

Early Evaluation of the Cray X1

Part 1.5

Patrick H. Worley
Thomas H. Dunigan Jr.
Mark R. Fahey
James B. White III

Oak Ridge National Laboratory

SC03

Cray Exhibitor Booth
November 18, 2003
Phoenix Convention Center
Phoenix, Arizona

Outline

Part 1 - Technical Paper, 2:00PM, Wednesday

- Standard kernel benchmarks (unmodified)
- Parallel application success stories
 - POP ocean code
 - GYRO fusion code

Part 1.5 - Cray Booth, 11:00AM, Tuesday

- Parallel application details
 - POP ocean code
 - GYRO fusion code
- Kernel measurements of communication performance
 - COMMTEST
 - HALO

Acknowledgements

- Research sponsored by 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.

Evaluation of Early Systems

A project that attempts to evaluate *quickly* the promise of “early” (possibly immature) systems:

- Verifying advertised functionality and performance
- Quantifying performance impact of unique system characteristics
- Providing guidance to (early) users
 - What performance to expect
 - Performance quirks and bottlenecks
 - Performance optimization tips

Evaluation Methodology

“Measure early, measure often, analyze just in time”

- 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 ways of using system, evaluating *both* traditional and alternative programming paradigms
 - Collecting data with *both* standard and custom benchmarks

Phoenix

Cray X1 with 64 SMP nodes

- 4 Multi-Streaming Processors (MSP) per node
- 4 Single Streaming Processors (SSP) per MSP
- Two 32-stage 64-bit wide vector units running at 800 MHz and one 2-way superscalar unit running at 400 MHz per SSP
- 2 MB Ecache per MSP
- 16 GB of memory per node for a total of 256 processors (MSPs), 1024 GB of memory, and 3200 GF/s peak performance.



OAK RIDGE NATIONAL LABORATORY
U. S. DEPARTMENT OF ENERGY


UT-BATTELLE

Other Platforms

- Earth Simulator: 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: 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 an SP Switch2 with two to eight network adapters per node.
- IBM SP at the National Energy Research Supercomputer Center (NERSC): 184 Nighthawk II 16-way SMP nodes (375MHz POWER3-II) and an SP Switch2 with two network adapters per node.
- SGI Altix 3700 at ORNL: 2 128-way SMP nodes and NUMAflex fat-tree interconnect. Each processor is a 1.5 GHz Itanium 2 with a 6 MB L3 cache
- SGI Origin 3000 at Los Alamos National Laboratory (LANL): 512-way SMP node. Each processor is a 500 MHz MIPS R14000.

Caveats

- These are EARLY results (even on the Cray after 6 months), resulting from sporadic benchmarking on evolving system software and hardware configurations.
- Performance characteristics are still changing, due to continued evolution of OS and compilers and libraries.

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)
- Ported to the Earth Simulator by Dr. Yoshikatsu Yoshida of the Central Research Institute of Electric Power Industry (CRIEPI).
- Initial port to the Cray X1 by John Levesque of Cray, using Co-Array Fortran for conjugate gradient solver.
- X1 and Earth Simulator ports merged and modified by Pat Worley and Trey White of Oak Ridge National Laboratory.
- Optimization on the X1 ongoing.

