Performance Evaluation of the Cray XT3

Jeffrey S. Vetter, Sadaf R. Alam,
Thomas H. Dunigan, Jr., Mark R. Fahey,
Philip C. Roth, and Patrick H. Worley

Oak Ridge National Laboratory

Oak Ridge, TN 37831

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

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

http://www.nccs.gov

Early Evaluation at ORNL

- Several systems have been evaluated with an eye toward important DOE applications; recent examples:
  - Cray X1E (See Pat Worley's talk later this session)
  - SGI Altix 3700
  - Cray XD1
- 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
- Hierarchical, staged approach
  - Microbenchmarks
  - Kernels
  - Applications from important DOE application areas

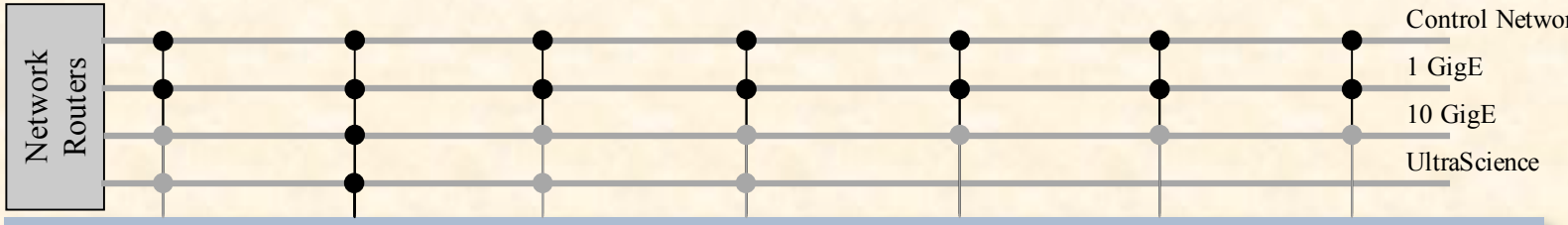
Our Team and Collaborators

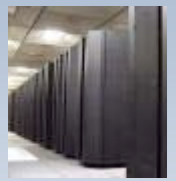
- Future Technologies Group
  - Sadaf Alam
  - Richard Barrett
  - Nikhil Bhatia
  - Jeff Kuehn
  - Collin McCurdy
  - Jeremy Meredith
  - Ken Roche
  - Philip Roth
  - Olaf Storaasli
  - Jeffrey Vetter
- Collaborators
  - Pat Worley
  - Melissa Smith
  - Hong Ong
  - Pratul Agarwal
  - SciDAC PERC Team
  - DARPA HPCS Team
  - DoD HPCMP
  - Georgia Tech CSE Dept, CERCS
  - Vendors
  - Many others…

- Performance analysis, evaluation, and modeling of architectures in support of scientific computation
- Research and development of software and algorithms for HPC
- ExCL (Experimental Computing Lab) for examining new technologies
  - FPGAs
  - Array processors
  - Optical processors
  - Multicore processors
  - …
- http://www.csm.ornl.gov/ft

National Center for Computational Sciences Resources

September 2005 Summary



Cray XT3 Jaguar	Cray X1E Phoenix	SGI Altix Ram	IBM SP4 Cheetah	IBM Linux NSTG	Visualization Cluster	IBM HPSS
						
5294 2.4GHz 11TB Memory	1024 .5GHz 2 TB Memory	(256) 1.5GHz 2TB Memory	(864) 1.3GHz 1.1TB Memory	(56) 3GHz 76GB Memory	(128) 2.2GHz 128GB Memory	Many Storage Devices Supported

Shared Disk	120TB	32TB	36TB	32TB	4.5TB	9TB	5TB

Scientific Visualization Lab

- 27 projector Power Wall

Test Systems

- 96 processor Cray XT3
- 32 processor Cray X1E*
- 16 Processor SGI Altix

Evaluation Platforms

- 144 processor Cray XD1 with FPGAs
- SRC Mapstation
- Clearspeed
- BlueGene (at ANL)

Backup Storage 5PB

7 Systems	
Supercomputers 7,622 CPUs 16TB Memory 45 TFlops	
238.5 TB	
5 PB	

The Cray XT3: Hardware

- Massively parallel distributed memory system
- Successor to the T3D, T3E designs
  - 3D mesh/torus
  - Commodity processor
- Each node has:
  - AMD Opteron
  - 1 to 8 GB RAM
    - On-chip memory controller
    - Connected to Opteron with 6.4 GB/s link
  - One Cray SeaStar
    - Communication and routing ASIC
    - Six 7.6 GB/s links
    - Connected to Opteron with 6.4 GB/s HyperTransport

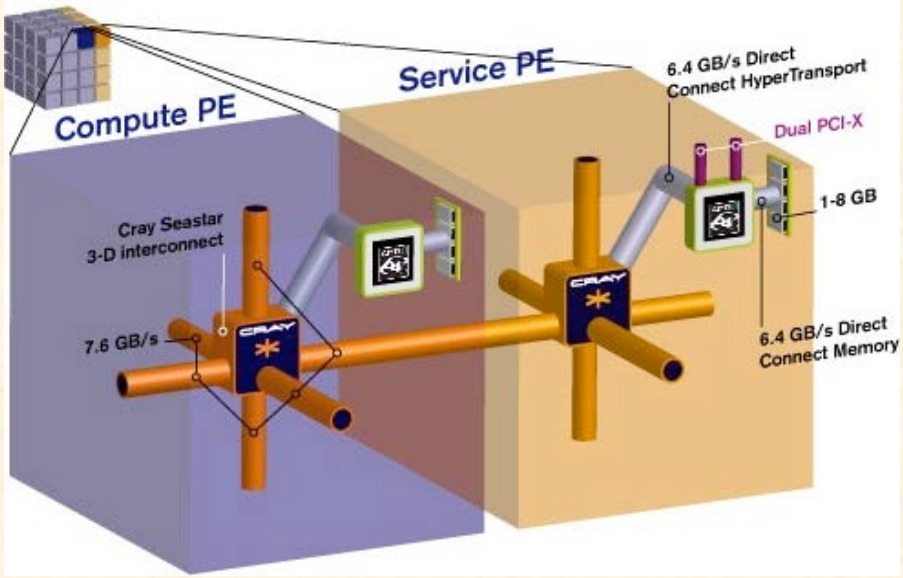
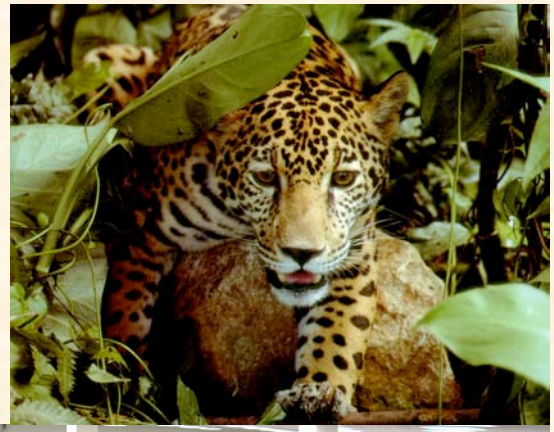


Image courtesy Cray Inc.

The Cray XT3: Software

- Catamount lightweight kernel on compute nodes
  - Single process (single thread)
  - No demand-paged virtual memory
  - POSIX-like system call interface
    - Some POSIX calls not supported (e.g., fork)
    - Some calls with POSIX signature but not POSIX semantics
- Linux on login and I/O nodes
- Portals data movement layer
  - Connectionless, reliable, in-order delivery
  - One-sided and two-sided communication models
- AMD Core Math Library (ACML) for BLAS, LAPACK, FFT
- Lustre parallel file system

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
  - Cables now certified for torus in second dimension
- 56 cabinets
- Also: rizzo 96-node test system

Microbenchmarks

- Goal: characterize specific architectural components
- Suite
  - Arithmetic performance (ACML 2.6 on the XT3)
  - Memory-hierarchy performance
  - Task and thread performance
  - Message-passing performance
  - System and I/O performance
  - Parallel I/O

