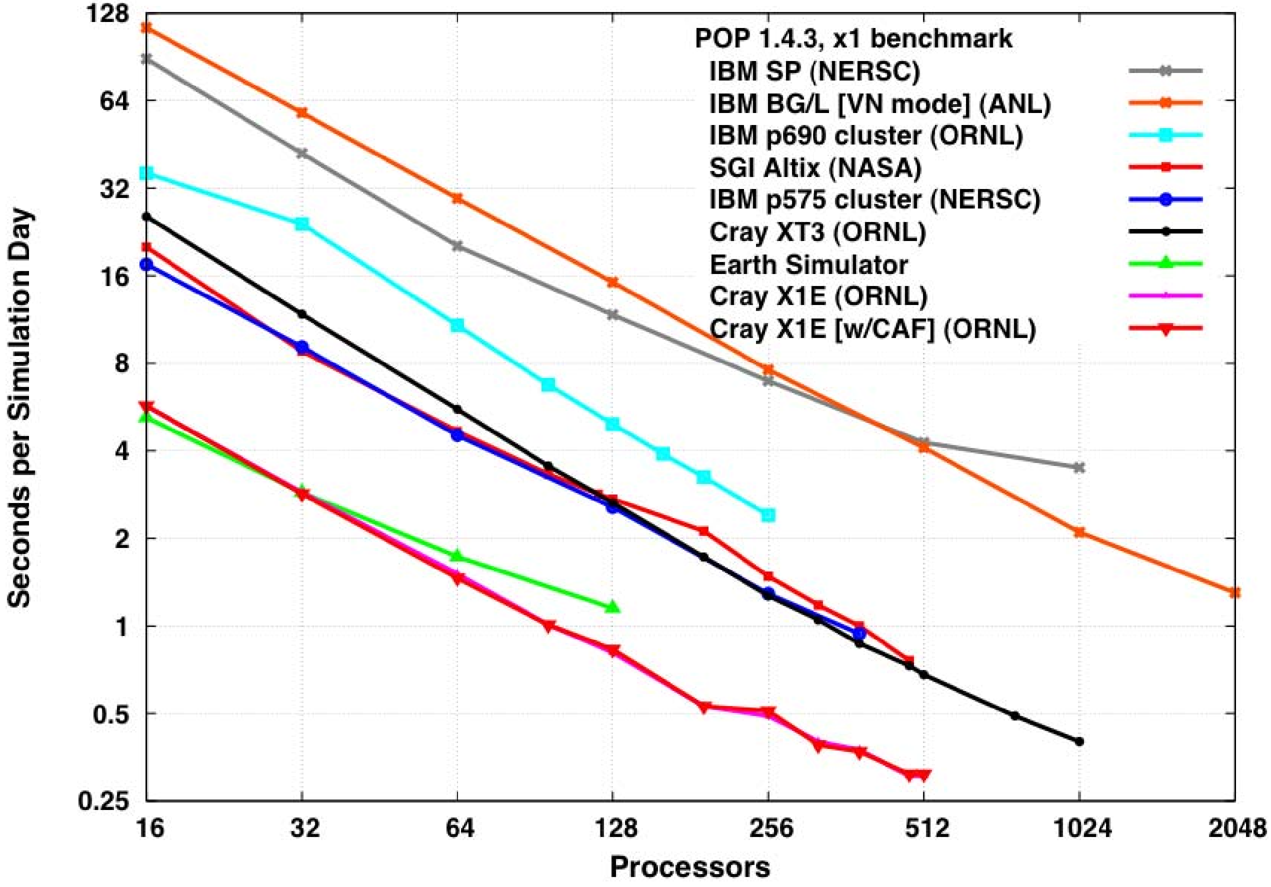


CCSM Ocean Model: POP



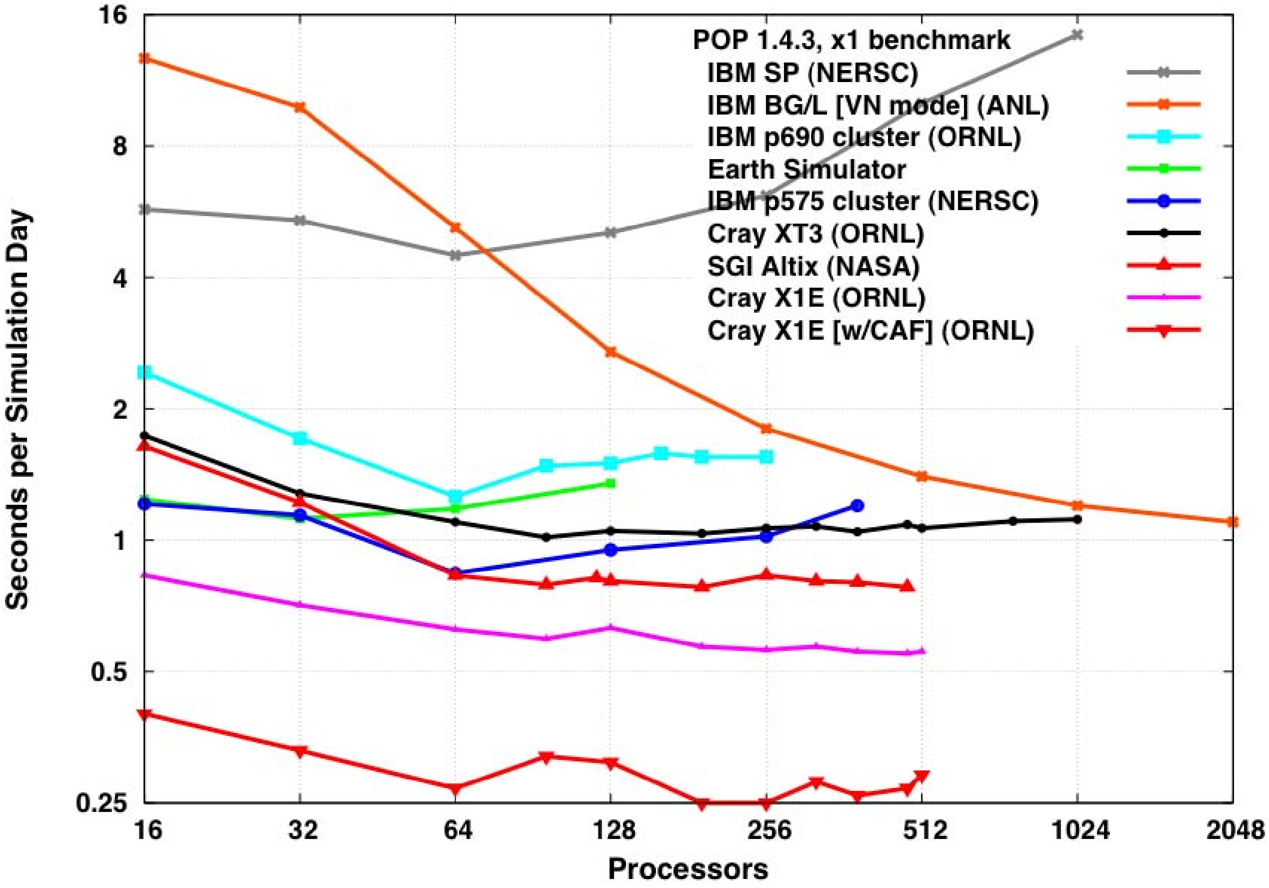
POP: Baroclinic phase

- Baroclinic phase
- Usually scales well
- XT3 performance similar to SGI Altix



POP: Barotropic phase

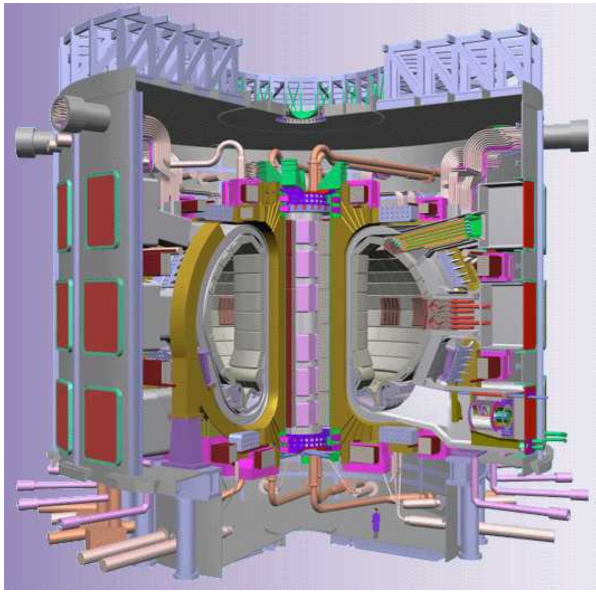
- Barotropic phase
