Performance Evaluation of the SGI Altix 3700

Tom Dunigan, Pat Worley, Jeffrey Vetter, Mark Fahey

Oak Ridge National Laboratory

THE CENTER FOR COMPUTATIONAL SCIENCES

Office of Science
U.S. DEPARTMENT OF ENERGY

ornl

Acknowledgements

- Research sponsored by the Atmospheric and Climate Research Division and 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.
- These slides have been authored by a contractor of the U.S. Government under contract No. DE-AC05-00OR22725. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes
- Oak Ridge National Laboratory is managed by UT-Battelle, LLC for the United States Department of Energy under Contract No. DE-AC05-00OR22725.

SGI Altix 3700 @ ORNL

Ram

SGI Altix 3700 with 256 processors

- **Itanium 2 processors running at 1.5 GHz**
  - 16K L1 instruction cache, 16K L1 data cache
  - 256K L2 cache
  - 6MB L3 cache.
- **8 GB of memory per processor, for a total of 2 TB of shared memory**
- **17 TB of disk space**
- **Four processor "compute brick" (C-brick), made up of two 2-processor + memory nodes**
- **Fat-tree NUMAflex network, with support for MPI and SHMEM communication**
- **Linux OS, Intel compilers**

C-brick Architecture

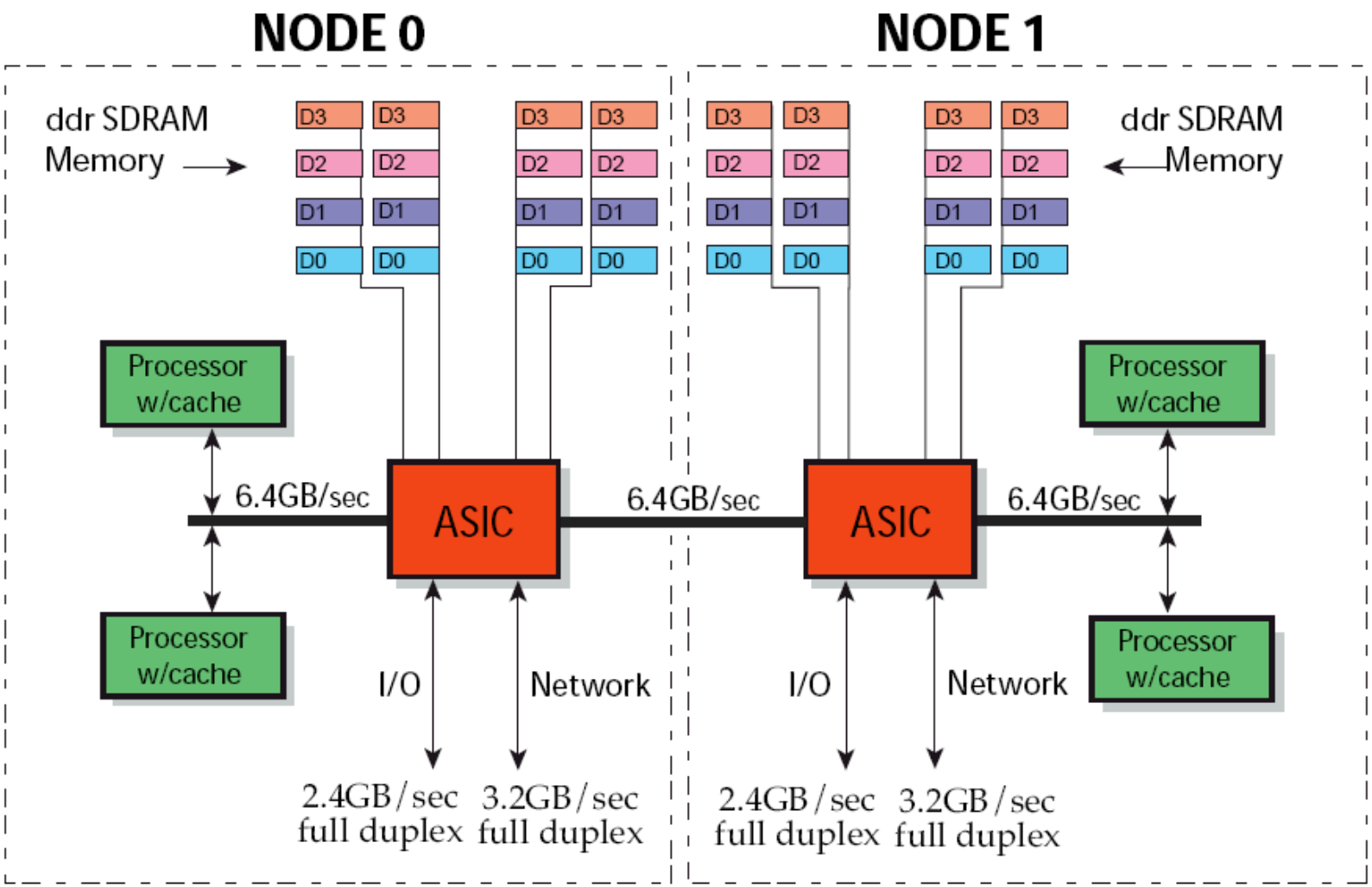


Image courtesy of SGI.

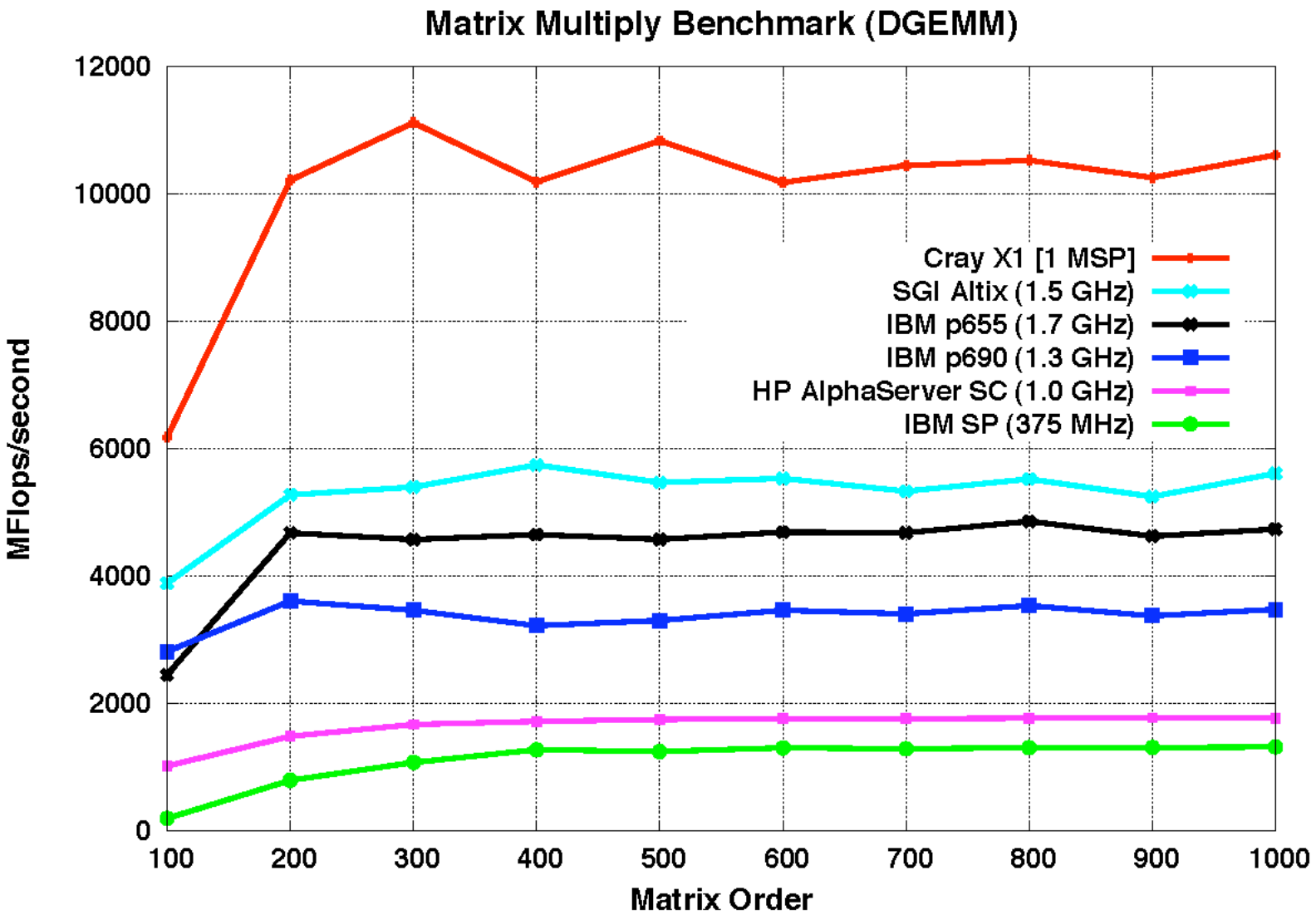
Evaluation Methodology

- **Hierarchical evaluation**
  - Microbenchmarks
  - Application-relevant kernels
  - Compact or full parallel application codes
- **Open evaluation**
  - Rapid posting of evaluation results
  - Systems available to external performance researchers
- **Fair evaluation**
  - Determining appropriate way of using system, evaluating both traditional and alternative programming paradigms
  - Collecting data with both standard and custom benchmarks