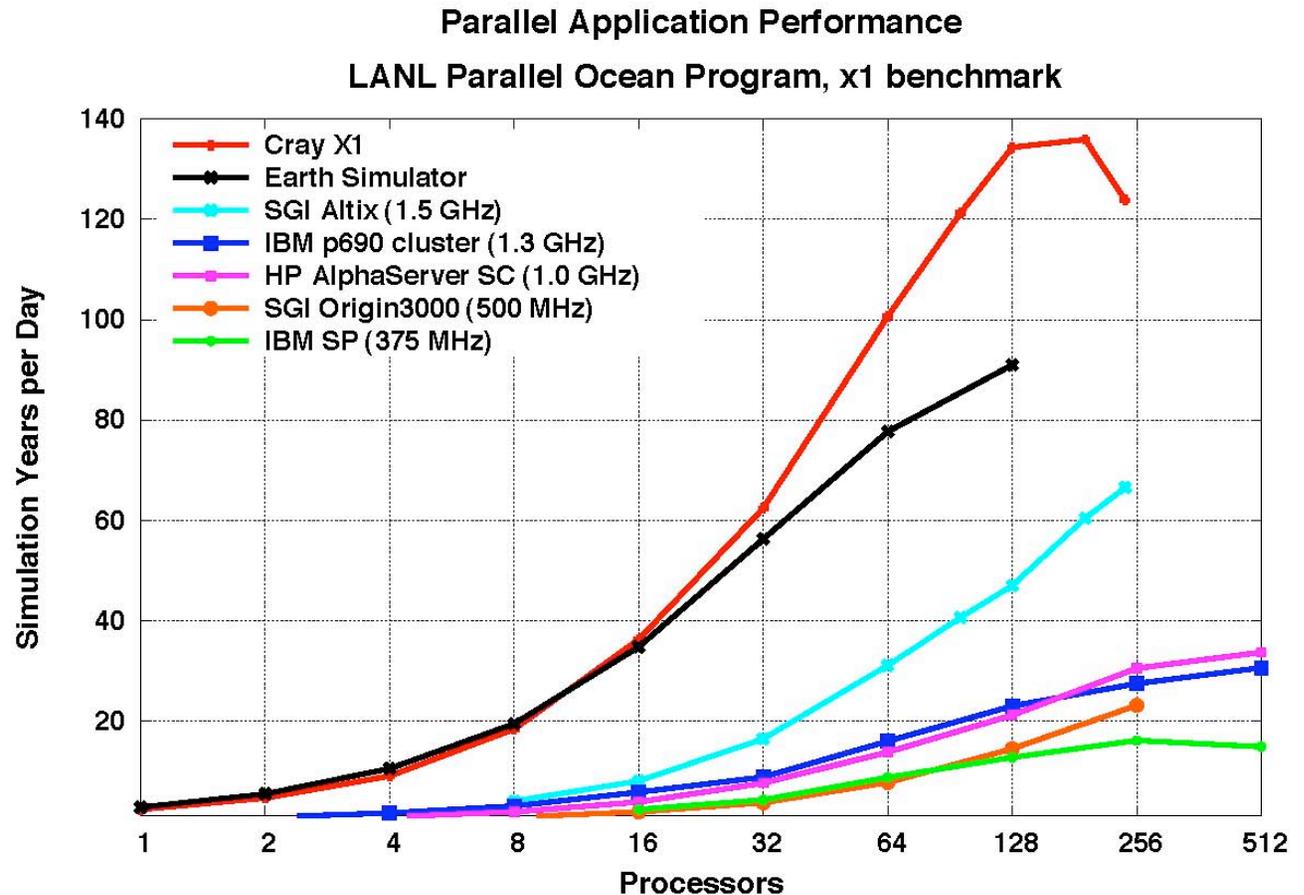
POP Experiment Particulars

- 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 benchmark problem size
 - 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.