- Usually scales poorly – it becomes latency bound
- Cost does not increase significantly for XT3

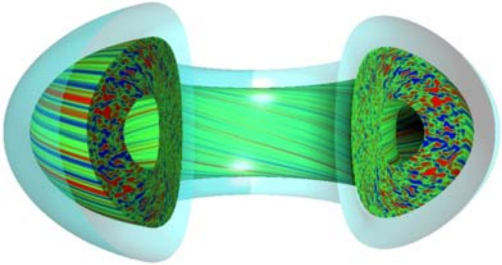


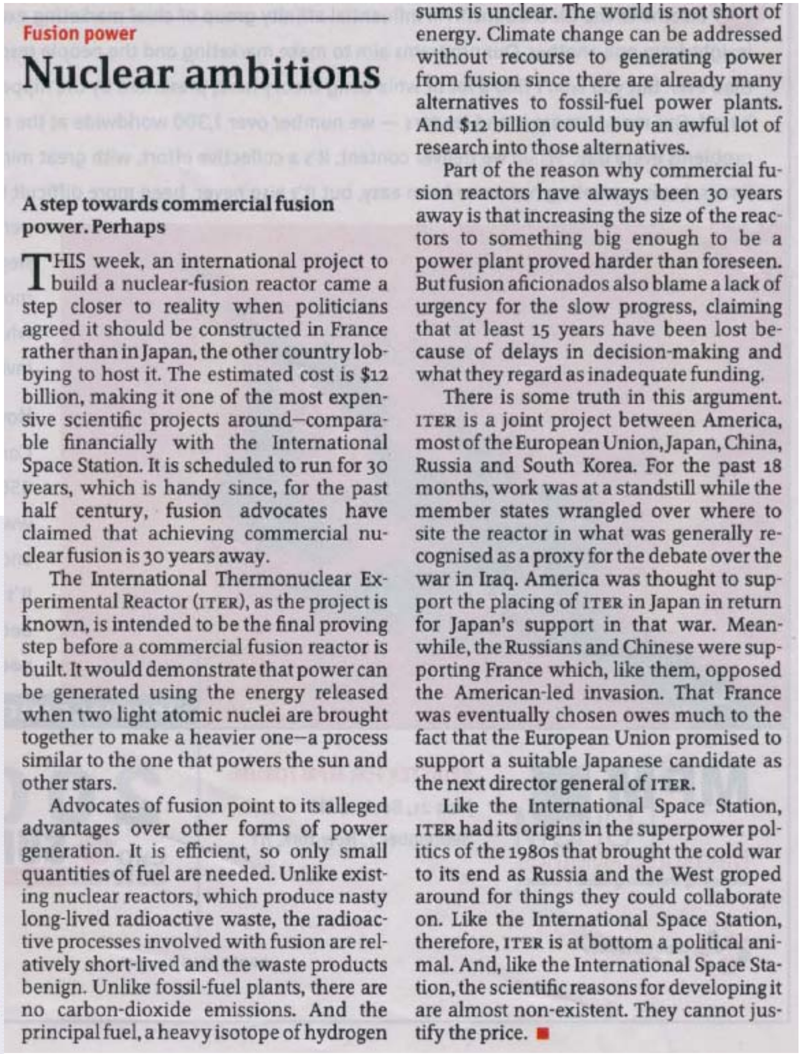
Fusion

- Modeling tokamak plasma behavior are necessary for the design of large scale reactor devices (like ITER)