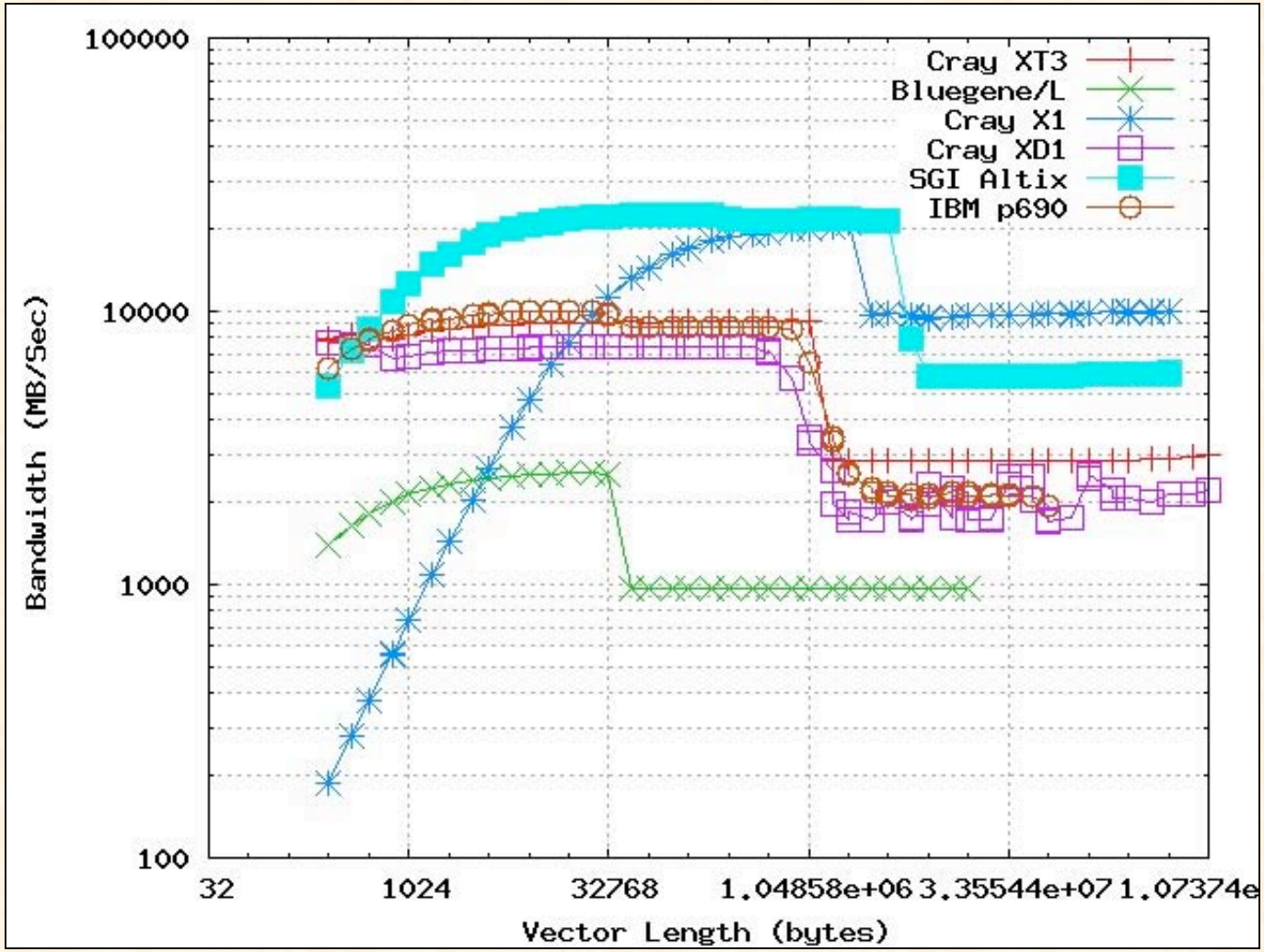
Microbenchmarks: Memory Hierarchy

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

- Opterons have on-chip memory controller, Intel Xeon has off-chip "Northbridge"
- Opteron 150 does not support SMP configurations (no coherent HyperTransport support)

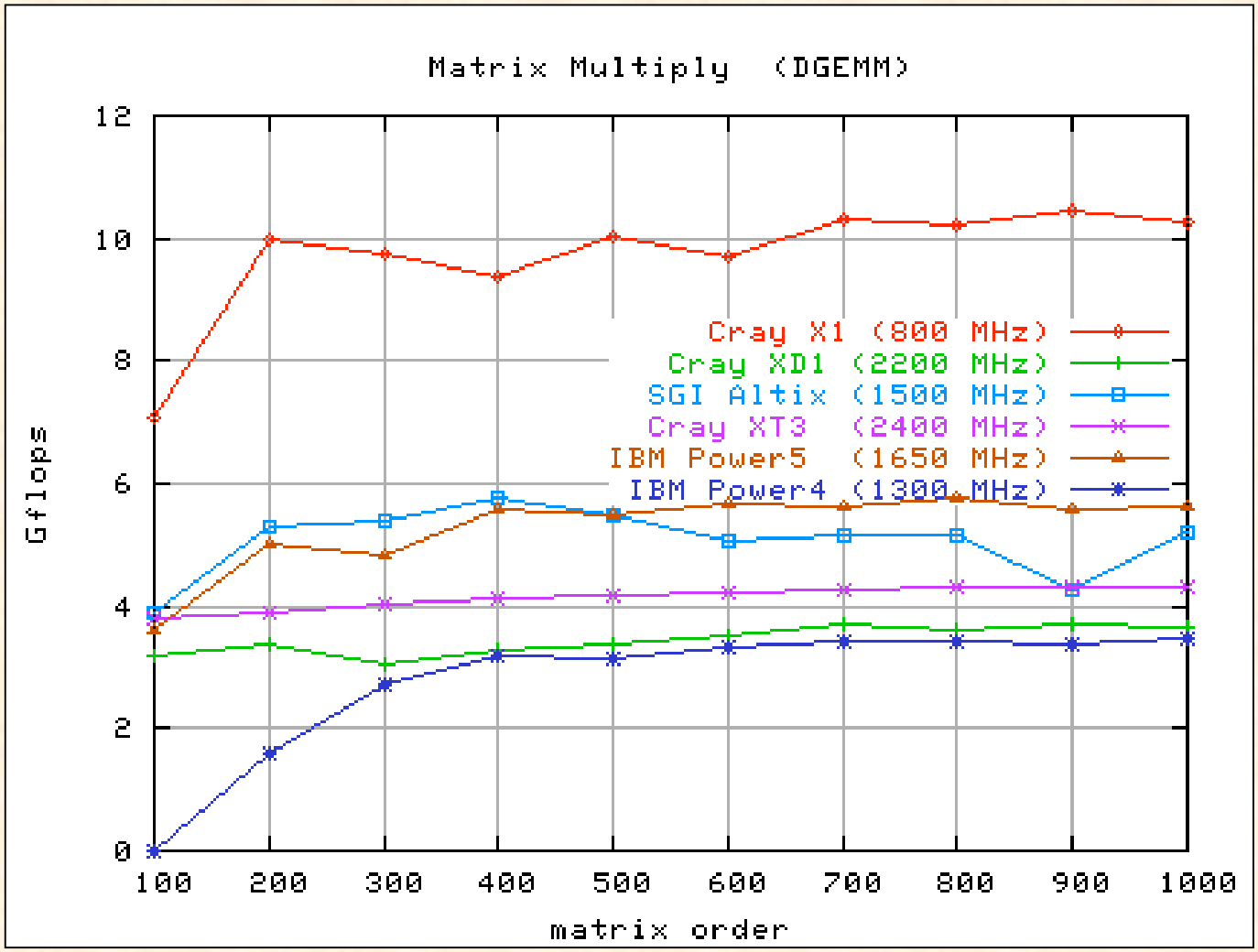
Microbenchmarks: Memory Hierarchy

- Cachebench read bandwidth
- XT3 achieves approximately 9 GB/s from L2 cache, 3 GB/s from main memory