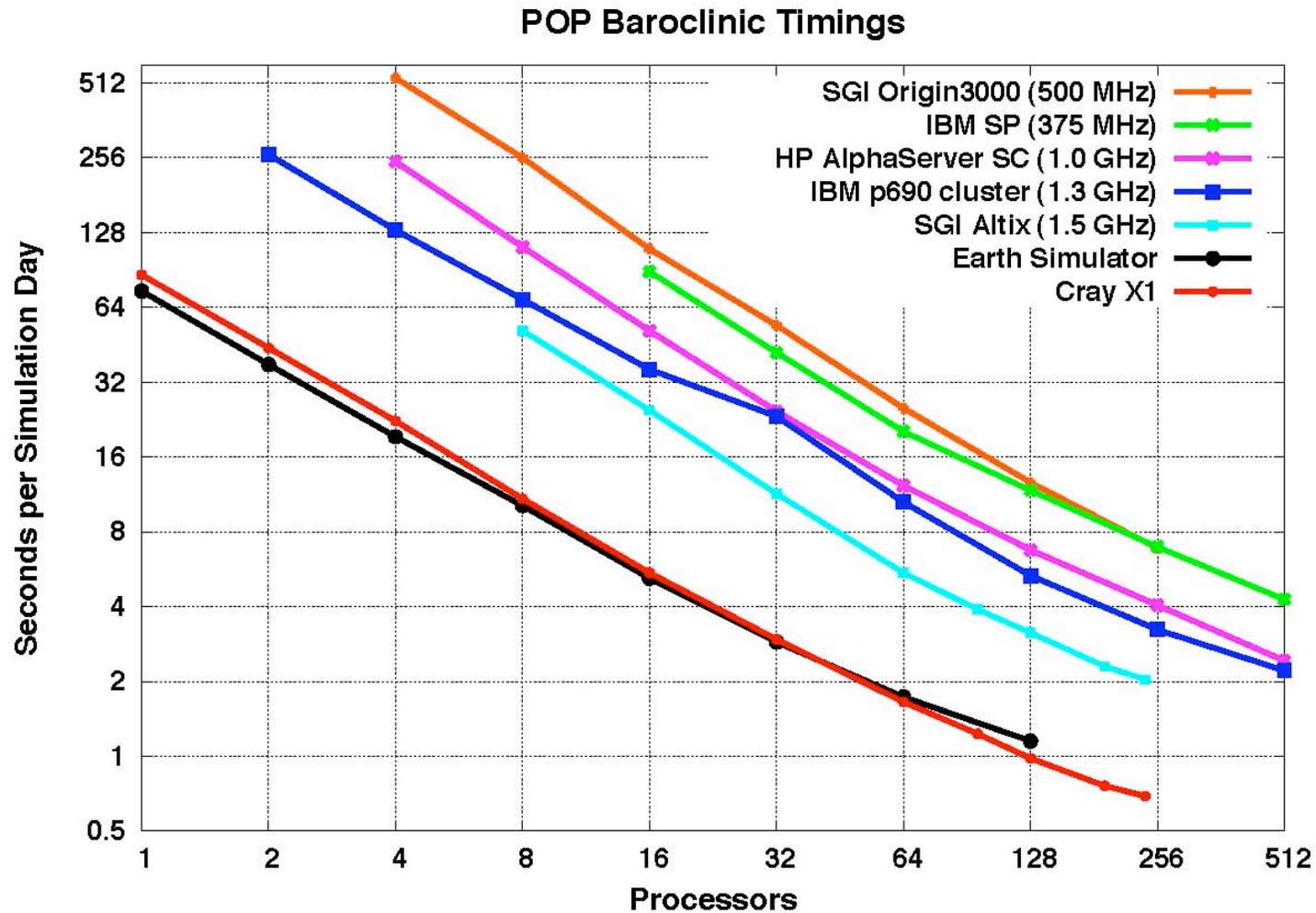
POP Platform Comparison

Comparing performance and scaling across platforms.

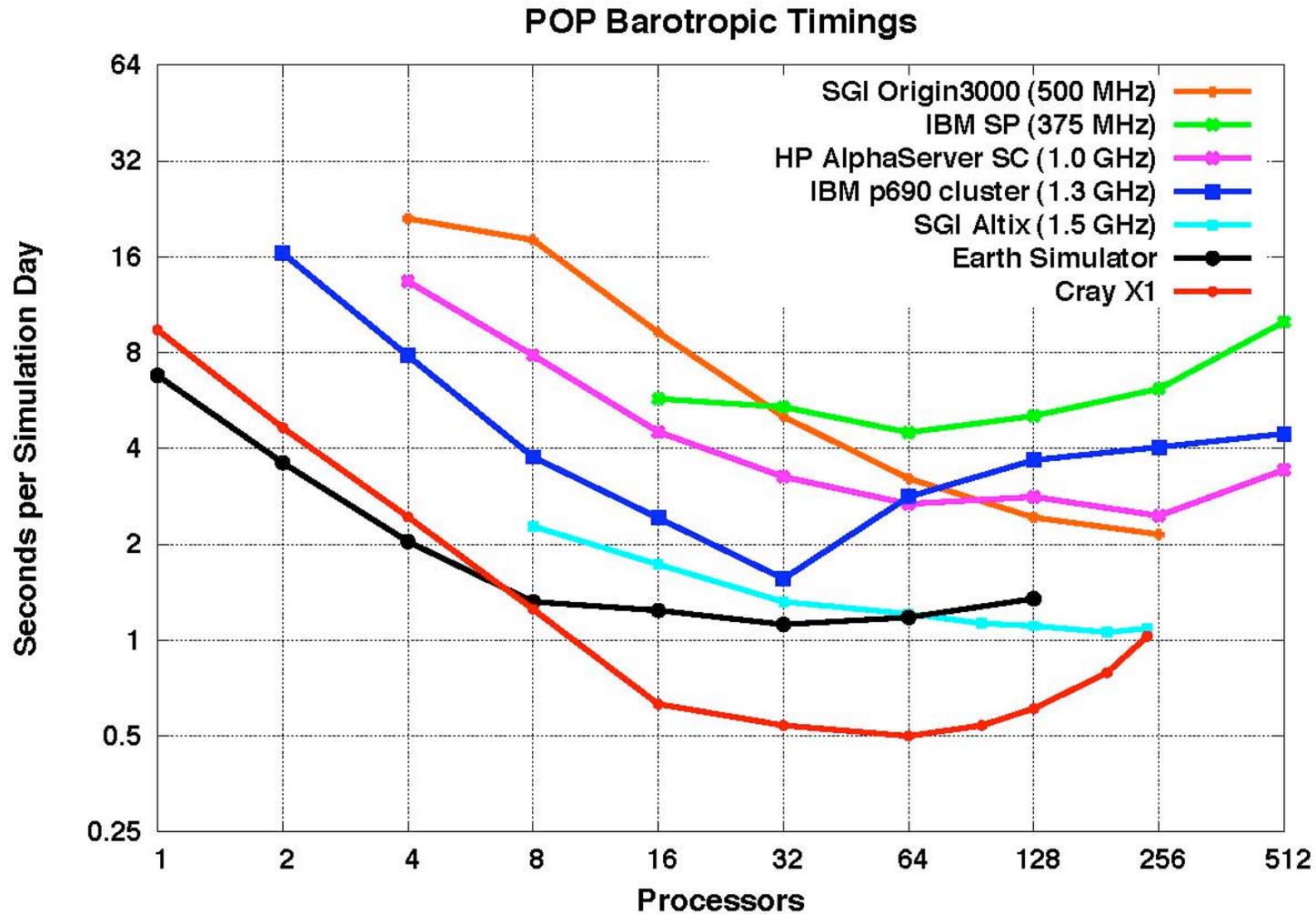
- Earth Simulator results courtesy of Dr. Y. Yoshida of the Central Research Institute of Electric Power Industry (CRIEPI).
- SGI Origin results courtesy of Dr. P. Jones of LANL.
- IBM SP results courtesy of Dr. T. Mohan of Lawrence Berkeley National Laboratory (LBNL)