- Multiple apps simulate various phenomena w/ different algorithms
  - GYRO
  - NIMROD
  - AORSA3D
  - GTC







Fusion power

Nuclear ambitions

A step towards commercial fusion power. Perhaps

THIS week, an international project to build a nuclear-fusion reactor came a step closer to reality when politicians agreed it should be constructed in France rather than in Japan, the other country lobbying to host it. The estimated cost is $12 billion, making it one of the most expensive scientific projects around—comparable financially with the International Space Station. It is scheduled to run for 30 years, which is handy since, for the past half century, fusion advocates have claimed that achieving commercial nuclear fusion is 30 years away.

The International Thermonuclear Experimental Reactor (ITER), as the project is known, is intended to be the final proving step before a commercial fusion reactor is built. It would demonstrate that power can be generated using the energy released when two light atomic nuclei are brought together to make a heavier one—a process similar to the one that powers the sun and other stars.

Advocates of fusion point to its alleged advantages over other forms of power generation. It is efficient, so only small quantities of fuel are needed. Unlike existing nuclear reactors, which produce nasty long-lived radioactive waste, the radioactive processes involved with fusion are relatively short-lived and the waste products benign. Unlike fossil-fuel plants, there are no carbon-dioxide emissions. And the principal fuel, a heavy isotope of hydrogen sums is unclear. The world is not short of energy. Climate change can be addressed without recourse to generating power from fusion since there are already many alternatives to fossil-fuel power plants. And $12 billion could buy an awful lot of research into those alternatives.

Part of the reason why commercial fusion reactors have always been 30 years away is that increasing the size of the reactors to something big enough to be a power plant proved harder than foreseen. But fusion aficionados also blame a lack of urgency for the slow progress, claiming that at least 15 years have been lost because of delays in decision-making and what they regard as inadequate funding.

There is some truth in this argument. ITER is a joint project between America, most of the European Union, Japan, China, Russia and South Korea. For the past 18 months, work was at a standstill while the member states wrangled over where to site the reactor in what was generally recognised as a proxy for the debate over the war in Iraq. America was thought to support the placing of ITER in Japan in return for Japan's support in that war. Meanwhile, the Russians and Chinese were supporting France which, like them, opposed the American-led invasion. That France was eventually chosen owes much to the fact that the European Union promised to support a suitable Japanese candidate as the next director general of ITER.

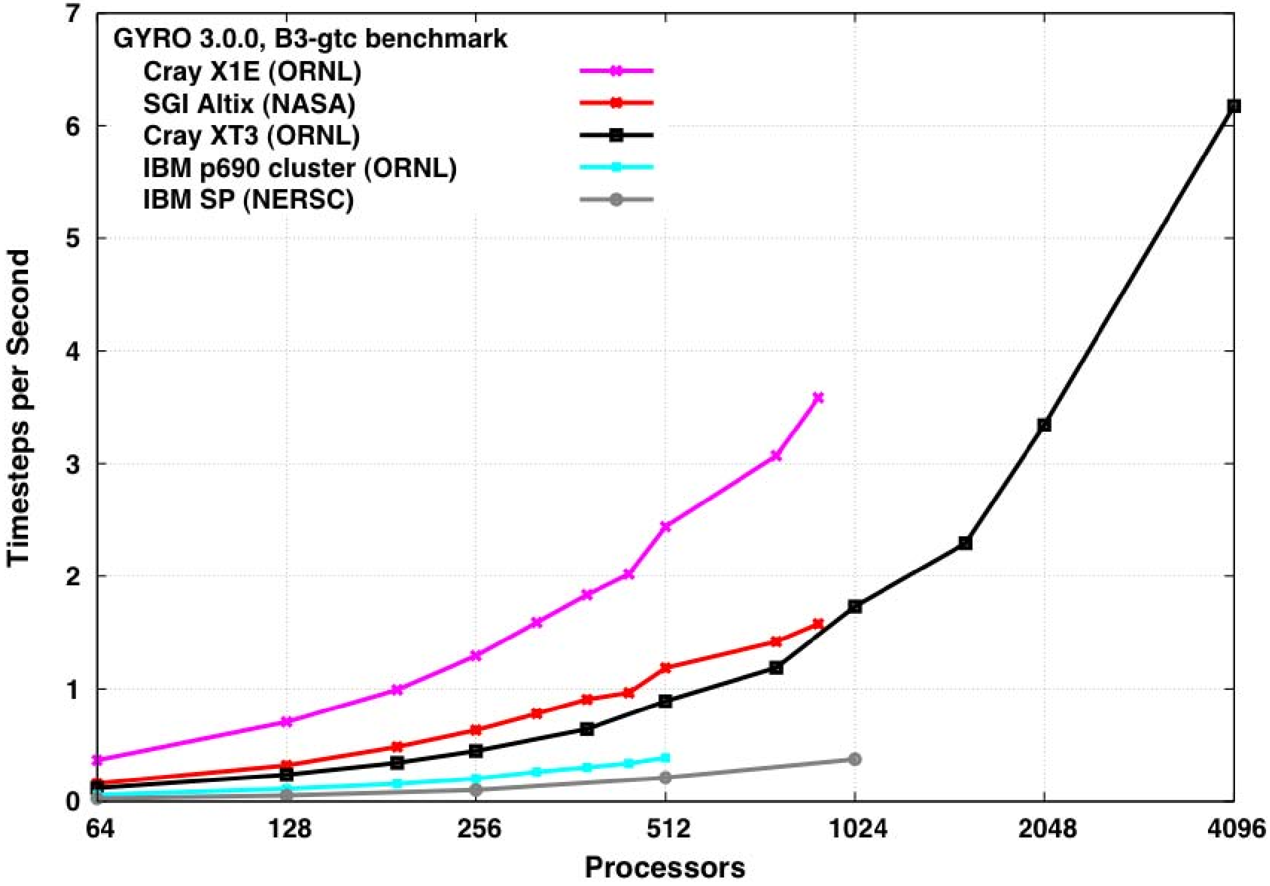
Like the International Space Station, ITER had its origins in the superpower politics of the 1980s that brought the cold war to its end as Russia and the West groped around for things they could collaborate on. Like the International Space Station, therefore, ITER is at bottom a political animal. And, like the International Space Station, the scientific reasons for developing it are almost non-existent. They cannot justify the price. ■