Other Platforms

- **Cray X1**: 128 SMP nodes and a torus interconnect. 4 Multi-Streaming Processors (MSP) per node. 4 Single Streaming Processors (SSP) per MSP. 2 64-bit floating point vector units running at 800 MHz per SSP.
- **Earth Simulator (ES)**: 640 8-way vector SMP nodes and a 640x640 single-stage crossbar interconnect. Each processor has 8 64-bit floating point vector units running at 500 MHz.
- **HP/Compaq AlphaServer SC at Pittsburgh Supercomputing Center (SC45)**: 750 ES45 4-way SMP nodes (1GHz Alpha EV68) and a Quadrics QsNet interconnect with two network adapters per node.
- **IBM p690 cluster at ORNL**: 27 32-way p690 SMP nodes (1.3 GHz POWER4) and a HPS interconnect with two to two-port network adapters per node. Also 8 4-way p655 SMP nodes (1.7 GHZ POWER4+).
- **IBM SP at the National Energy Research Supercomputer Center (SP3)**: 184 Nighthawk II 16-way SMP nodes (375MHz POWER3-II) and an SP Switch2 with two network adapters per node.

Serial DGEMM



Serial Streams

MB/sec

	copy	scale	add	triad
Altix	3183	3118	3795	3807
ES45	1946	1941	1978	1978
P655	3069	3255	2436	2477
P690	1774	1860	2098	2119
SP3	486	494	601	601
X1	22111	21634	23658	23752

copy: x(i)=y(i) **scale:** x(i)=A*y(i)

add: z(i)=x(i)+y(i) **triad:** z(i)=x(i)+A*y(i)

MPI Latency and Bandwidth

	Intra node / Inter node	
	latency (min 1 way us)	**bandwidth (MB/sec)**
Altix	1.5/2.8	1090/ 568
P690	3.0/7.0	2186/ 1400
SC45	5.0/4.6	746/ 265
X1	7.3/7.3	12125/11776

- Latency is for 8 byte message. Bandwidth is for 1 MB message.
- Latency is 1-way (half of RTT).
- Altix node is a C-brick.

MPI Exchange Experiments

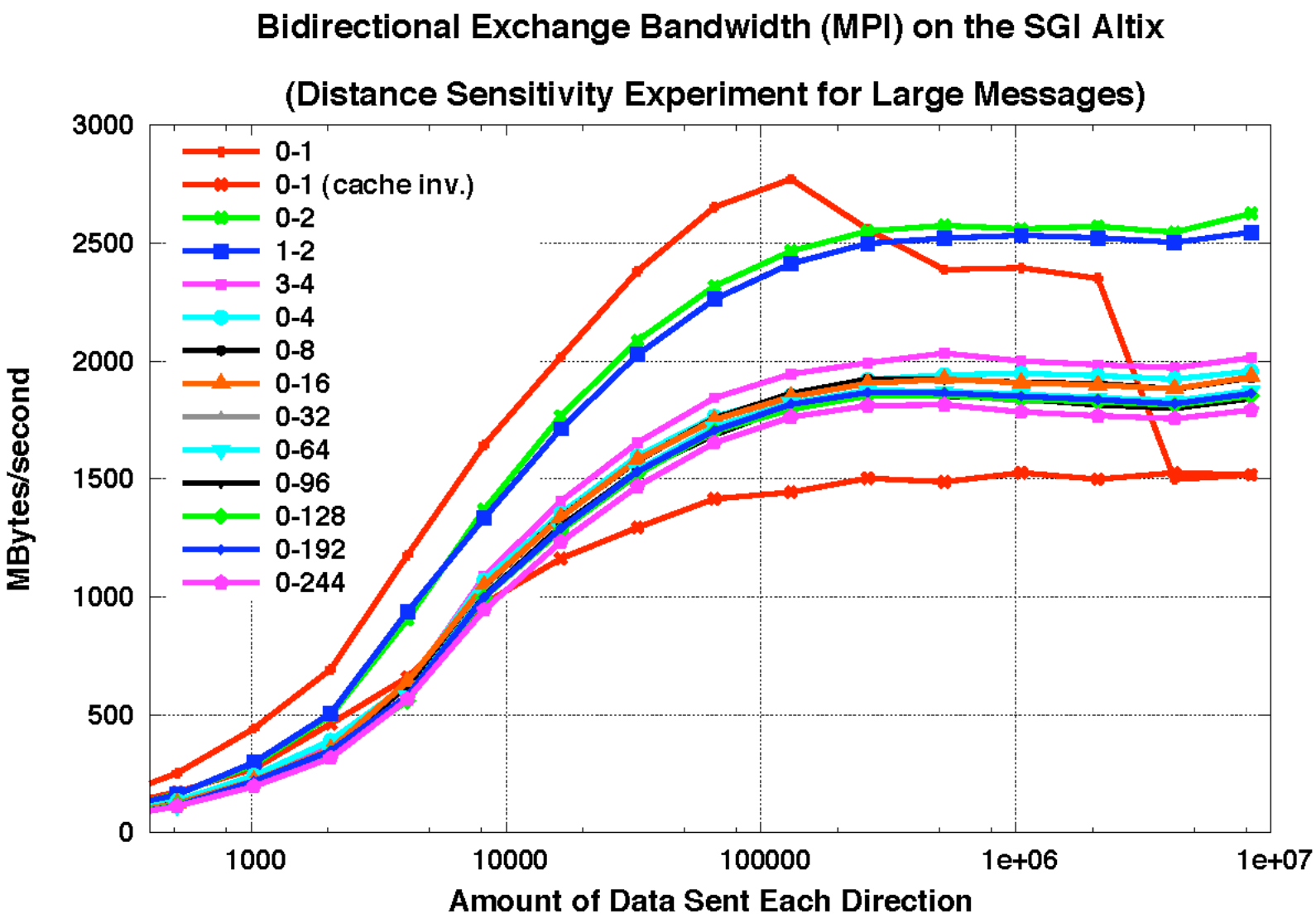
i-j

processor i swaps data with processor j. Depending on i and j, this can be within an SMP node or between SMP nodes.