POP Performance Diagnosis: Baroclinic



POP Performance Diagnosis: Barotropic



POP Performance Diagnosis

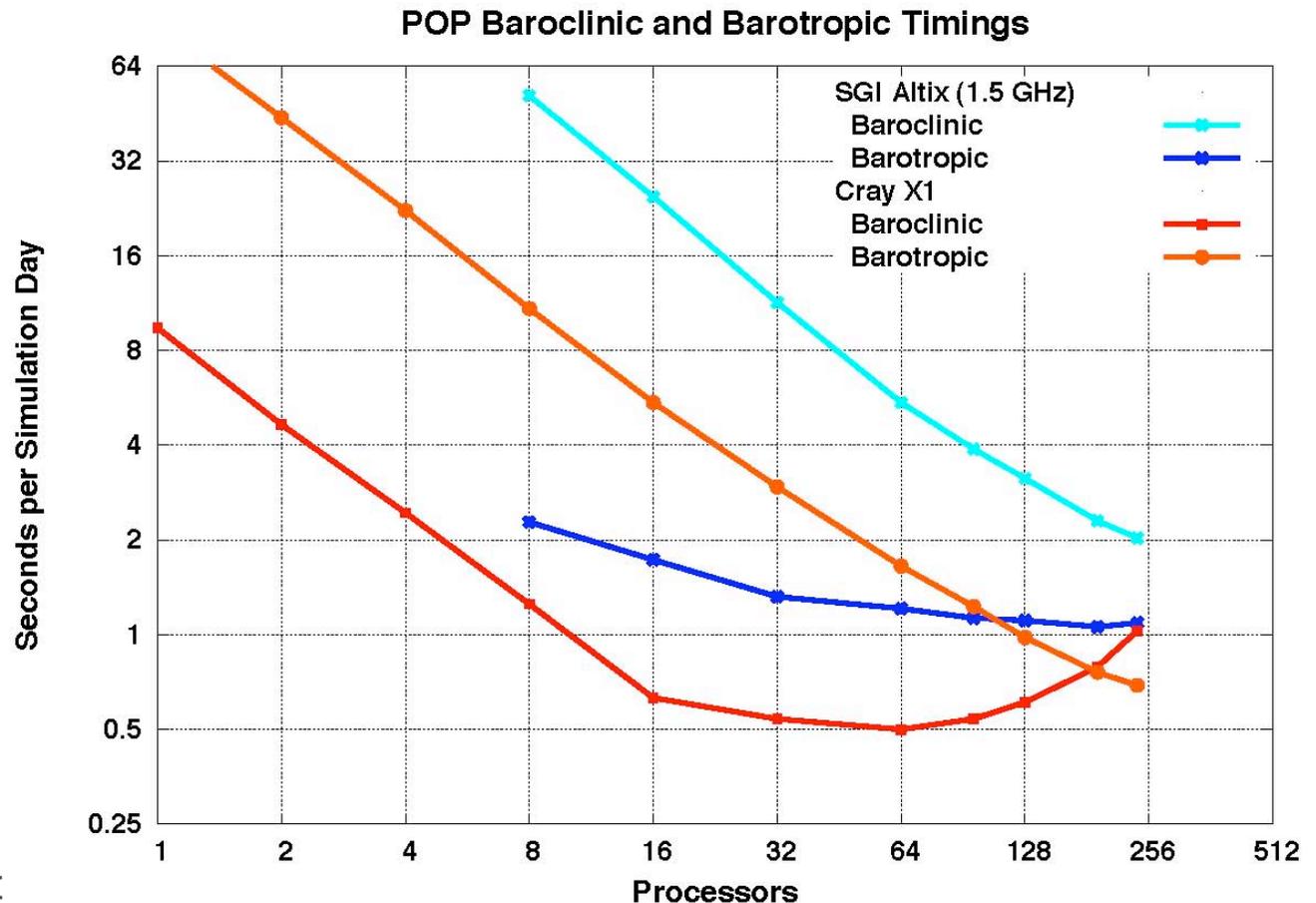
Cray X1

Communication-bound for more than 192 processors, with communication costs increasing. Communication algorithms