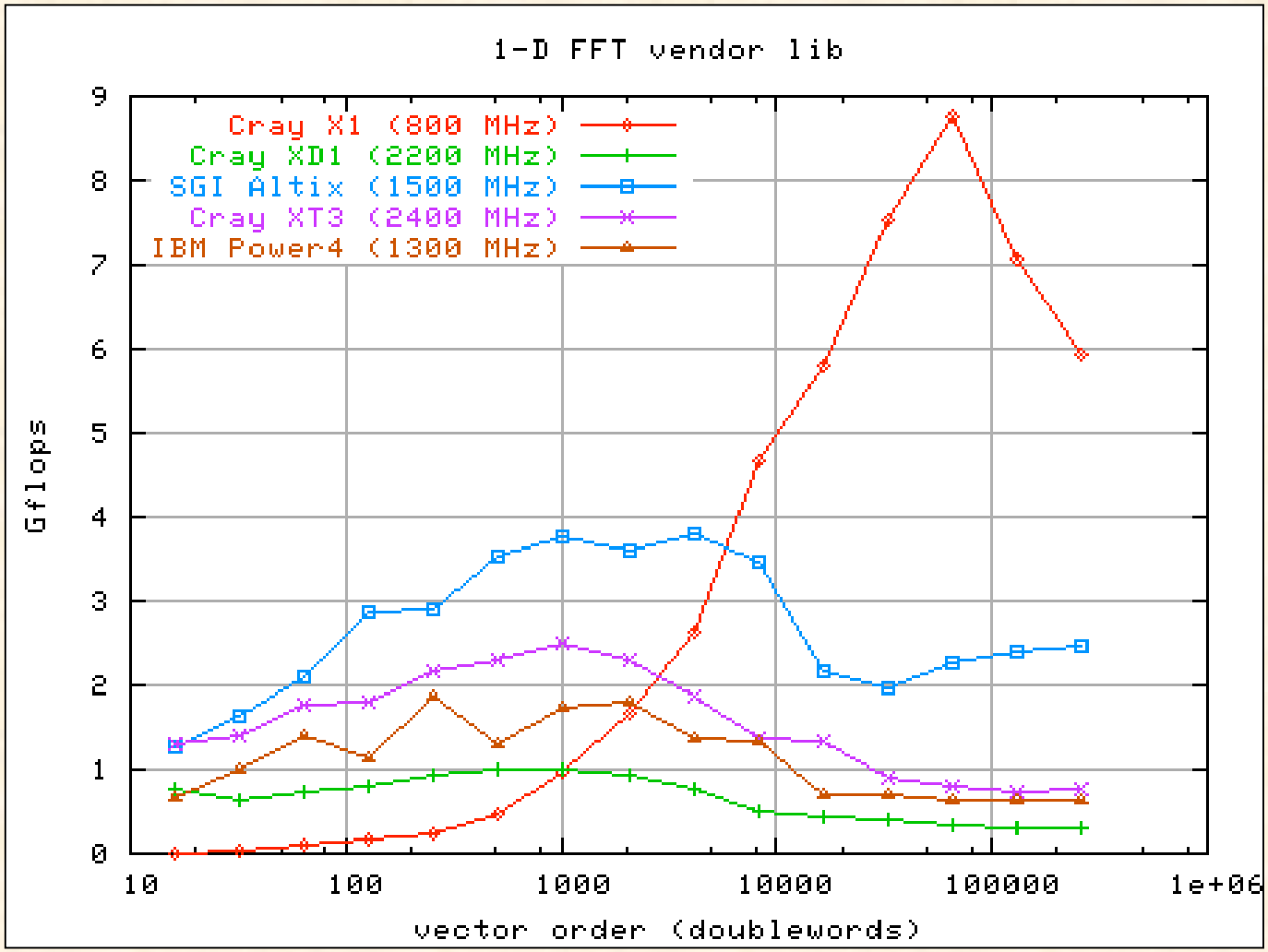
Microbenchmarks: Matrix Multiply

- DGEMM with vendor math library
- XT3 achieves about 4 Gflop/s (83% of peak) with ACML 2.6



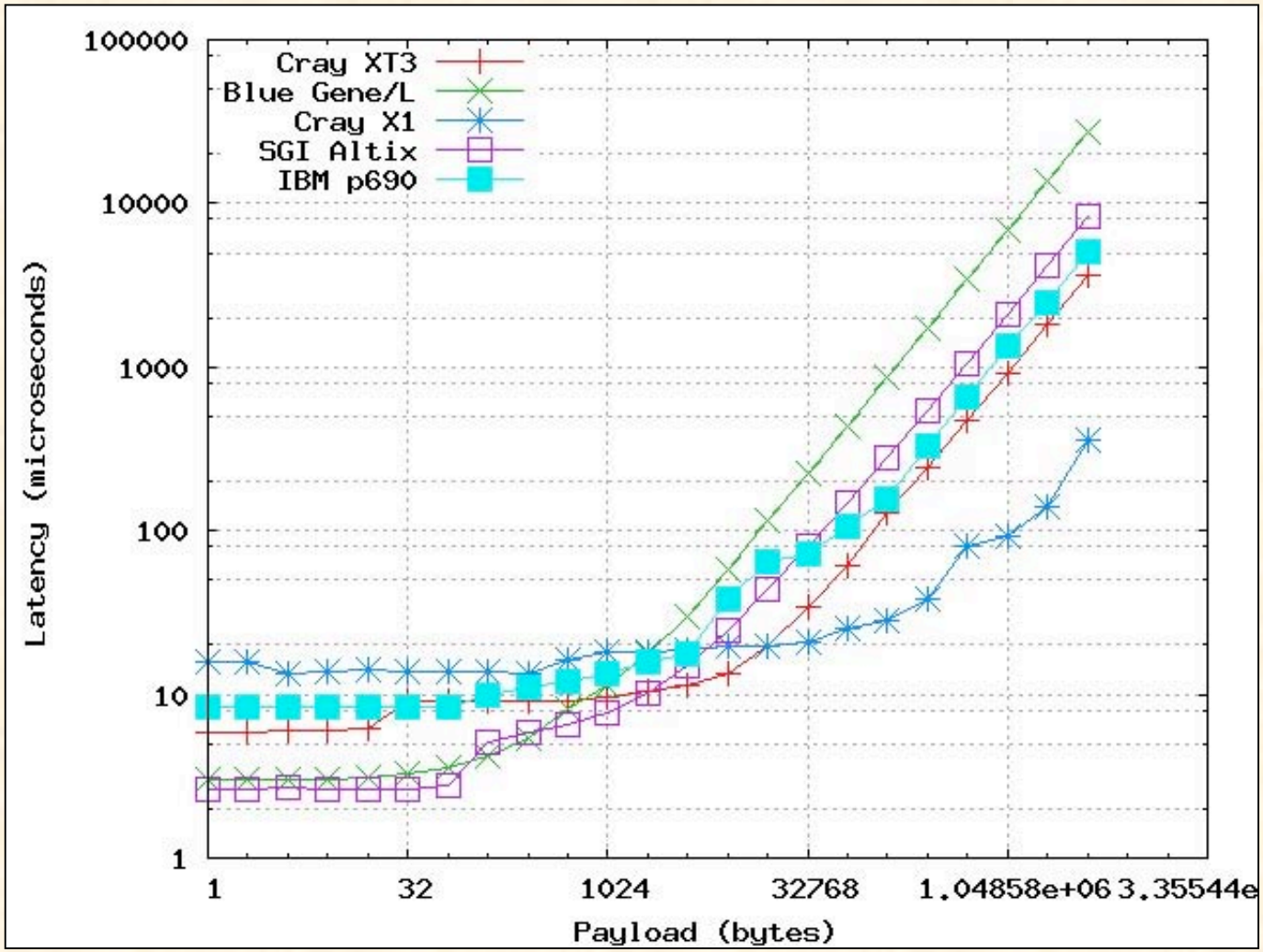
Microbenchmarks: Vendor FFT

- FFT using vendor math library
- XT3 better than X1 for short vector lengths, but outperformed by Altix at all vector lengths