Economist, 2 July 2005

Fusion: GYRO

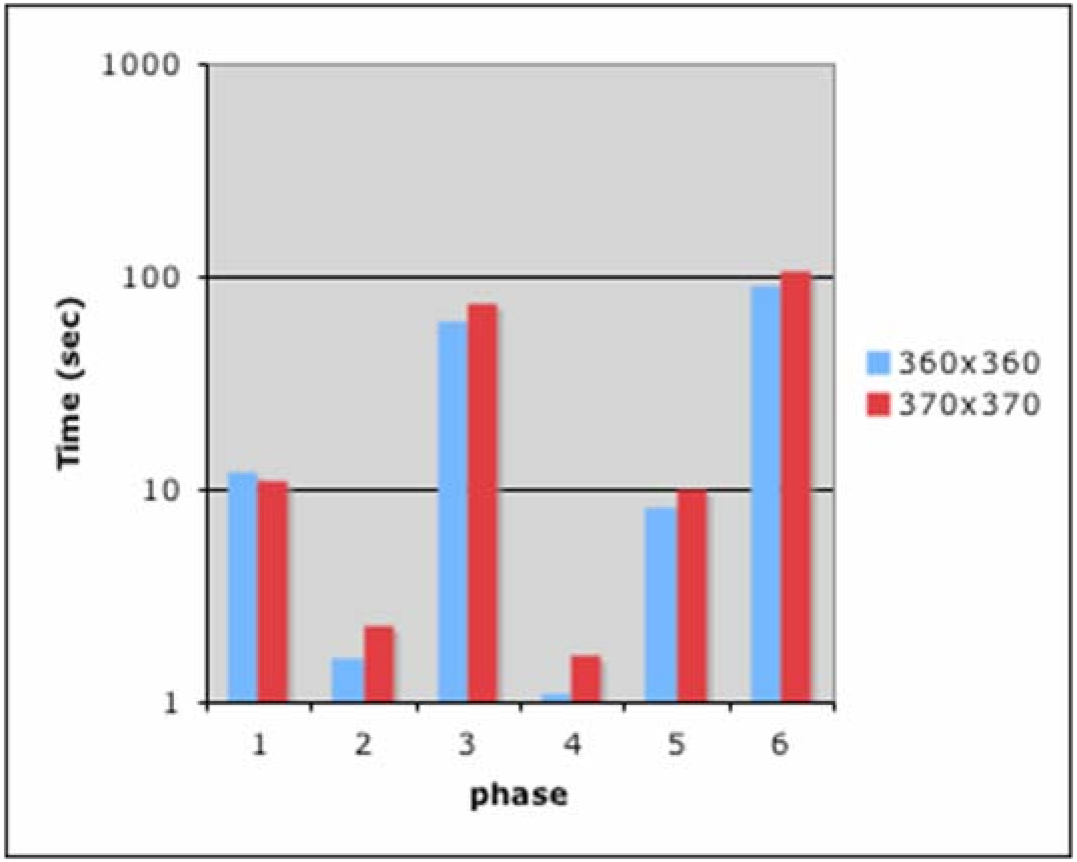
Micro-turbulence simulation

- Communication dominated by simultaneous all-to-alls over process sub-groups



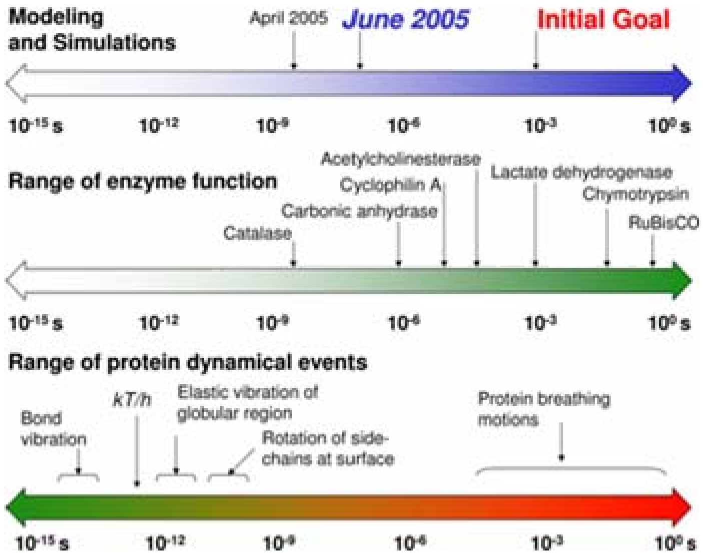
Fusion - AORSA<2D,3D>

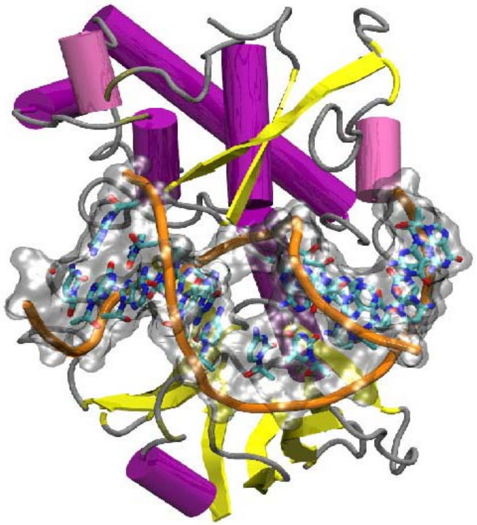
- Rf-heating of plasma in tokamak
  - All Orders Spectral Algorithm
- Time dominated by dense linear solver.
- SWIM



Computational Biology using Molecular Modeling

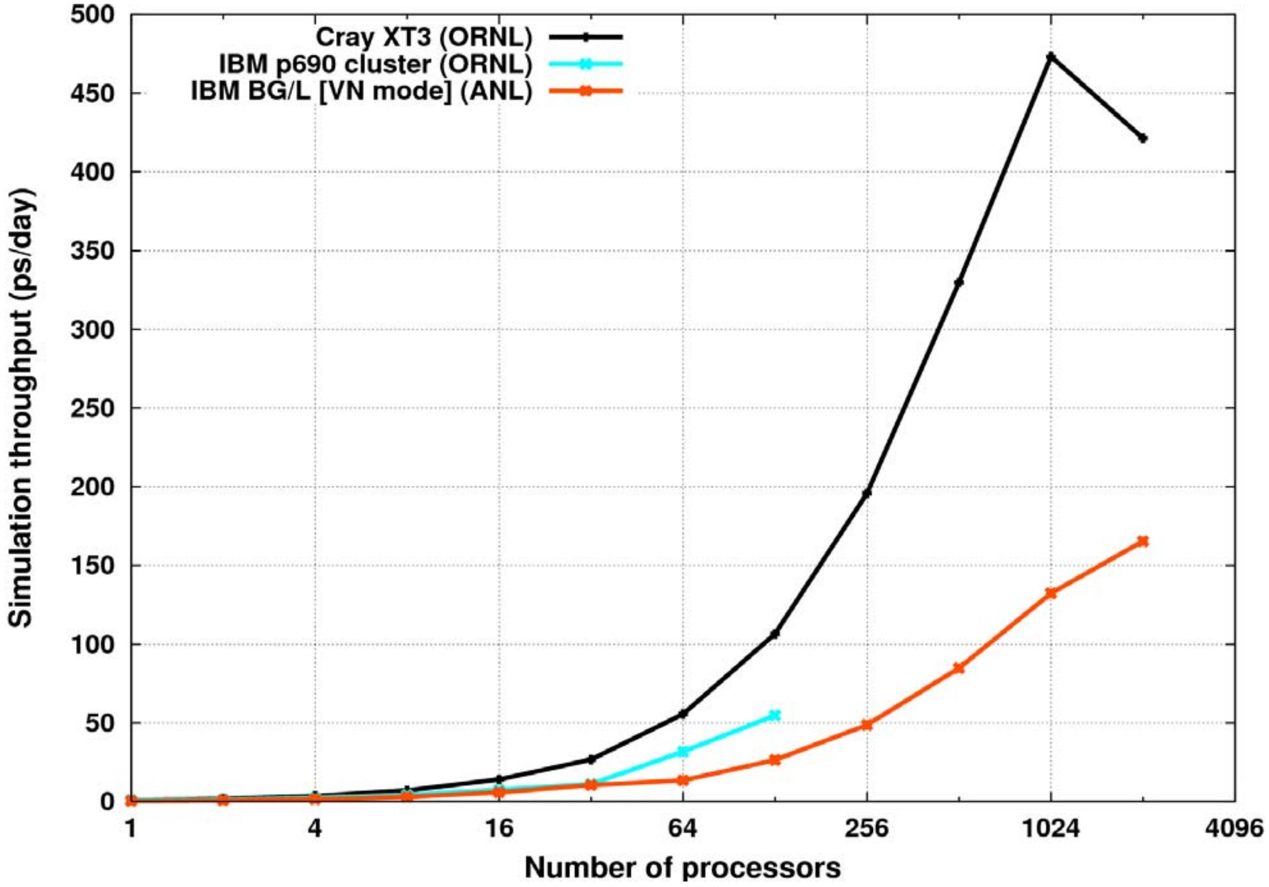
- The structure, dynamics and function of biomolecular complexes are inter-related
- Various aspects of biomolecules structure and function span multiple scales of time and length
- Wide community of biologist are interested in the multi-scale modeling of biomolecules
- *Multi-scale modeling of a real system may require 1 peta-flop/s for an entire year!*
- Scaling of existing software packages and algorithms is limited





AMBER

- Nearly 74K atoms
- Good XT3 throughput, but sharp knee at 1K processors



FAQ

- Do you consider TCO in your evaluation?
  - Yes, but we cannot share that information in this forum
- How do you know that your applications are fully optimized?
  - How does any developer know that he has finished optimizing an application? ☺
  - We work closely with the applications teams and systems developers to improve our probabilities
  - We are developing a performance modeling and analysis toolkit to help developers understand potential optimizations
- How did you select the applications?
  - We selected applications based on their role in DOE, availability, portability, and ability to represent computational characteristics
  - Other applications have demonstrated similar results.
  - Our workload is constantly evolving in contrast to some other agencies
- The Cray X1(E) appears to perform well on the applications you listed, so why aren't you pursuing that option for leadership computing?
  - Vector supercomputers can provide substantial performance benefits for some applications
  - Those applications that do not vectorize well, generally perform *very* poorly
  - In other words, the system has poor performance stability across a diverse range of applications