known to have scaling problems, and alternatives being developed.

SGI Altix

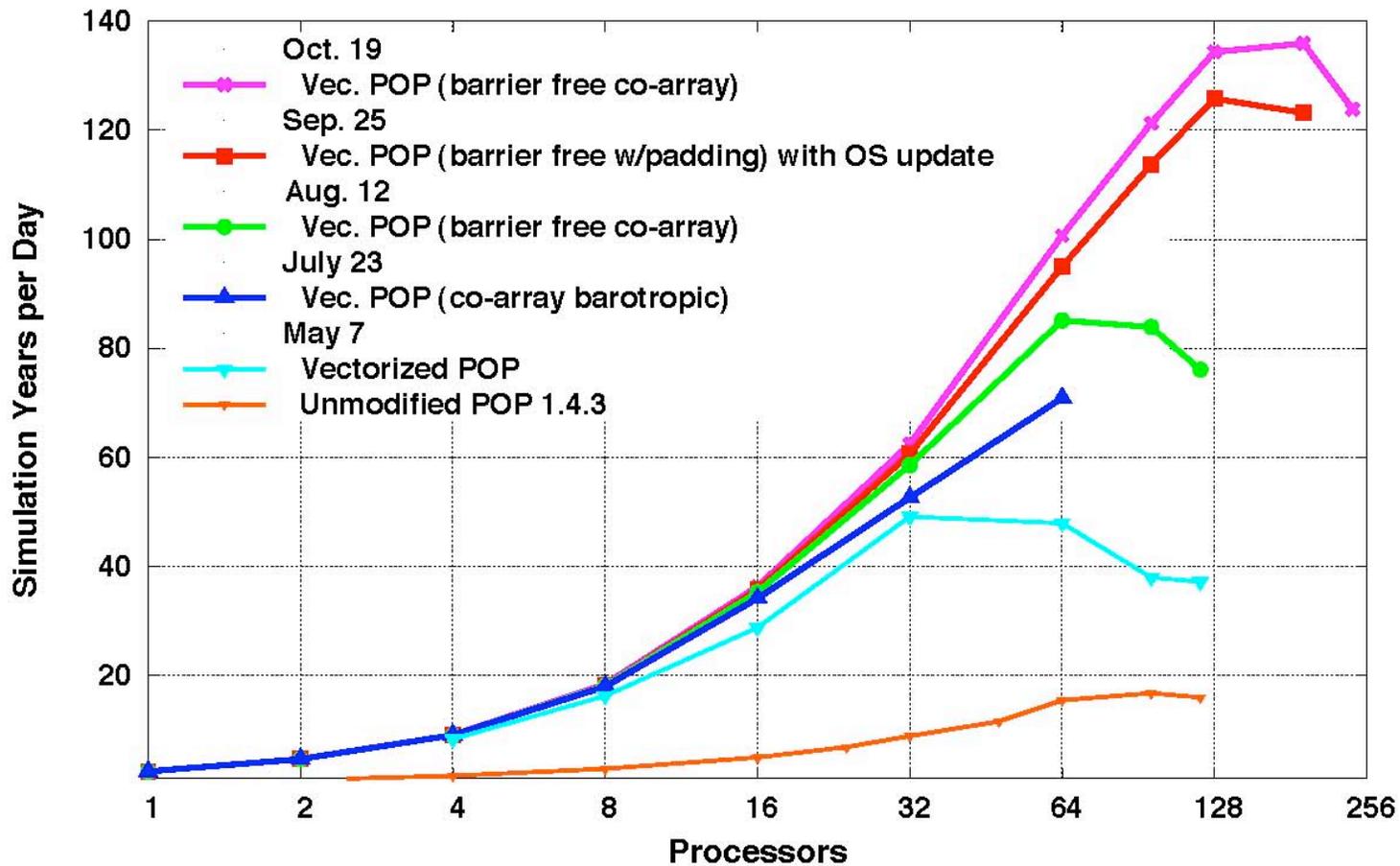
Not yet communication bound. Using MPI point-to-point and collectives for barotropic. Initial experiments SHMEM with do not show significant improvement.