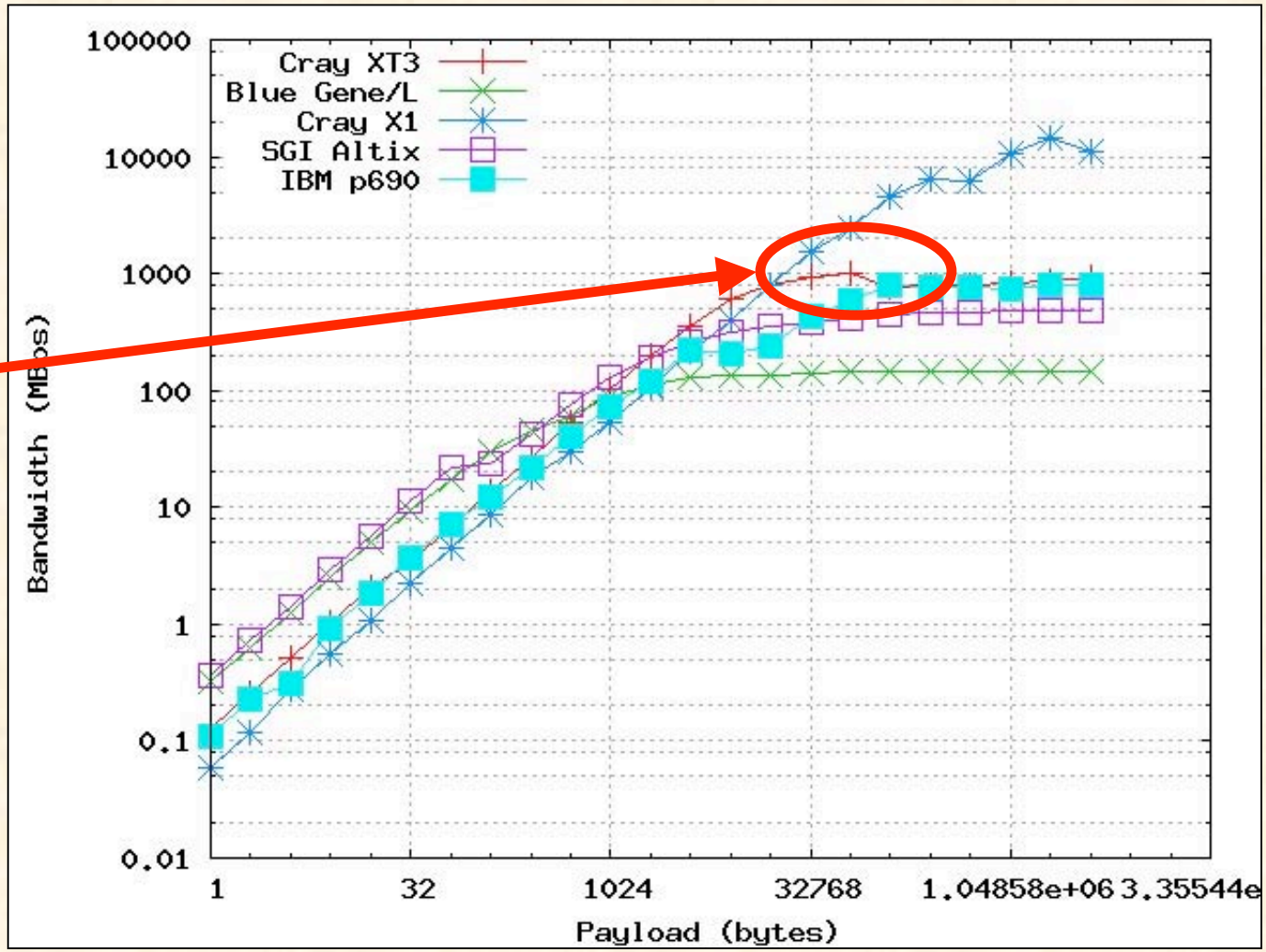
MPI: PingPong Latency

- Pallas MPI Benchmark Suite
- Approximately 8 μs latency for 4-byte messages



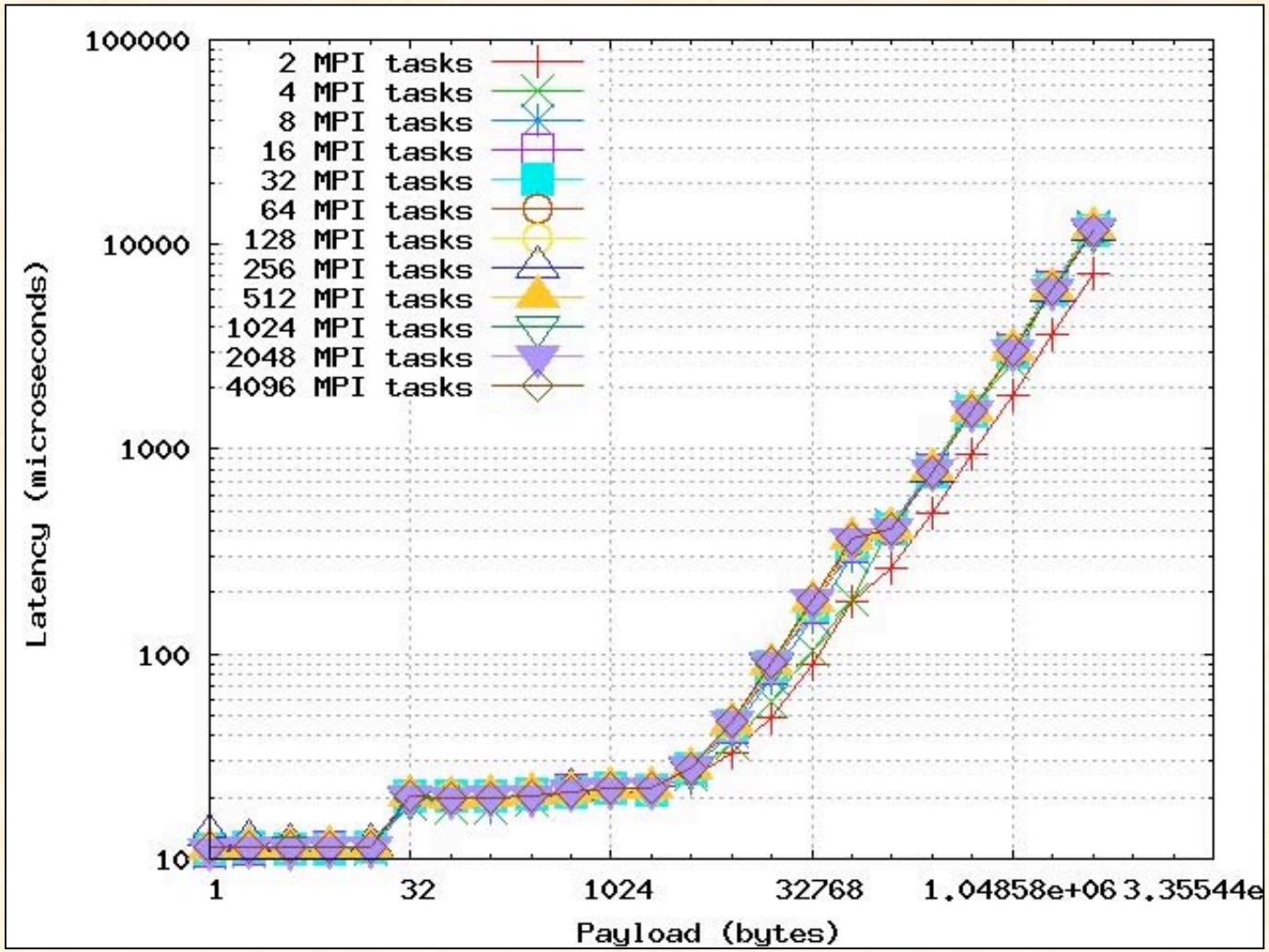
MPI: PingPong Bandwidth

- Pallas MPI Benchmark Suite
- XT3 achieves max bandwidth at about 64 KB payload



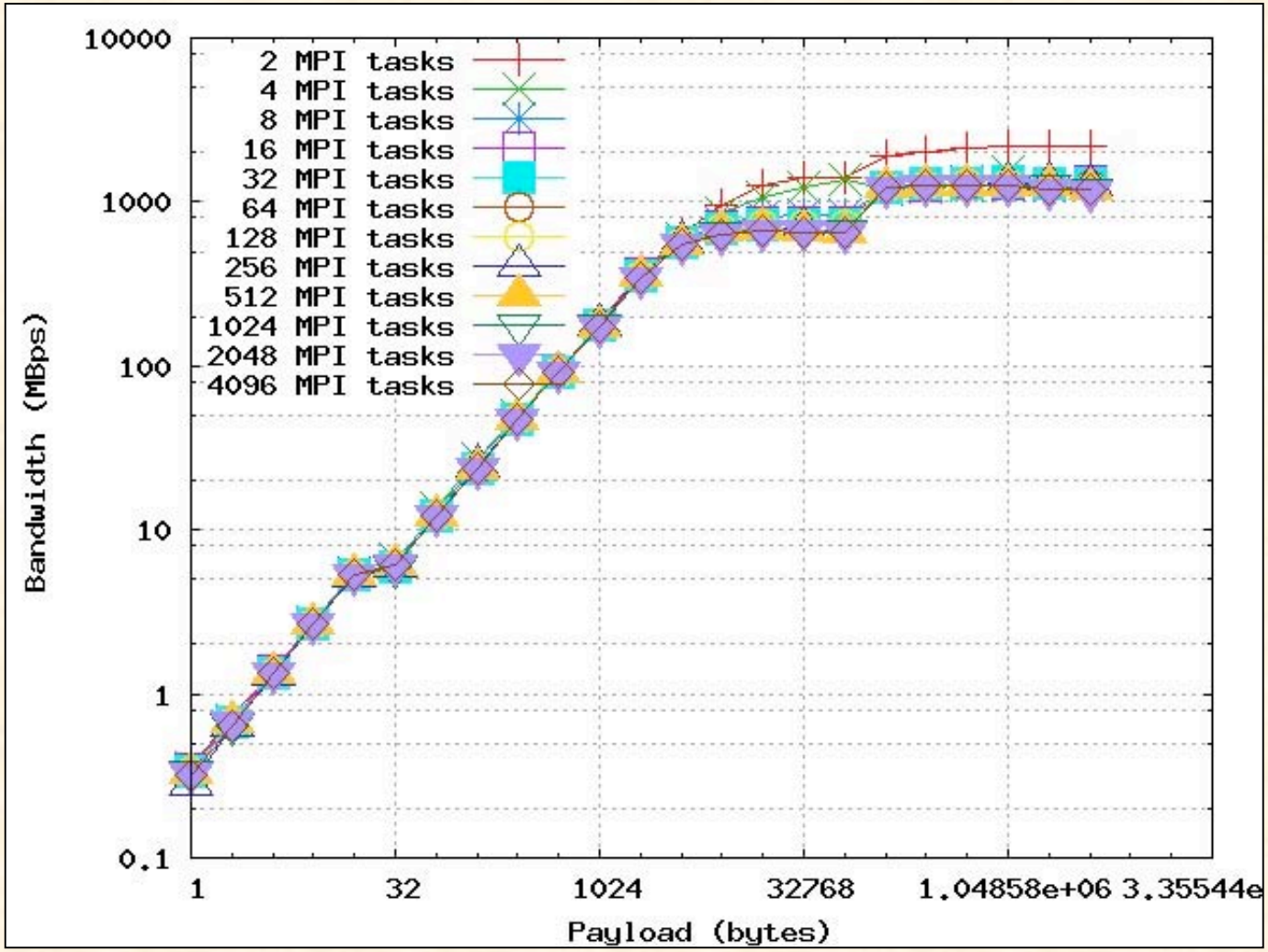
MPI: Exchange Latency

- Pallas MPI Benchmark Suite