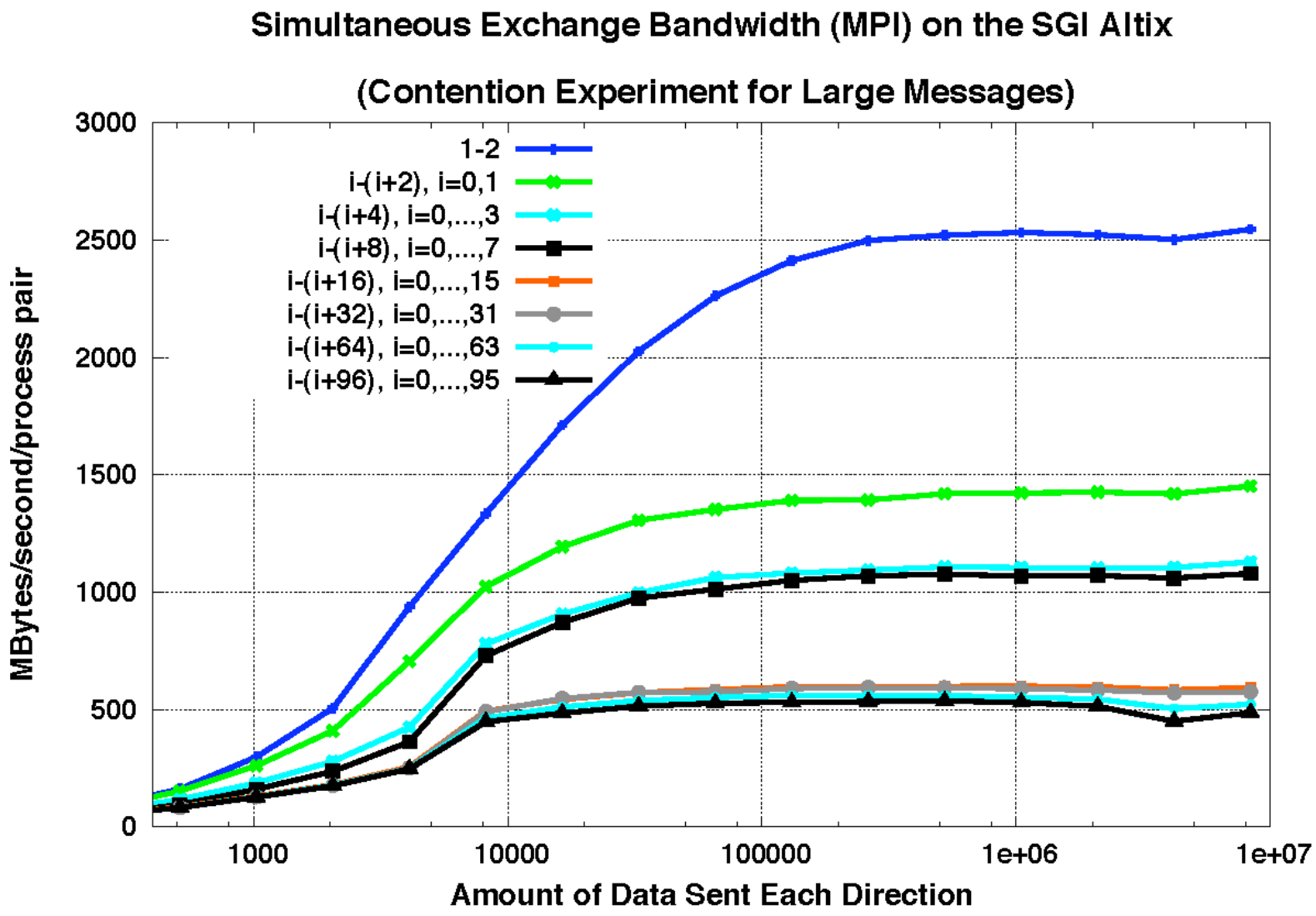
i-(i+j), i=1,n

n processor pairs (i,j) swap data simultaneously. Depending on j, this will be within an SMP node or between SMP nodes (or both).

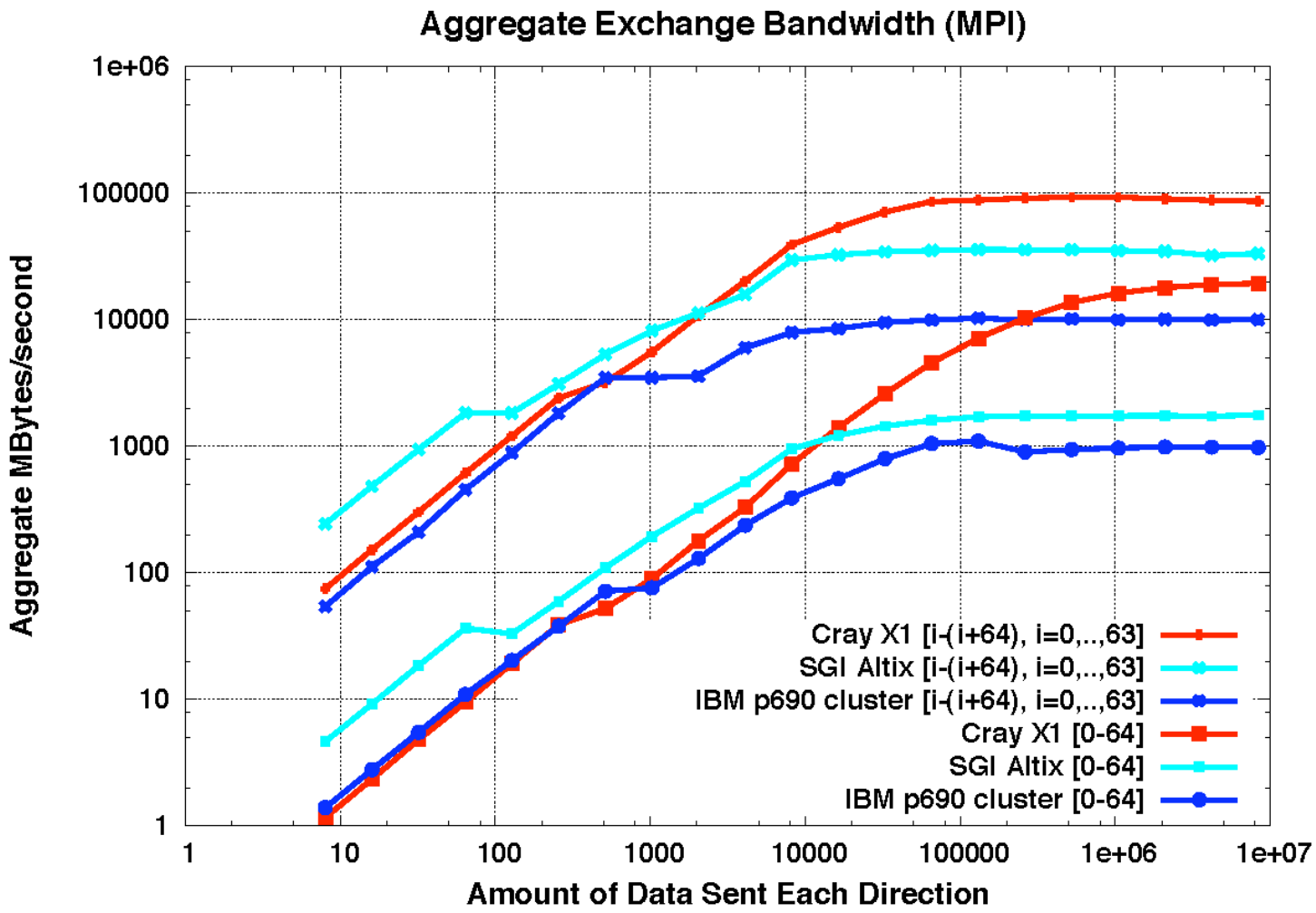
MPI Exchange: Distance Experiments



MPI Exchange: Contention Experiments



MPI Exchange: Platform Comparison



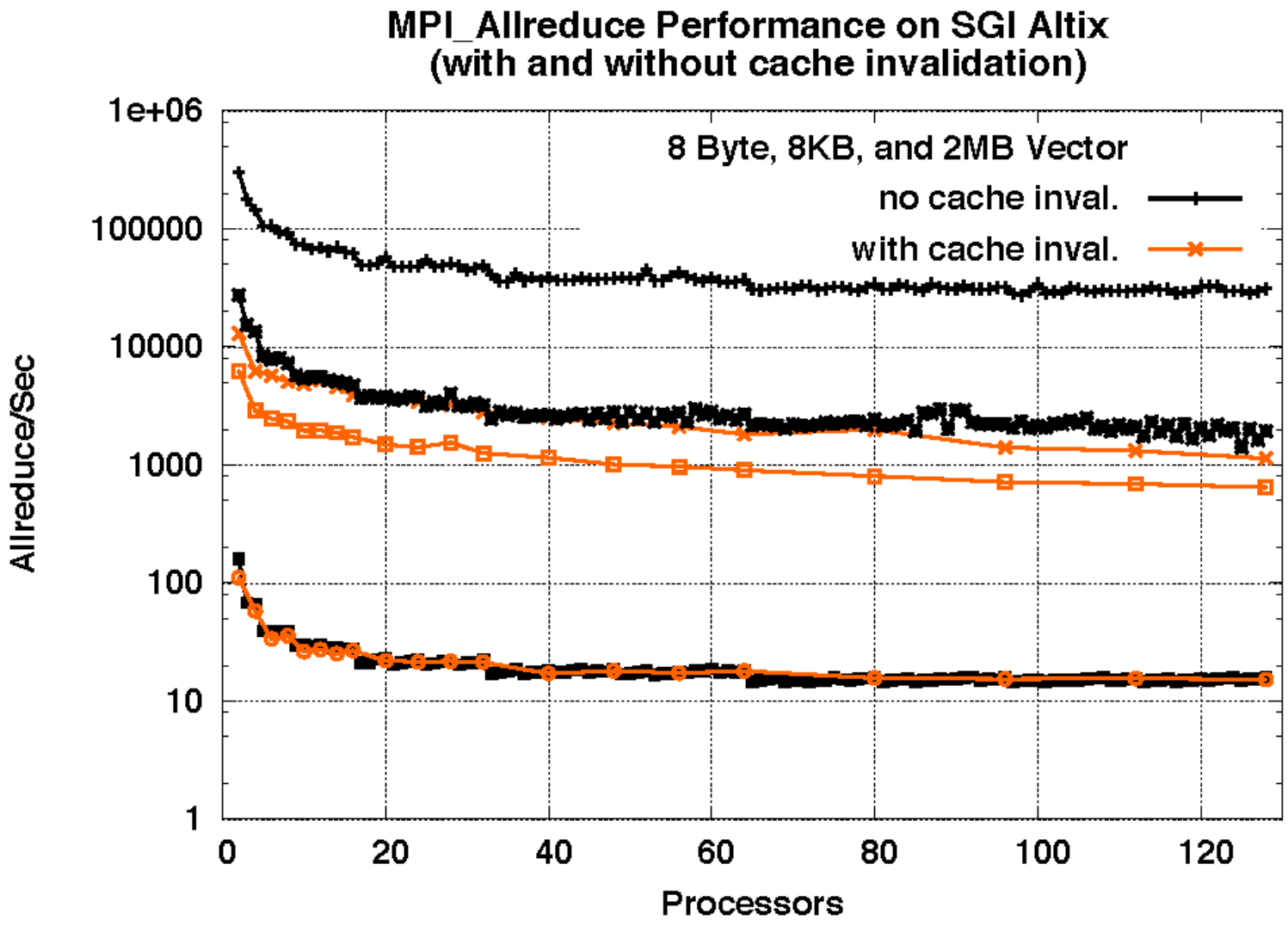
Comparing performance of exchange for single processor pair a distance of 64 apart and for 64 processor pairs, each a distance of 64 apart, simultaneously.