OAK RIDGE NATIONAL LABORATORY
U. S. DEPARTMENT OF ENERGY



POP Performance Evolution on the X1

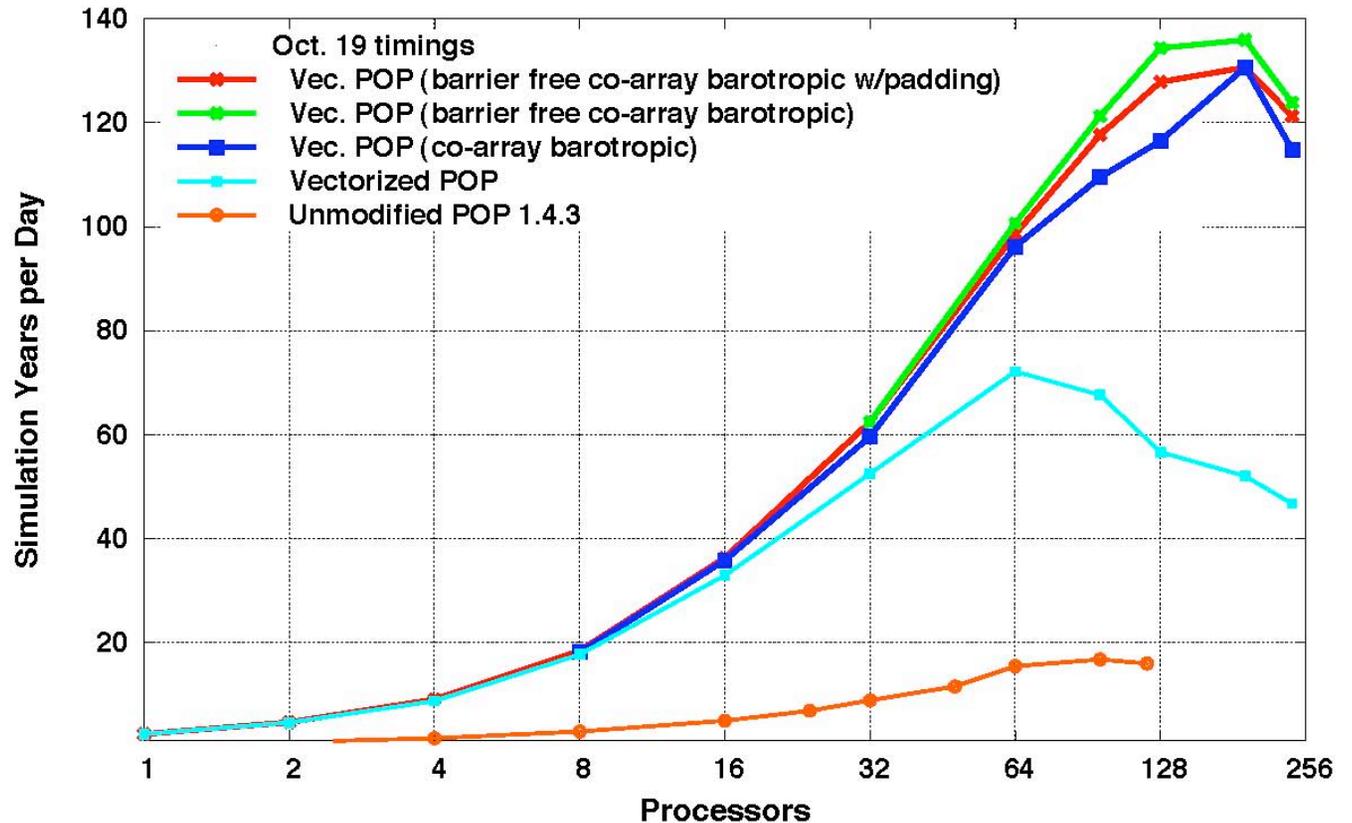
LANL Parallel Ocean Program
POP 1.4.3, x1 benchmark