Summary

- DOE has deployed a 5,294-processor, 25 TFlop Cray XT3 at Oak Ridge National Laboratory
- We evaluated the system using micro-benchmarks, kernel, and applications from important DOE areas
  - Competitive scalar performance
  - Strong interconnect bandwidth with respect to other microprocessor-based systems
  - Good scaling behavior
- The Cray XT3 is a well-balanced platform and it is demonstrating strong performance for diverse DOE application workloads
- Continued work
  - Optimizations to software, system software
  - Additional applications
  - Preparation for 100T, 250T, 1PF systems

Acknowledgements

- 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-Battelle, 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, including Supercomputing Center of Excellence
  - Jeff Beckleheimer
  - John Levesque
  - Nathan Wichmann
  - Jim Schwarzmeier
  - Luiz DeRose

- ORNL
  - Don Maxwell
  - Jeff Kuehn

- Thanks to other organizations for machine access, benchmark results, and information exchange
  - Sandia National Laboratory
  - Pittsburgh Supercomputing Center
  - Lawrence Berkeley National Laboratory

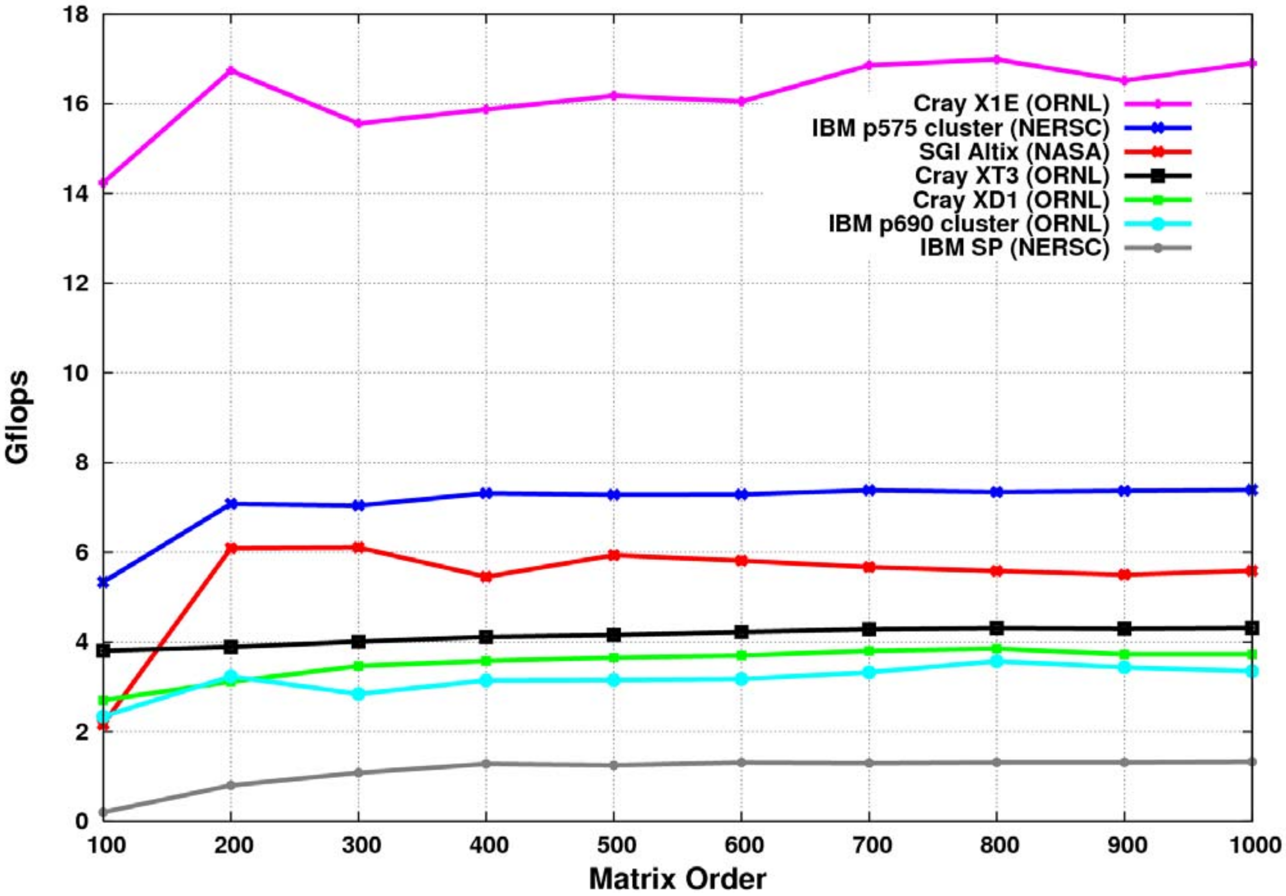
Compute Kernel Performance

(More results shown tomorrow at Jeff Kuehn's HPCC talk.)