Allreduce Experiments

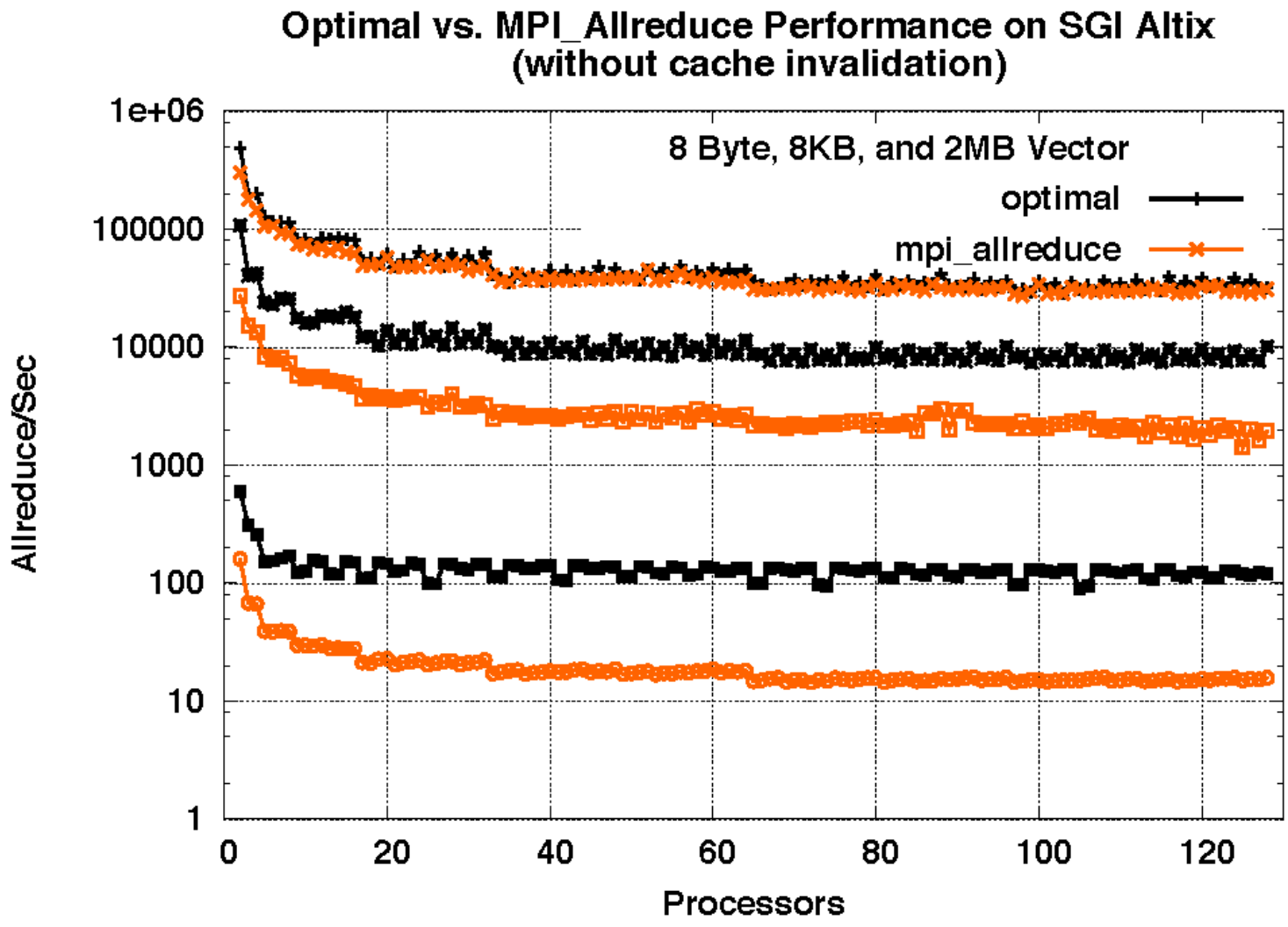
- Impact of cache invalidation on MPI_Allreduce performance
- Comparison of MPI_Allreduce performance with optimal point-to-point implementations (MPI-1 and SHMEM)
- Platform Comparisons

All with SUM operator and 3 different vector lengths: 8B, 8KB, 2MB.