- Simultaneous sendrecv between pairs of tasks
- More realistic than PingPong: latency measured with additional system activity
- Approximately 20 μs for 4-byte messages



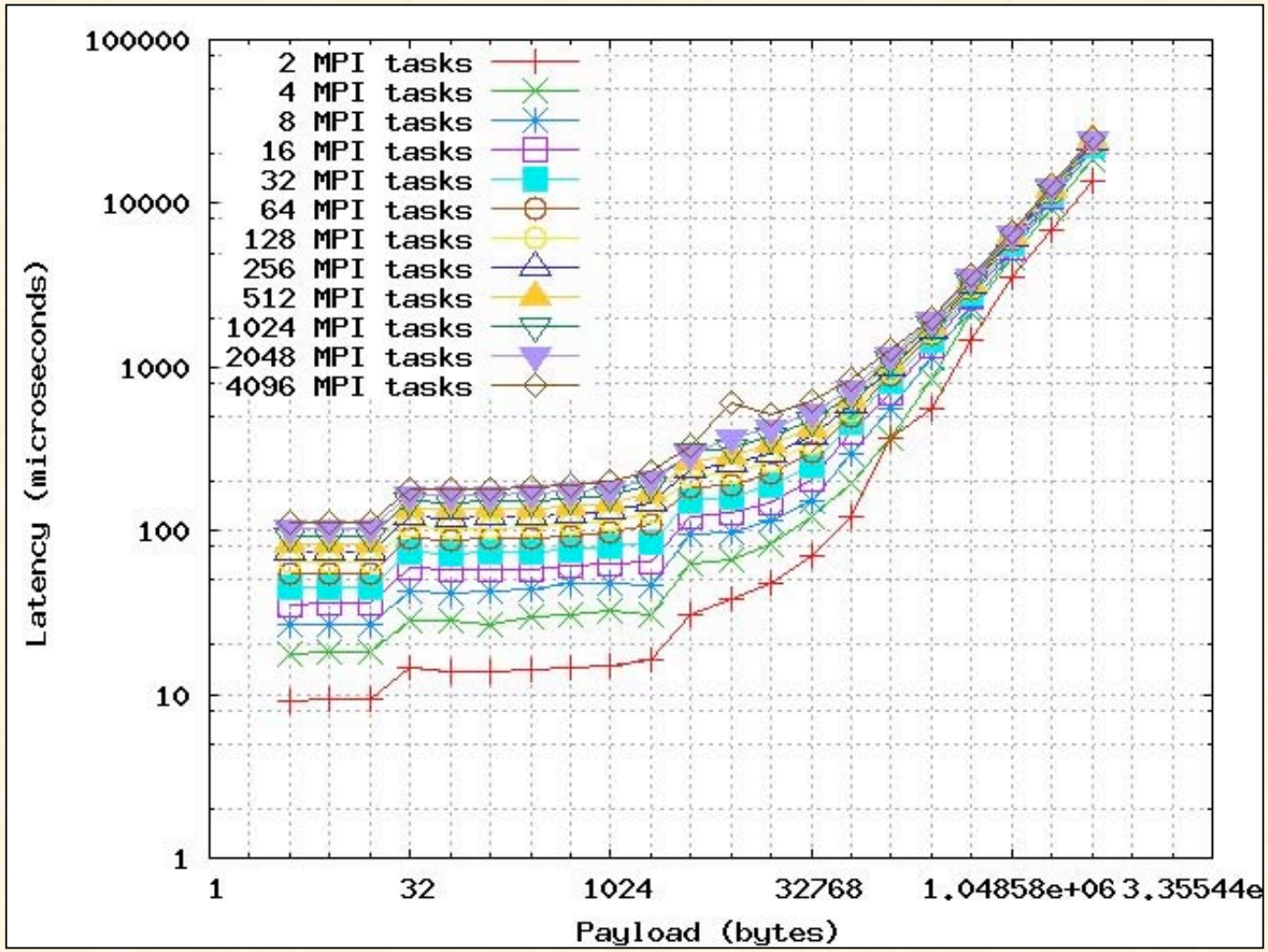
MPI: Exchange Bandwidth

- Pallas MPI Benchmark Suite
- Approx. 1 GB/s per transfer, with 2K simultaneous transfers
- Good scalability with increases in payload size