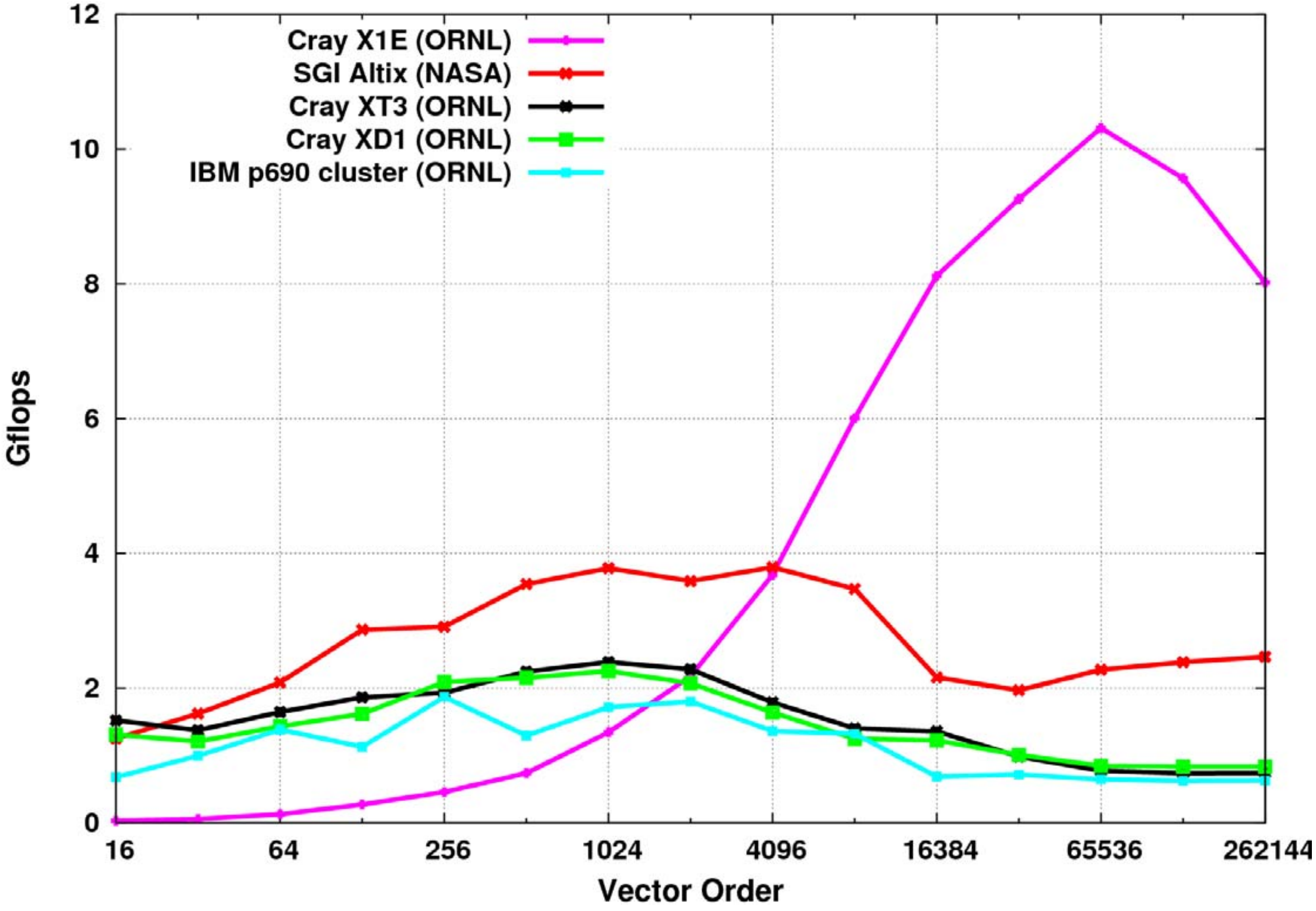
Matrix Multiply
Vendor DGEMM

4.4 GFLOPS

@matdim 1600;
91.6% peak



Vendor FFT

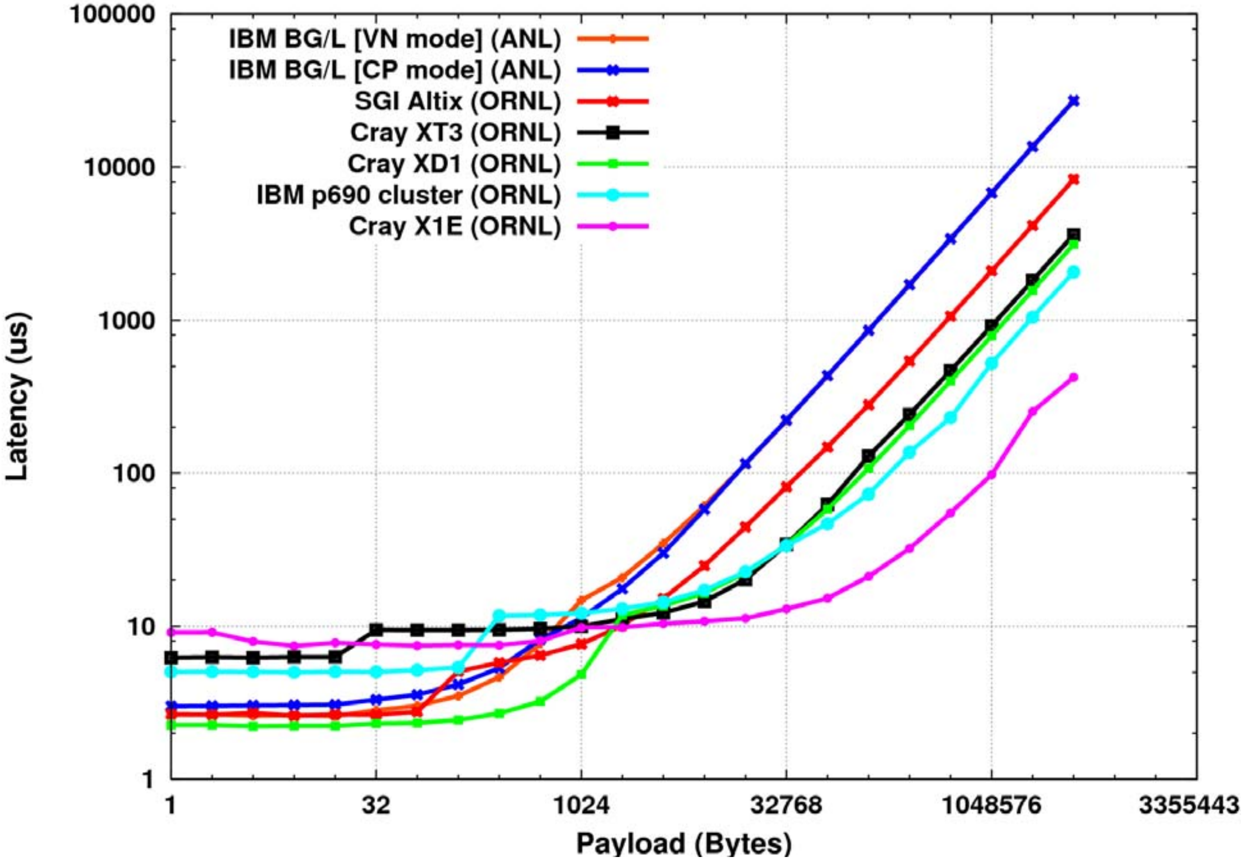

Cray X1E (ORNL)
SGI Altix (NASA)
Cray XT3 (ORNL)
Cray XD1 (ORNL)
IBM p690 cluster (ORNL)
Gflops
0
2
4
6
8
10
12
16
64
256
1024
4096
16384
65536
262144
Vector Order

Communication Kernel Performance