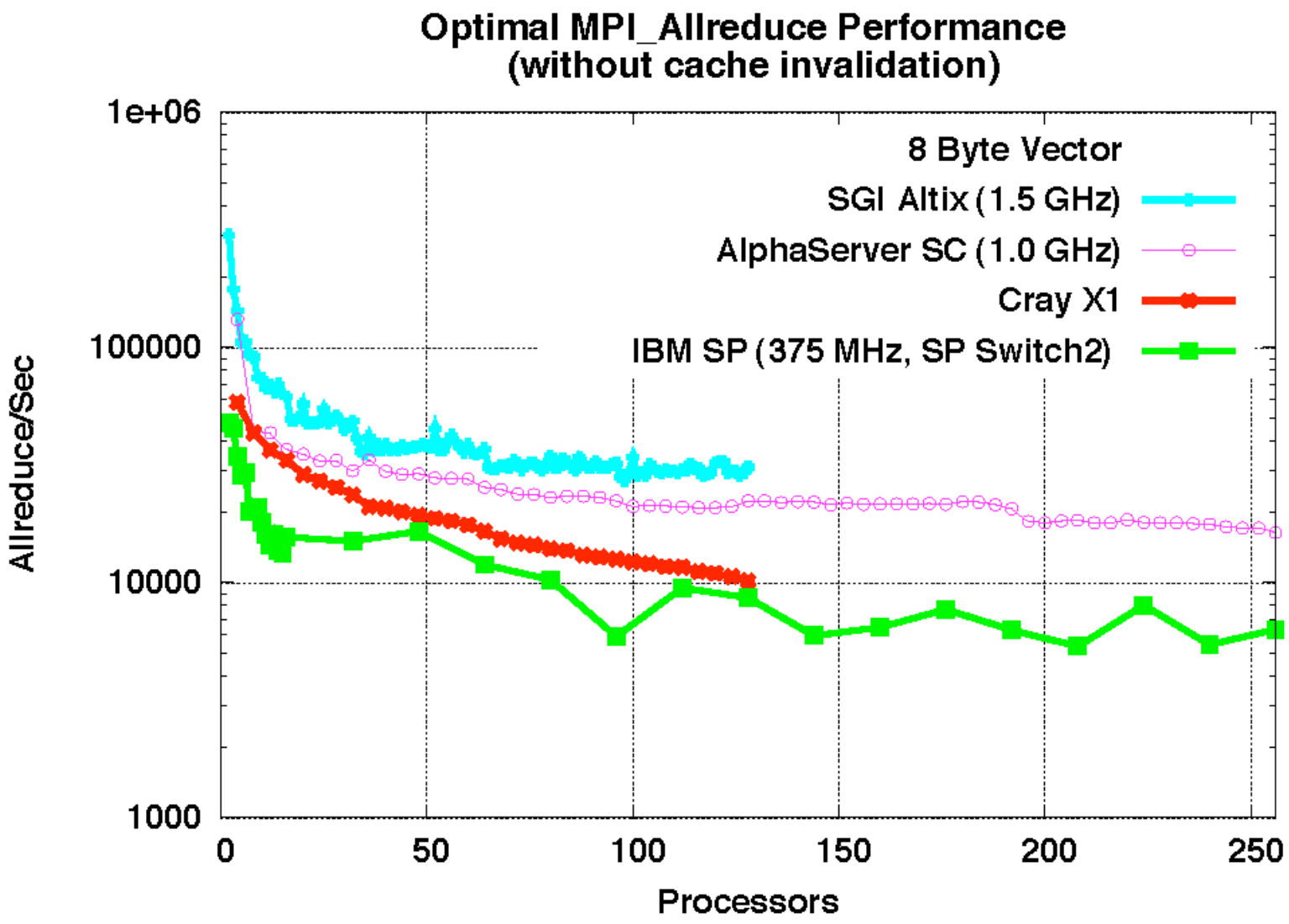
MPI_Allreduce: Cache Invalidation



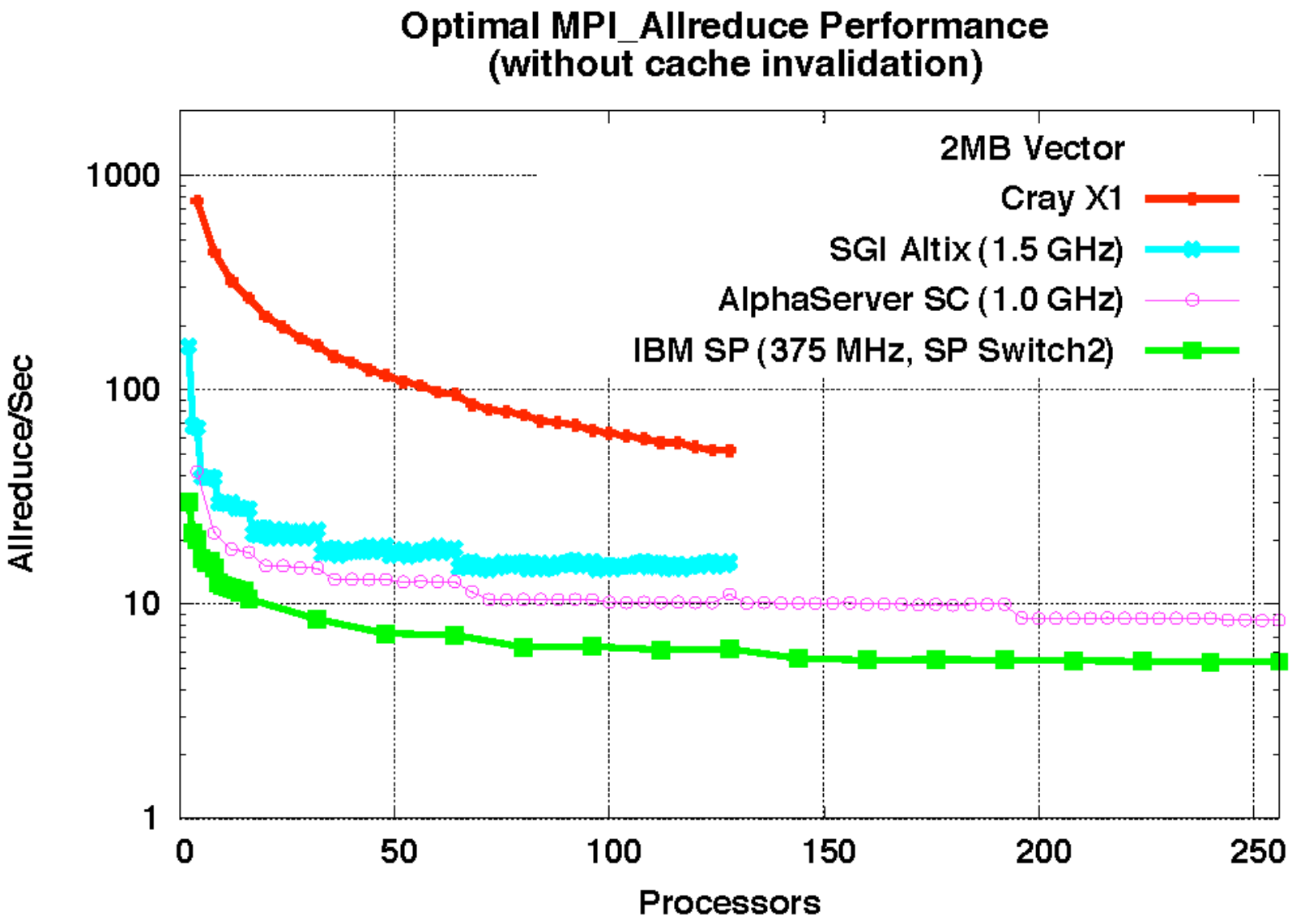
Allreduce: Optimal vs. MPI_Allreduce



MPI_Allreduce: Platform Comparison



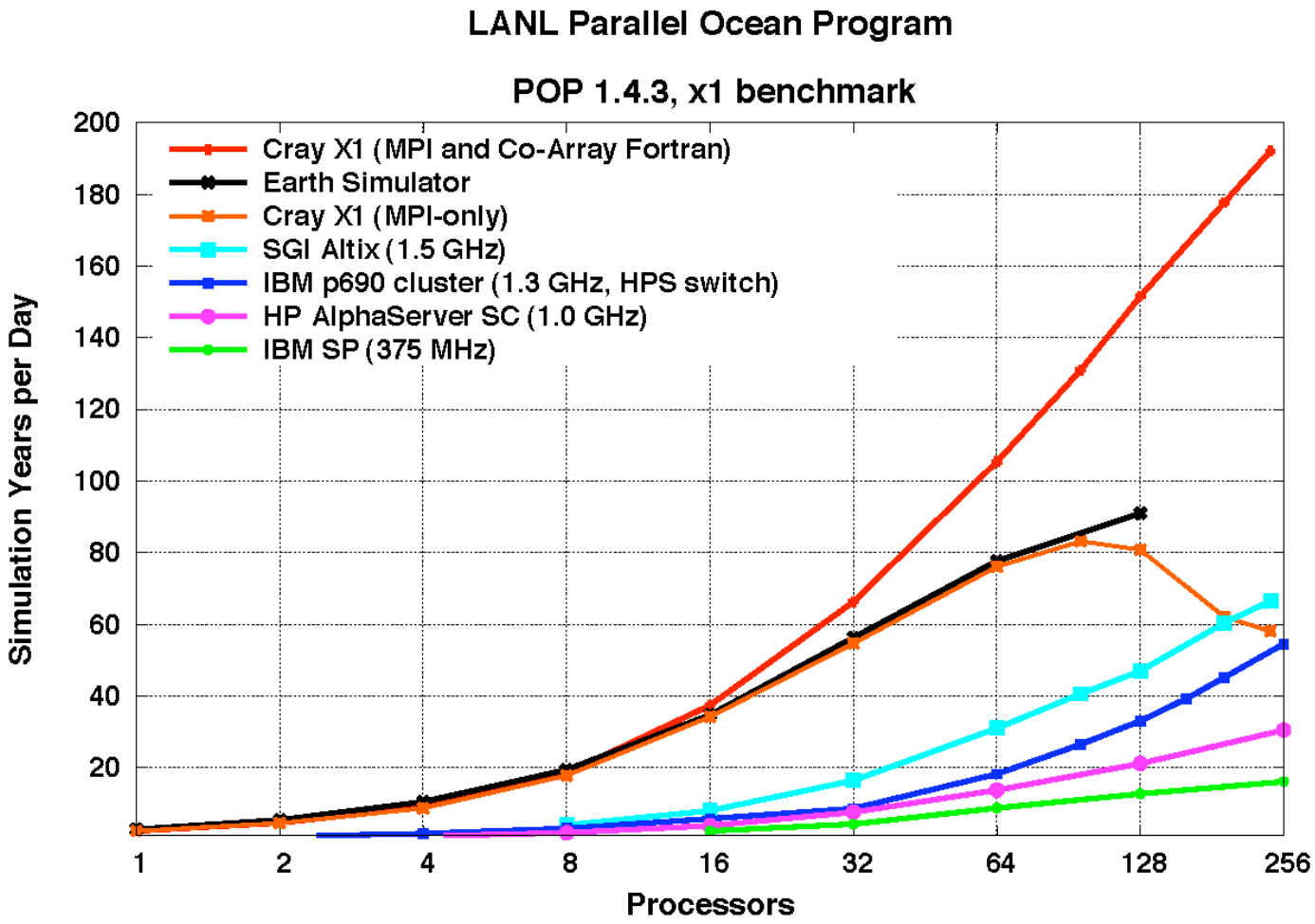
MPI_Allreduce: Platform Comparison



Parallel Ocean Program (POP)