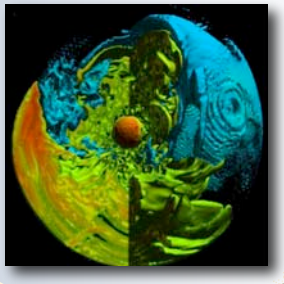
MPI: Allreduce

- Allreduce is a performance-limiting operation for many DOE applications

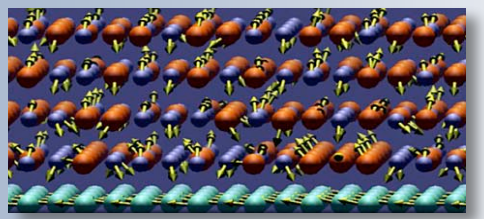


Applications at ORNL

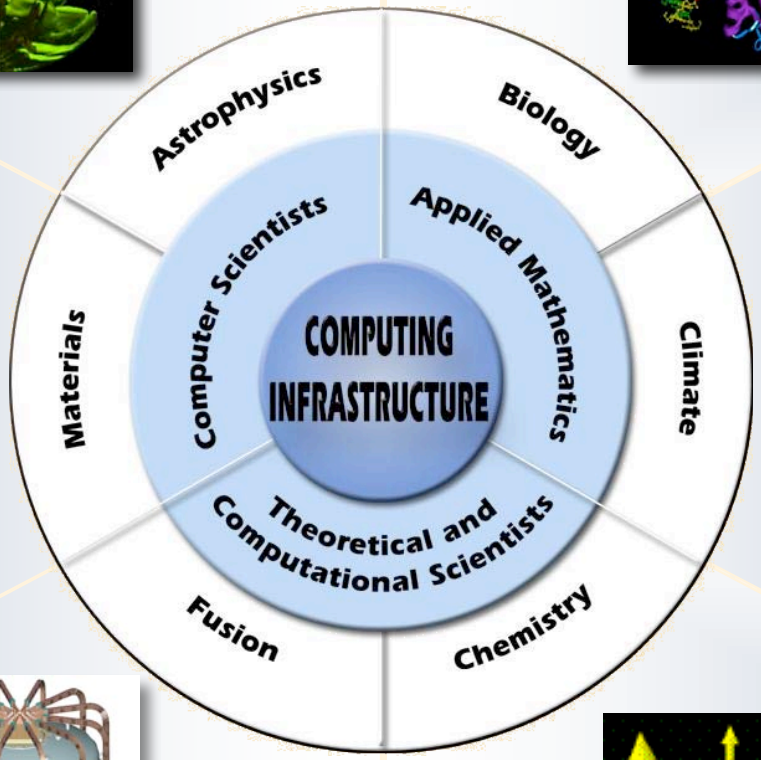
SciDAC Astrophysics

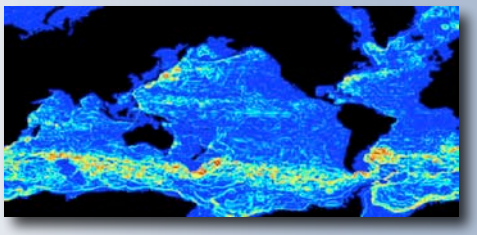


Genomes to Life



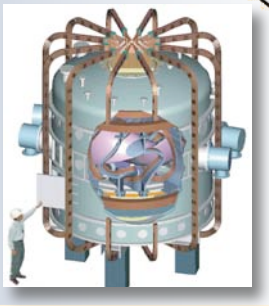
Nanophase Materials

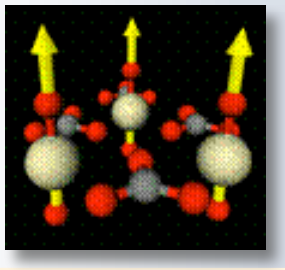




SciDAC Climate

SciDAC Fusion

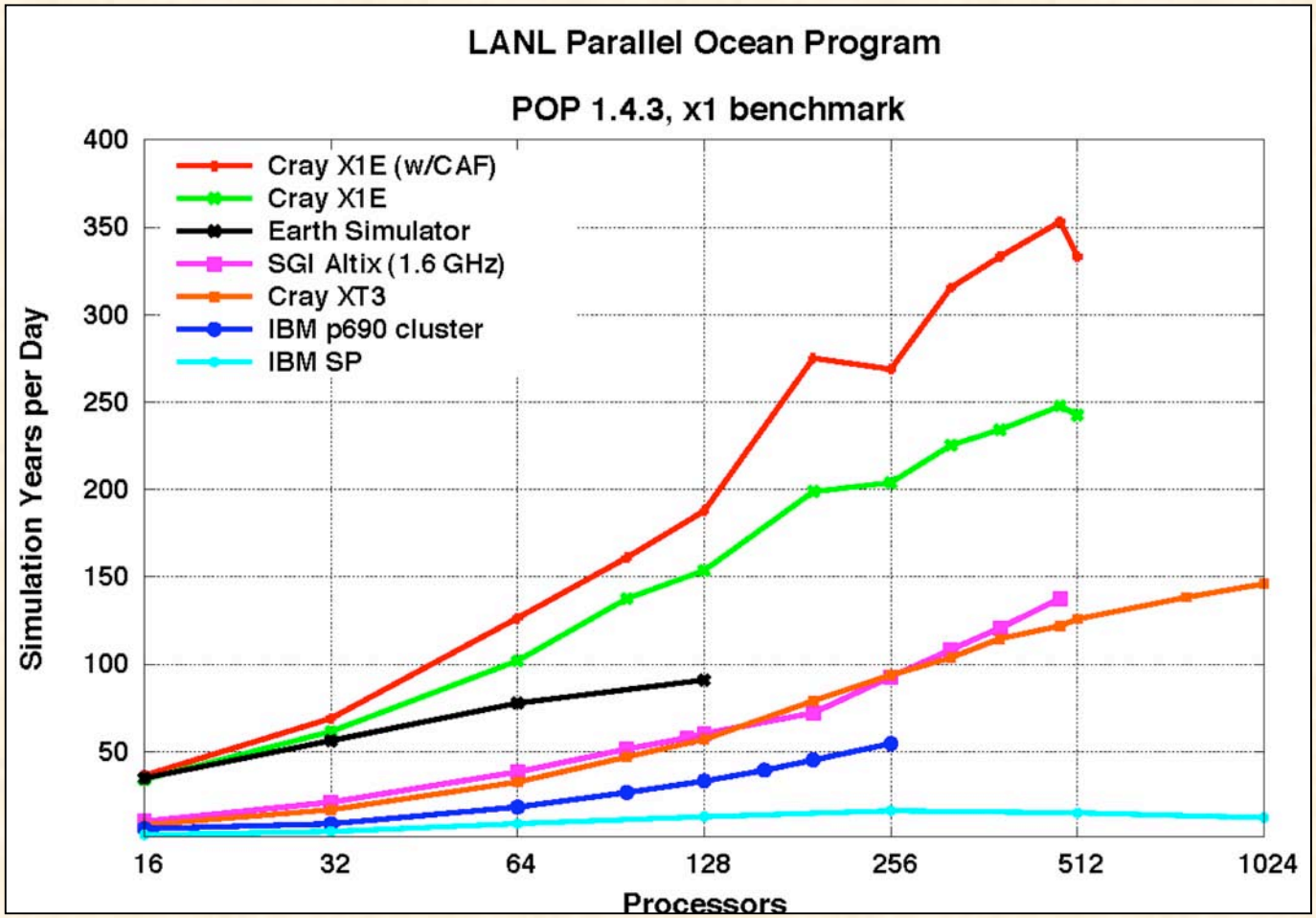




SciDAC Chemistry

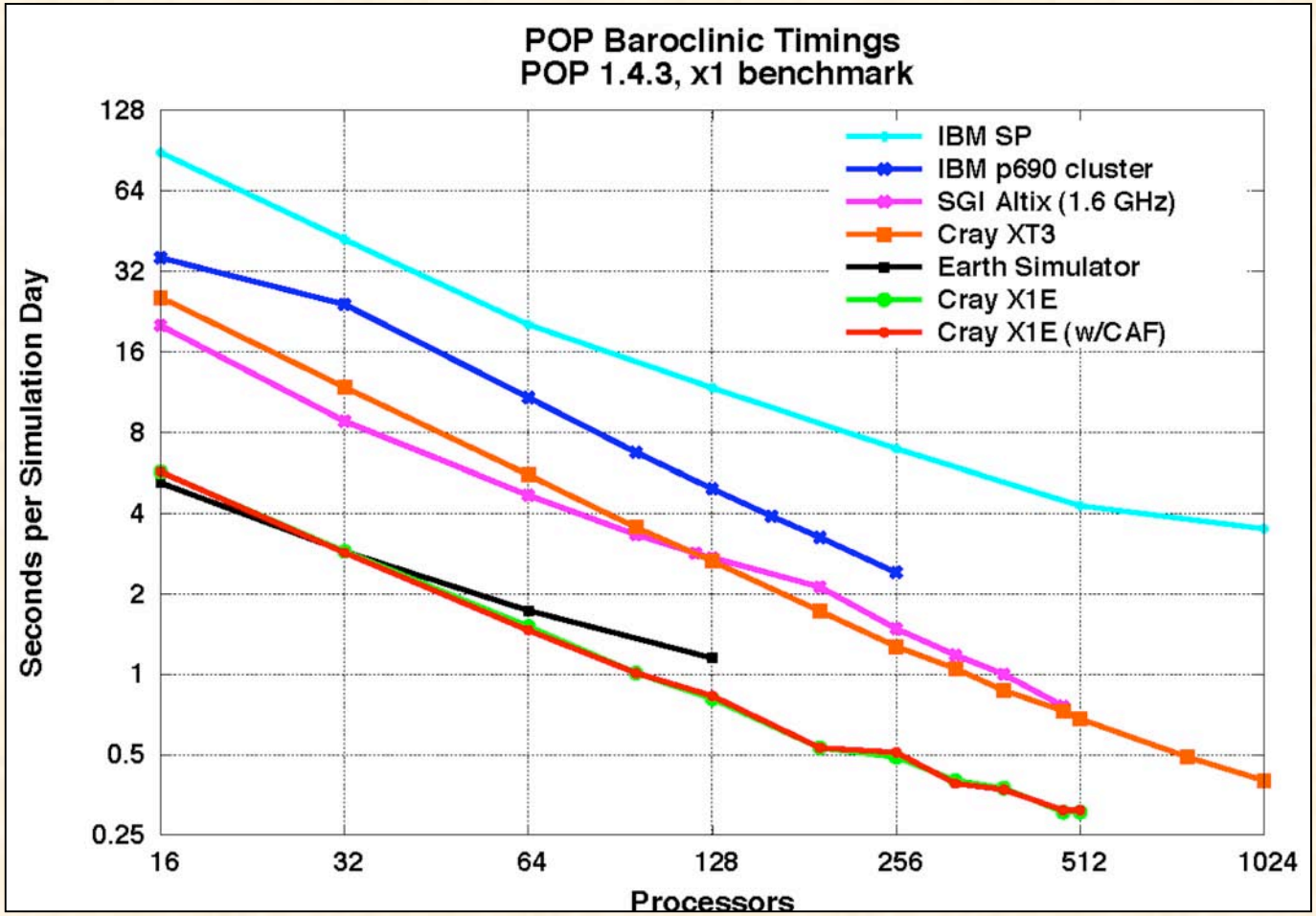
Applications: Climate (POP)