Using Intel MPI Benchmark Suite 2.3

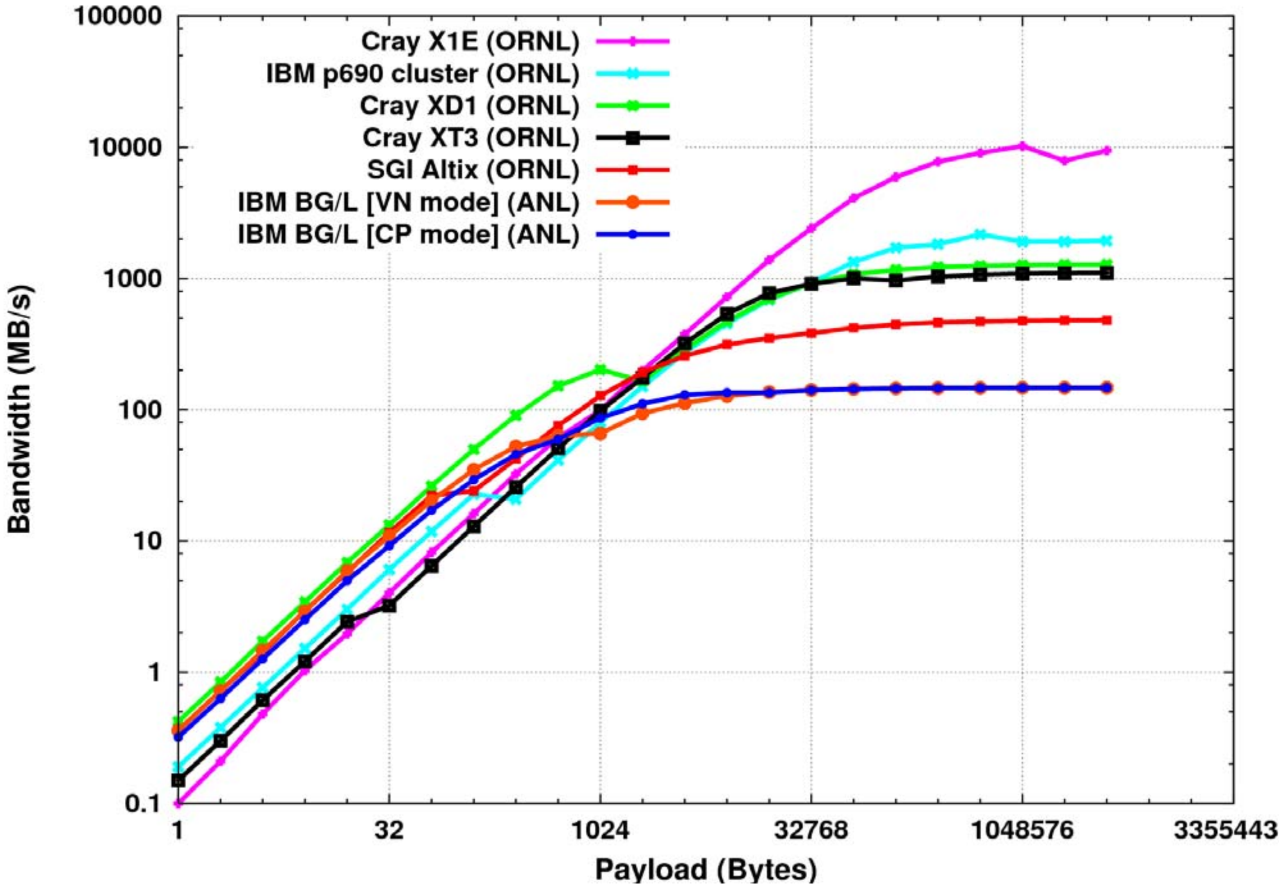
MPI: PingPong Latency

Null msg latency ~6us



MPI: PingPong Bandwidth

Max ~1.1 GB/s at ~64KB


Cray X1E (ORNL)
IBM p690 cluster (ORNL)
Cray XD1 (ORNL)
Cray XT3 (ORNL)
SGI Altix (ORNL)
IBM BG/L [VN mode] (ANL)
IBM BG/L [CP mode] (ANL)
Bandwidth (MB/s)
Payload (Bytes)