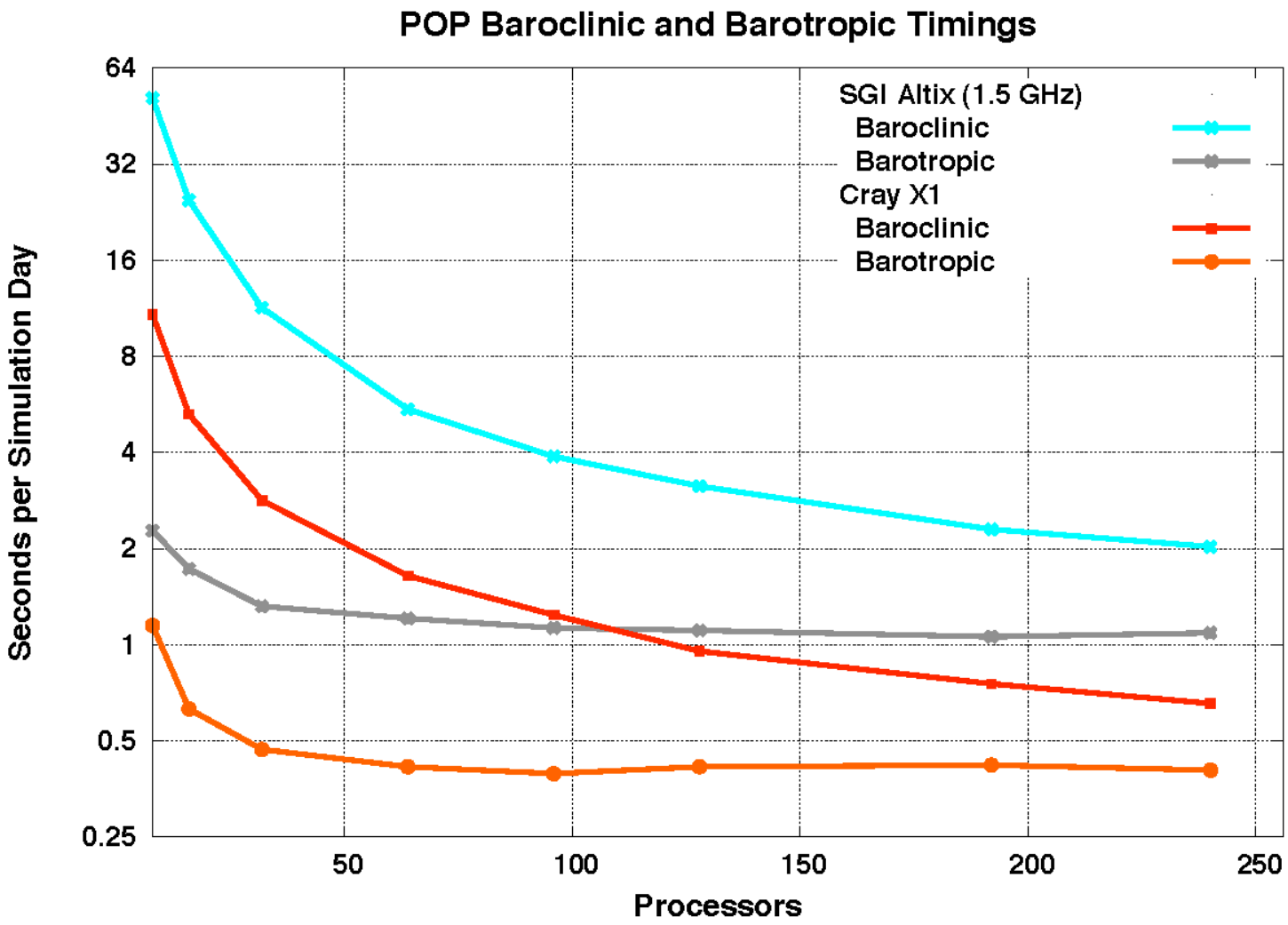
- Developed at Los Alamos National Laboratory. Used for high resolution studies and as the ocean component in the Community Climate System Model (CCSM).
- Two primary computational phases
  - Baroclinic: 3D with limited nearest-neighbor communication; scales well.
  - Barotropic: dominated by solution of 2D implicit system using conjugate gradient solves; scales poorly.
- One fixed size benchmark problem
  - One degree horizontal grid ("by one" or "x1") of size 320x384x40.
- Domain decomposition determined by grid size and 2D virtual processor grid. Results for a given processor count are the best observed over all applicable processor grids.

POP: Platform Comparisons



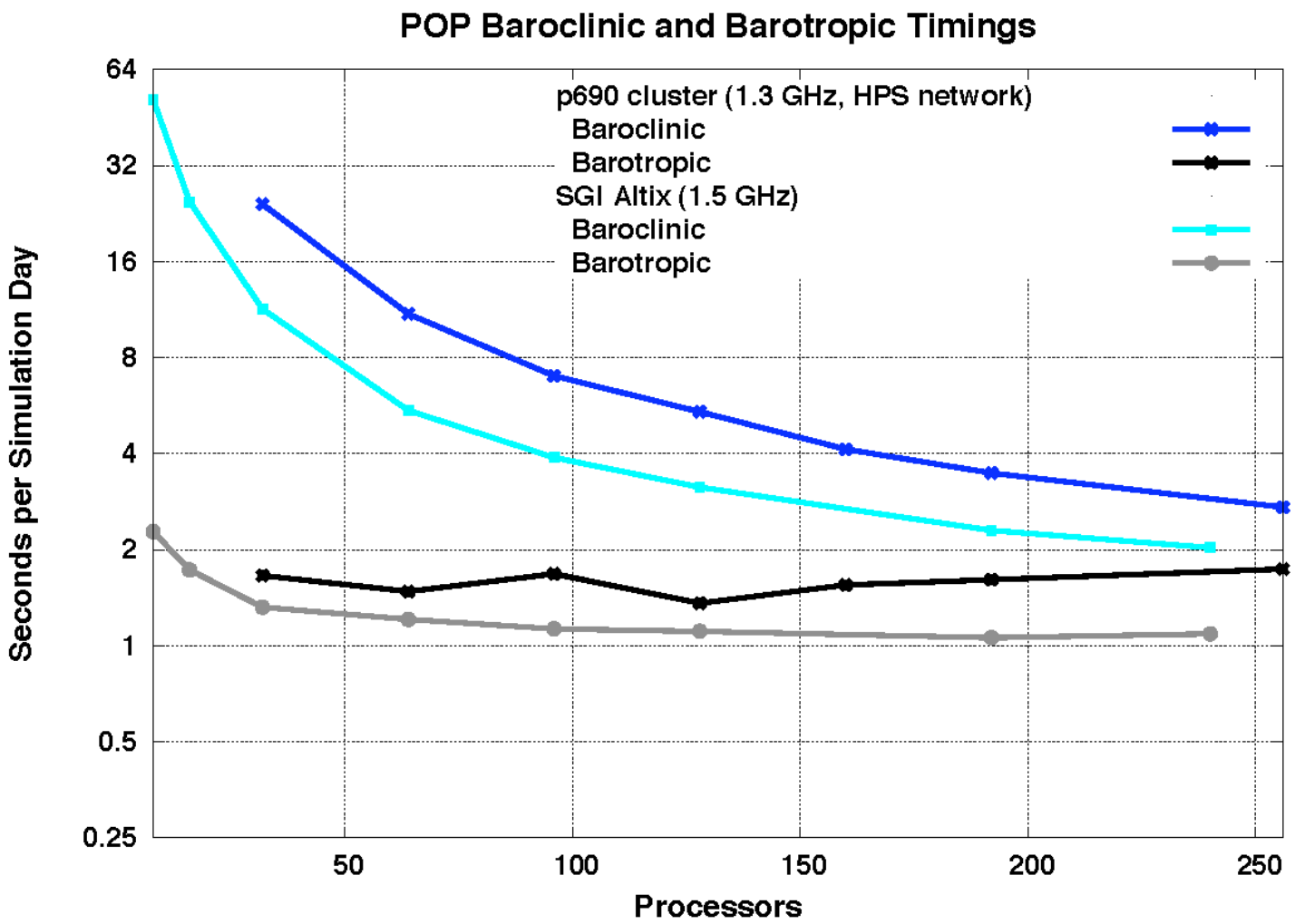
- Earth Simulator results courtesy of Dr. Y. Yoshida of CRIEPI.
- IBM SP results courtesy of Dr. T. Mohan of LBNL.

POP: Performance Diagnosis vs. X1



Altix is 3-4 times slower than X1 with similar scaling.