- LANL Parallel Ocean Program
- XT3 performs similarly to SGI Altix



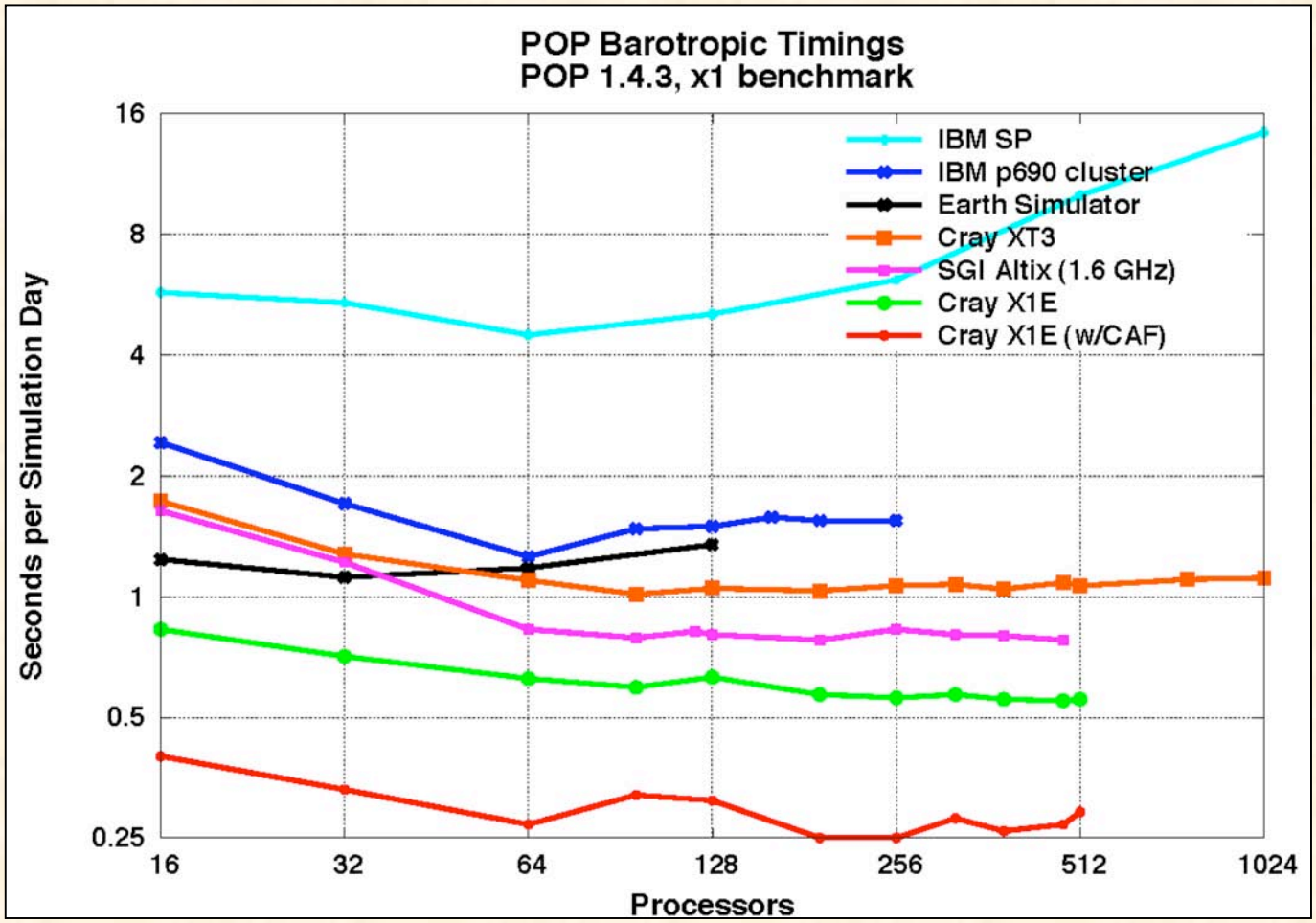
Applications: Climate (POP)

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



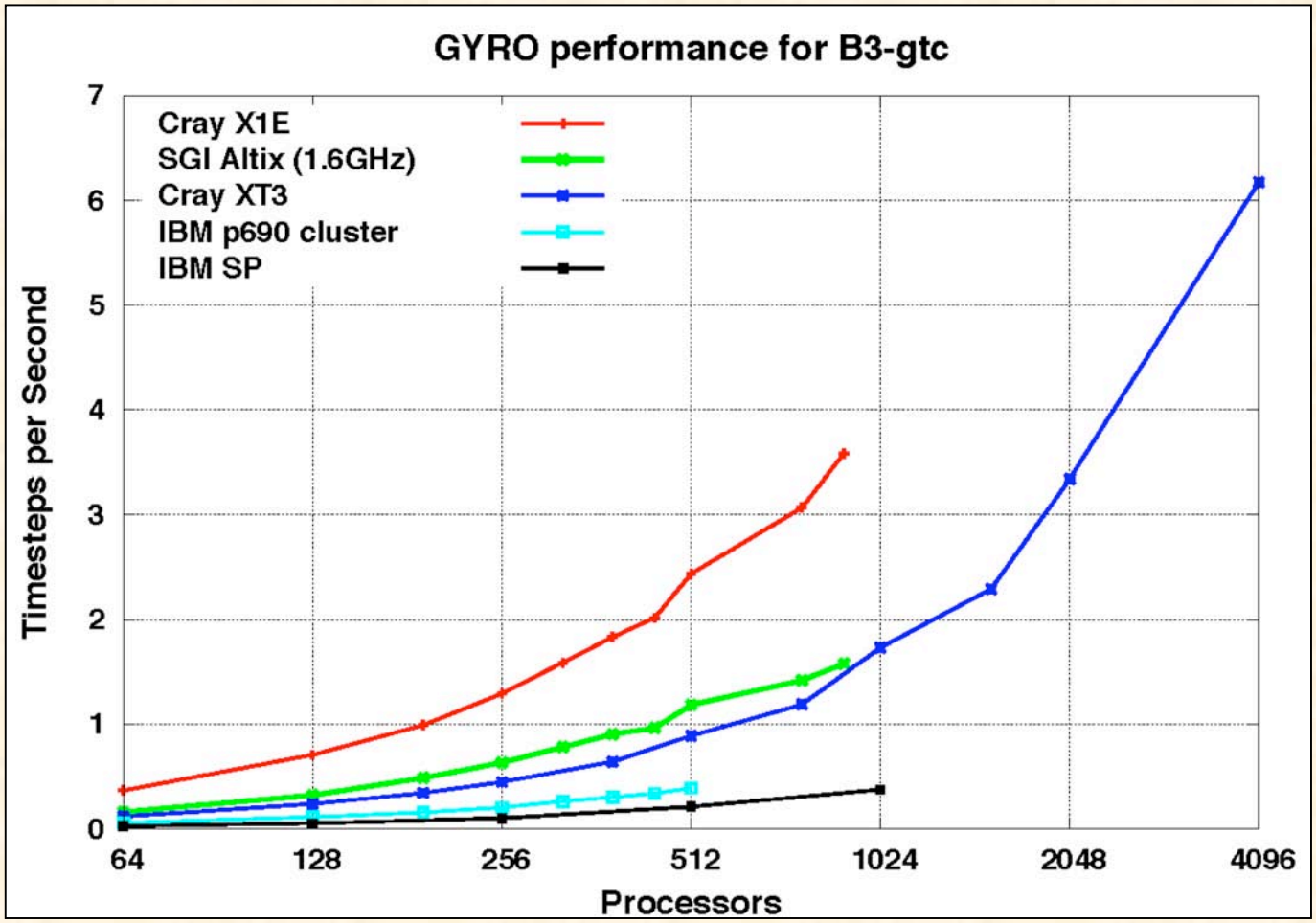
Applications: Climate (POP)

- Barotropic phase
- Usually scales poorly due to CG (including small payload allreduce) and halo exchange
- XT3 throughput remains stable with increases in number of processors