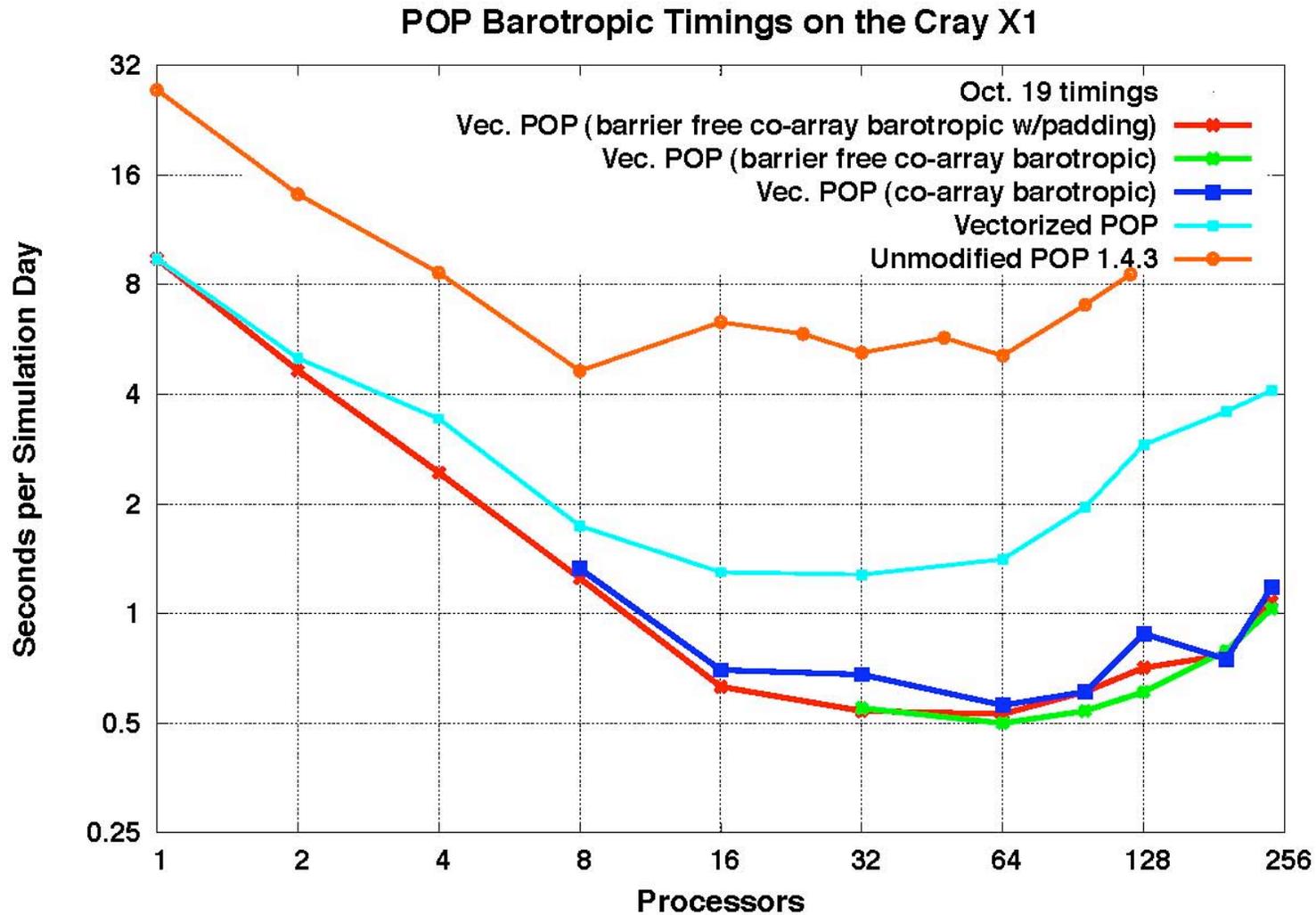
POP Implementation Comparison on the X1

Much of recent algorithm development driven by OS performance problems. Once OS problems solved, algorithmic differences less significant. MPI performance still poor for latency sensitive Algorithms, and restructuring for vectorization still vital.