POP: Performance Diagnosis vs. p690

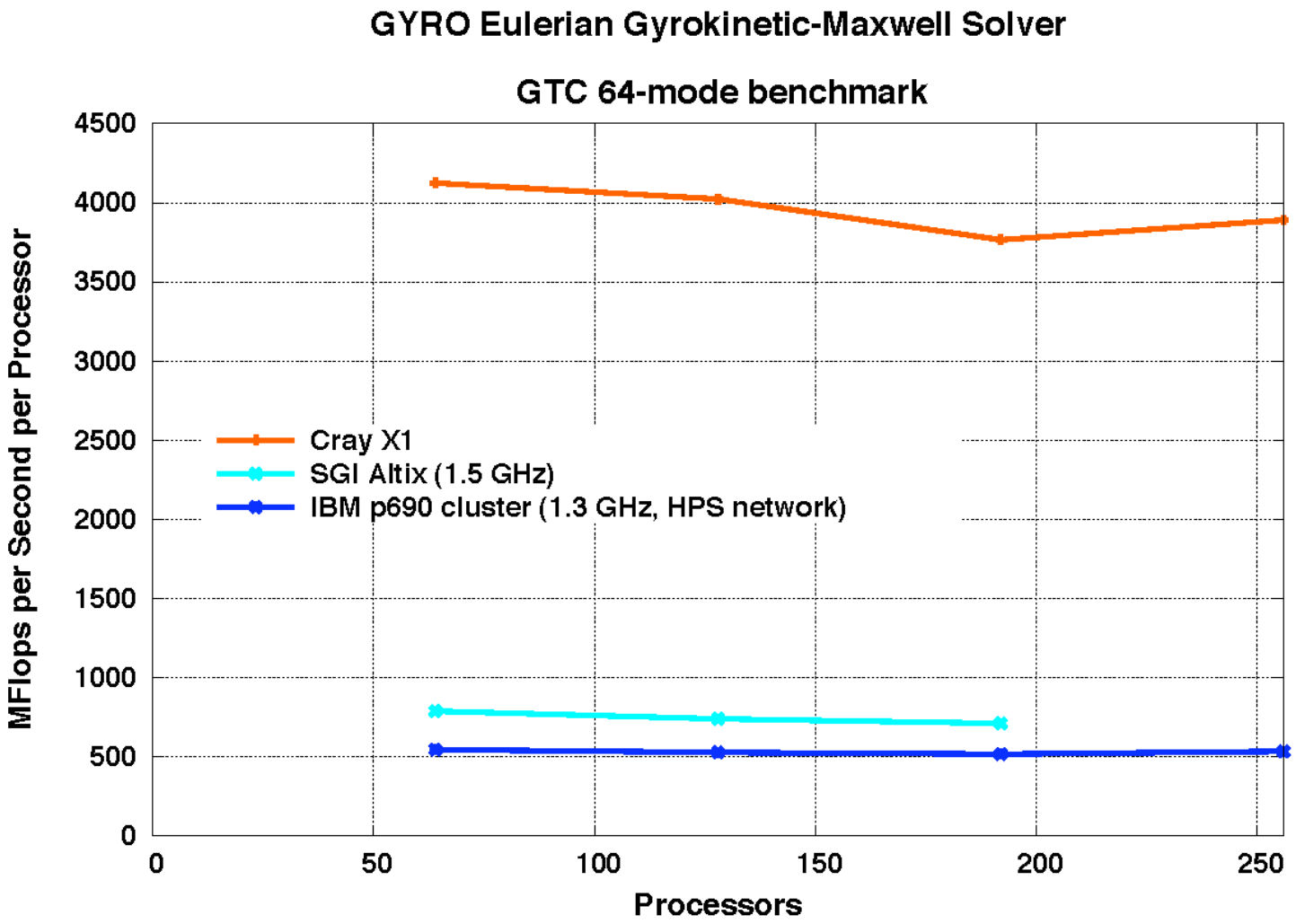


Altix is > 50% faster than p690 cluster with better scaling for barotropic.

GYRO

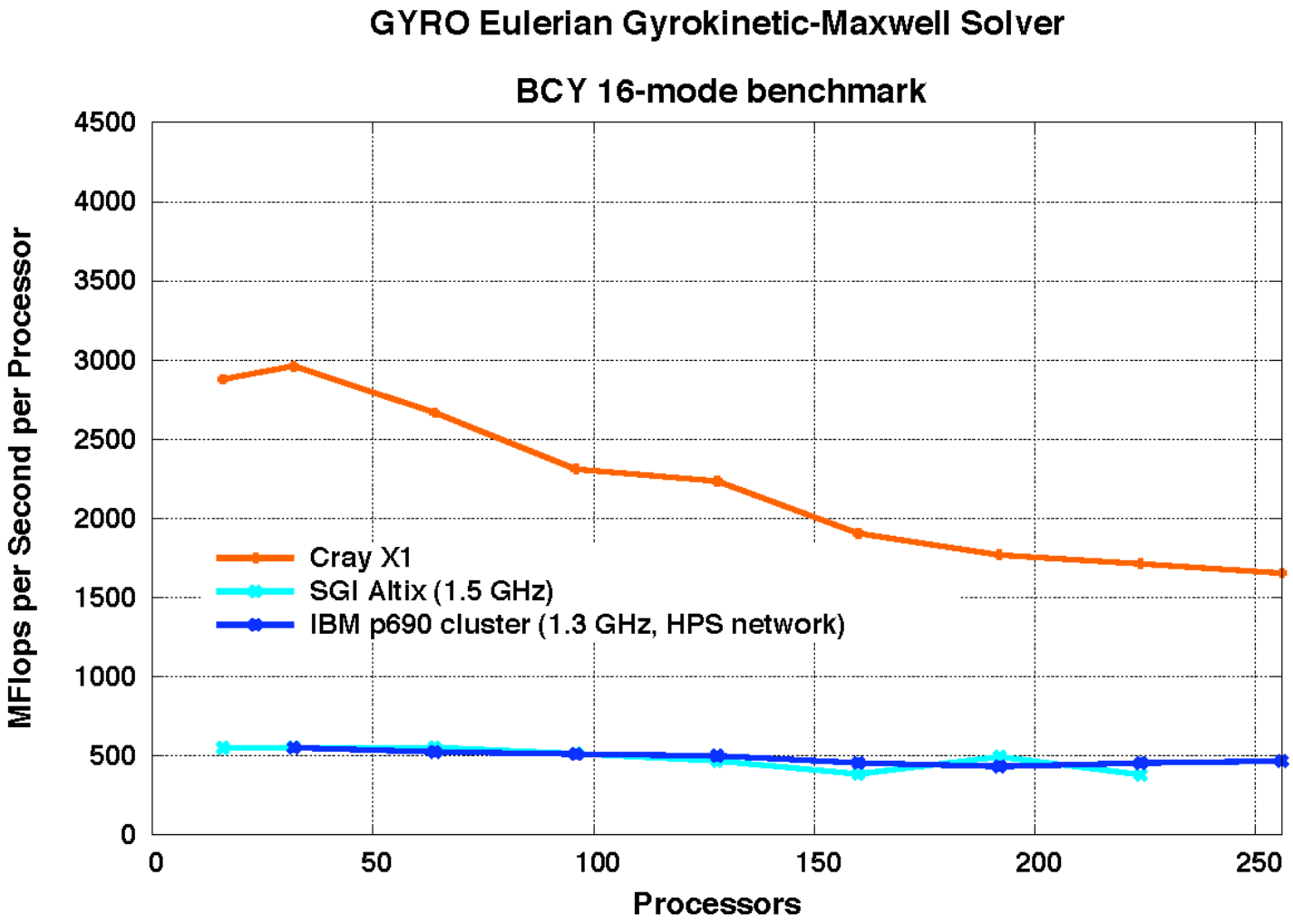
- GYRO is an Eulerian gyrokinetic-Maxwell solver developed by R.E. Waltz and J. Candy at General Atomics. It is used in the DOE SciDAC Fusion Energy project studying plasma microturbulence.
- Primary communication cost in this benchmark is calls to MPI_ALLTOALL to transpose distributed matrices.
- Two benchmark problems, both time dependent:
- 1. GTC.n64.500
  - 64-mode adiabatic electron case. It is run on multiples of 64 processors. Duration is 3 simulation seconds, representing 100 timesteps.
- 2. BCY.n16.b.25
  - 16-mode electromagnetic case. It is run on multiples of 16 processors. Duration is 8 simulation seconds, representing 1000 timesteps.