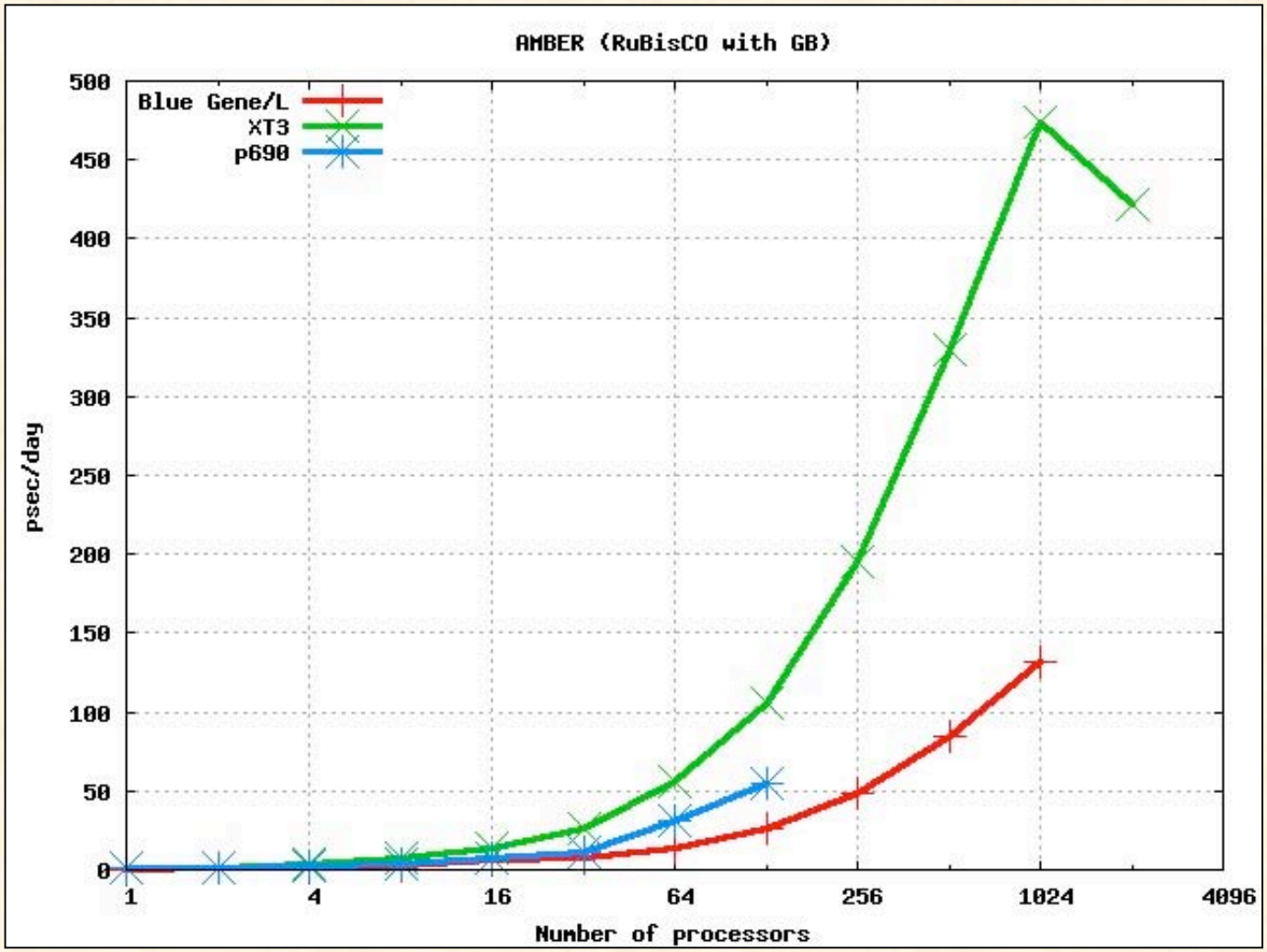
Applications: Fusion

- GYRO: tokamak micro-turbulence
- Communication dominated by simultaneous all-to-alls over process sub-groups



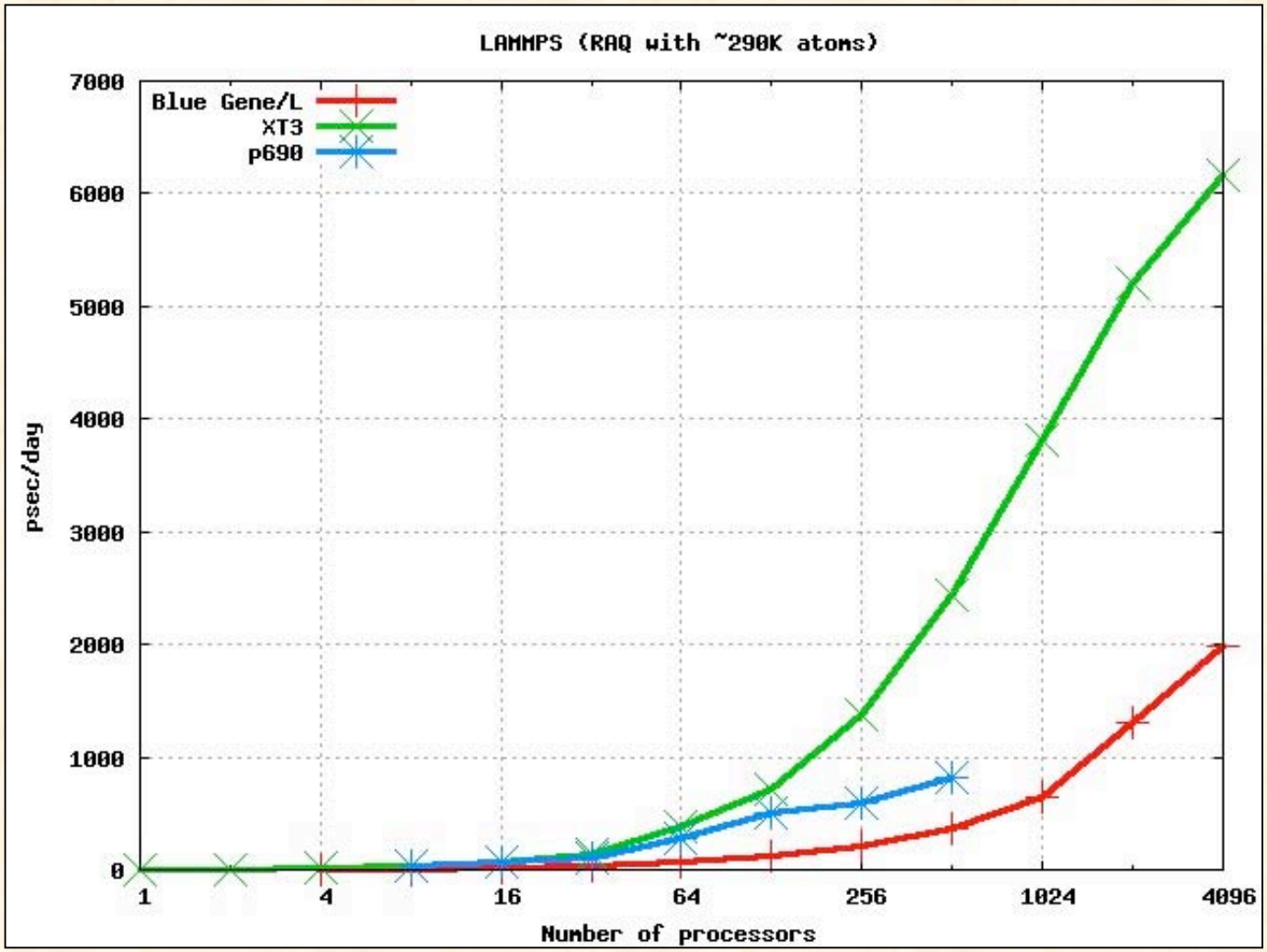
Applications: Molecular Dynamics (AMBER)

- Nearly 74K atoms
- Good XT3 throughput till 1K processors



Applications: Molecular Dynamics (LAMMPS)

- 290K atoms
- Explicit solvent
- XT3: over 40% efficiency with 4K processors



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

Summary

- Oak Ridge National Laboratory has deployed a 5294-processor, 25 TFlop Cray XT3
- We evaluated the system using micro-benchmarks, kernel, and applications from important DOE areas
  - Good scalar performance
  - High interconnect bandwidth with respect to other microprocessor-based systems

- XT3 shows strong performance stability across DOE application workload

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