GYRO Computation Rate: Large Benchmark



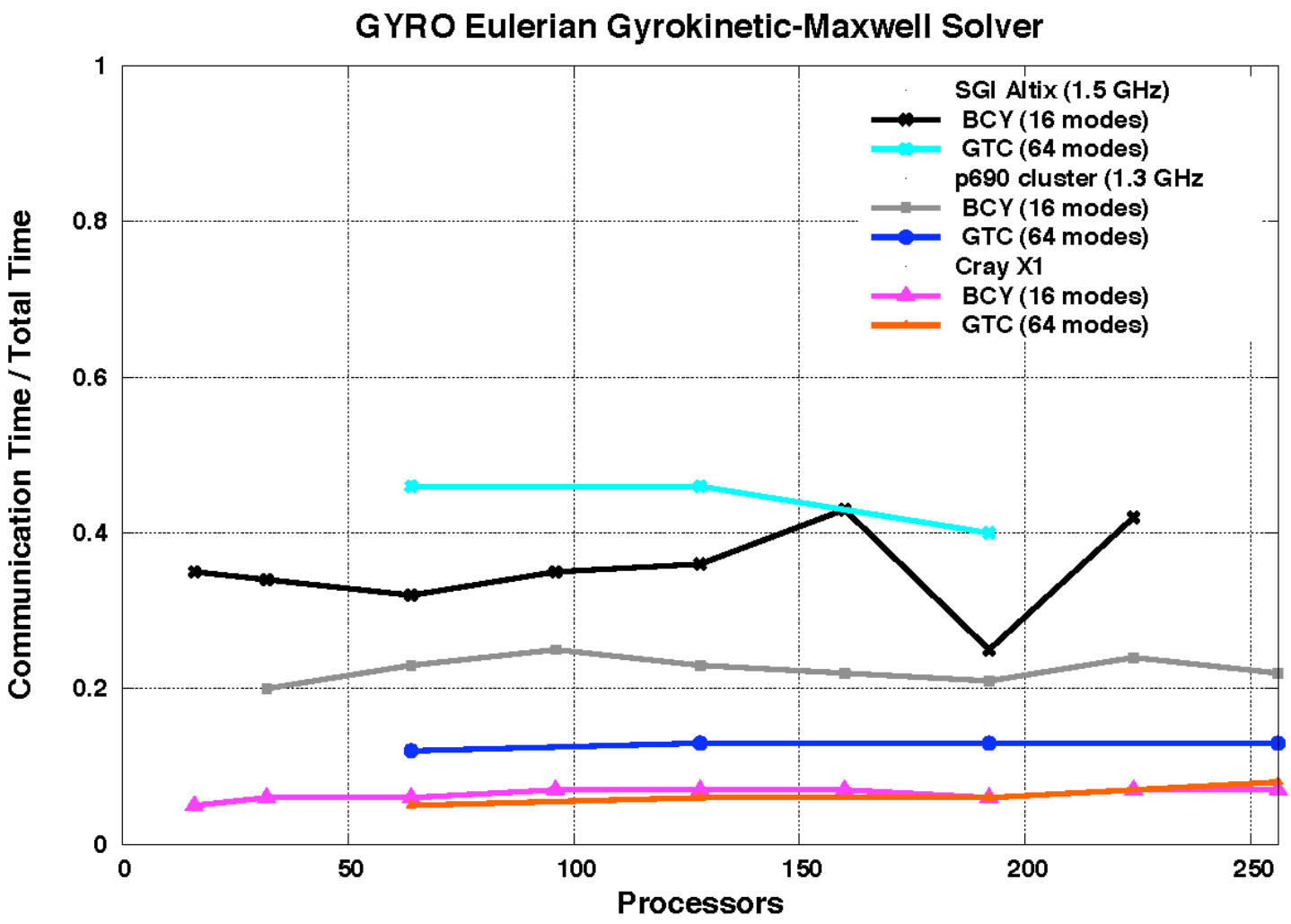
Altix is 50% faster than p690 and 5 times slower than X1.

GYRO Computation Rate: Small Benchmark



Altix does not show typical 50% advantage over p690 cluster.

GYRO Communication Overhead



Communication overhead high on the Altix compared to the Cray and IBM systems.

Other Applications

- **AORSA**
  - AORSA solves for the wave electric field and heating in a 3-D stellerator plasma heated by radio frequency waves using an all orders spectral algorithm. (Results courtesy of F. Jaeger, ORNL.)

- **NWCHEM**
  - NWCHEM is a computational chemistry package that is designed to run on high-performance parallel supercomputers as well as conventional workstation clusters. (Results courtesy of E. Apra, PNNL.)

AORSA Performance

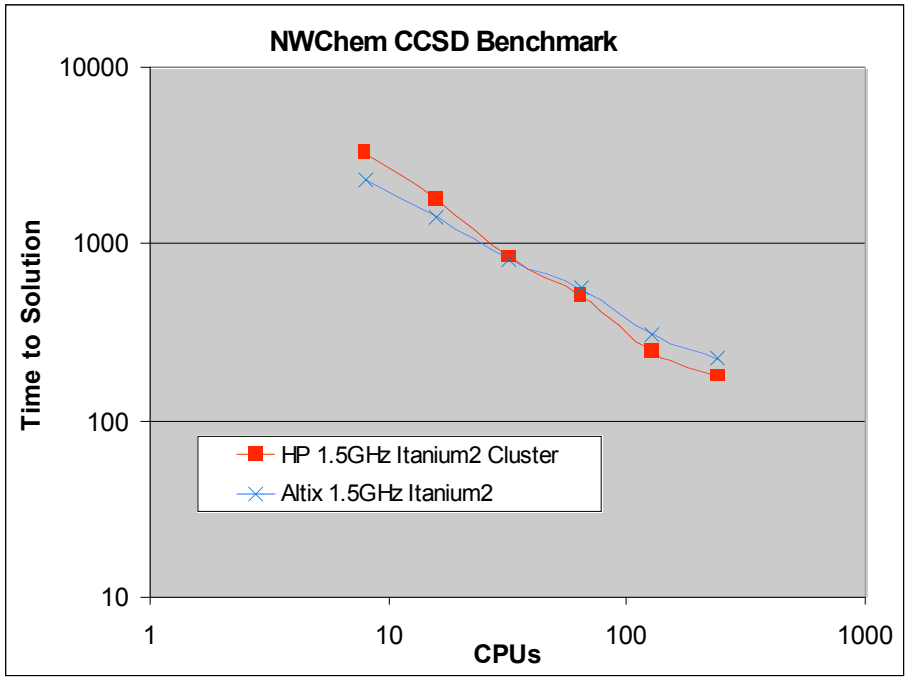
128 processors	Seaborg (IBM POWER3)	Cheetah-f (IBM POWER4)	RAM	Phoenix
Matrix load	207.7	37.3	10.2	13.7
ScaLAPACK	8.5	3.9	4.2	2.0
Plasma current	151.9	45.7	12.4	16.4
Absorption (wdot)	86.3	39.7	37.2	31.4
Total cpu time	456.1	127.4	64.2	64.1

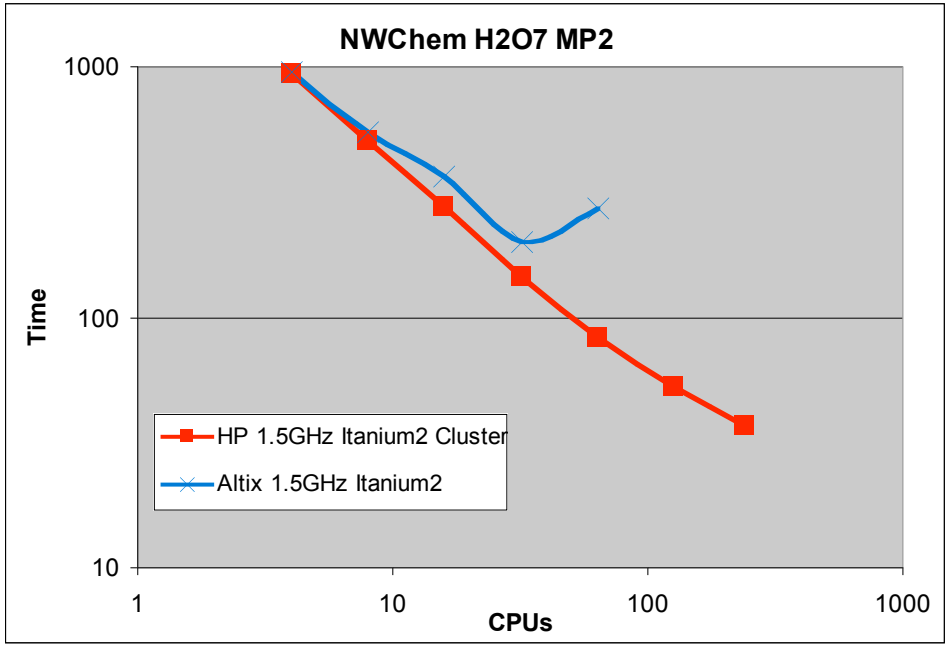
64 processors	Cheetah-f (IBM POWER4)	RAM	Phoenix
Matrix load	74.4	34.1	27.4
ScaLAPACK	7.3	7.8	3.1
Plasma current	76.2	20.8	27.1
Absorption (wdot)	52.8	56.9	50.4
Total cpu time	211.4	122.1	108.5

RAM is the ORNL SGI Altix. Phoenix is the ORNL Cray X1.

AORSA does not (yet) vectorize well. Performance on Altix is comparable to X1.

NWChem





Out of core solvers in NWCHEM require high performance file system. Itanium2 cluster at PNNL was configured to satisfy this requirement. I/O on ORNL Altix inhibits scalability for H207 MP2 benchmark.

Conclusions

- Itanium2 processor and C-brick memory controller perform well compared to other nonvector systems.
- MPI benchmarks show excellent latency and good bandwidth.
- Low latency enables good scaling for latency sensitive applications.
- Performance of (some) MPI collectives negatively impact performance of (some) bandwidth sensitive applications.
- Care is required in system configuration if high performance I/O is important.
- Transitioning to new compilers (7.1 to 8.1) will require revisiting performance studies. Early tests show 10-15% performance improvement and much better hybrid MPI/OpenMP performance.

For More Information …

- http://www.csm.ornl.gov/evaluation
- http://www.csm.ornl.gov/~dunigan/sgi
- *Evaluation of the SGI Altix 3700 at Oak Ridge National Laboratory,* J. Ahrens, P. Agarwal, D. Bauer, R. Bleck, P. Chandramohan, T. Dunigan, G. Fann, M. Guest, F. Jaeger, G. Kora, L. Oliker, R. Latham, V. Meunier, R. Ross, N. Samatova, A. Shoshani, B. Sumpter, R. Thakur, R. Toedte, J. Vetter, P. Worley, D. Xu, ORNL Technical Report.