LANL Parallel Ocean Program
POP 1.4.3, x1 benchmark

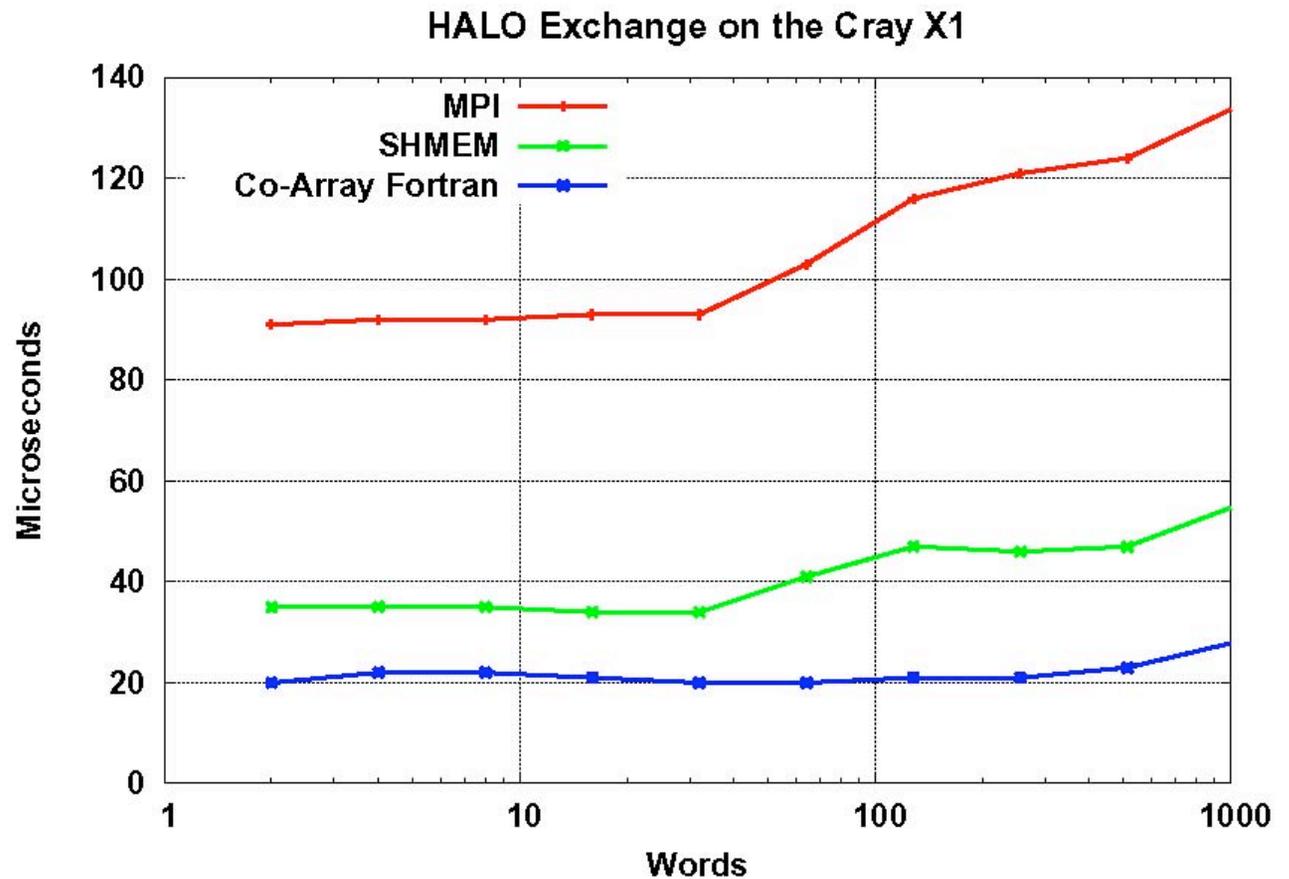


POP Implementation Comparison on the X1



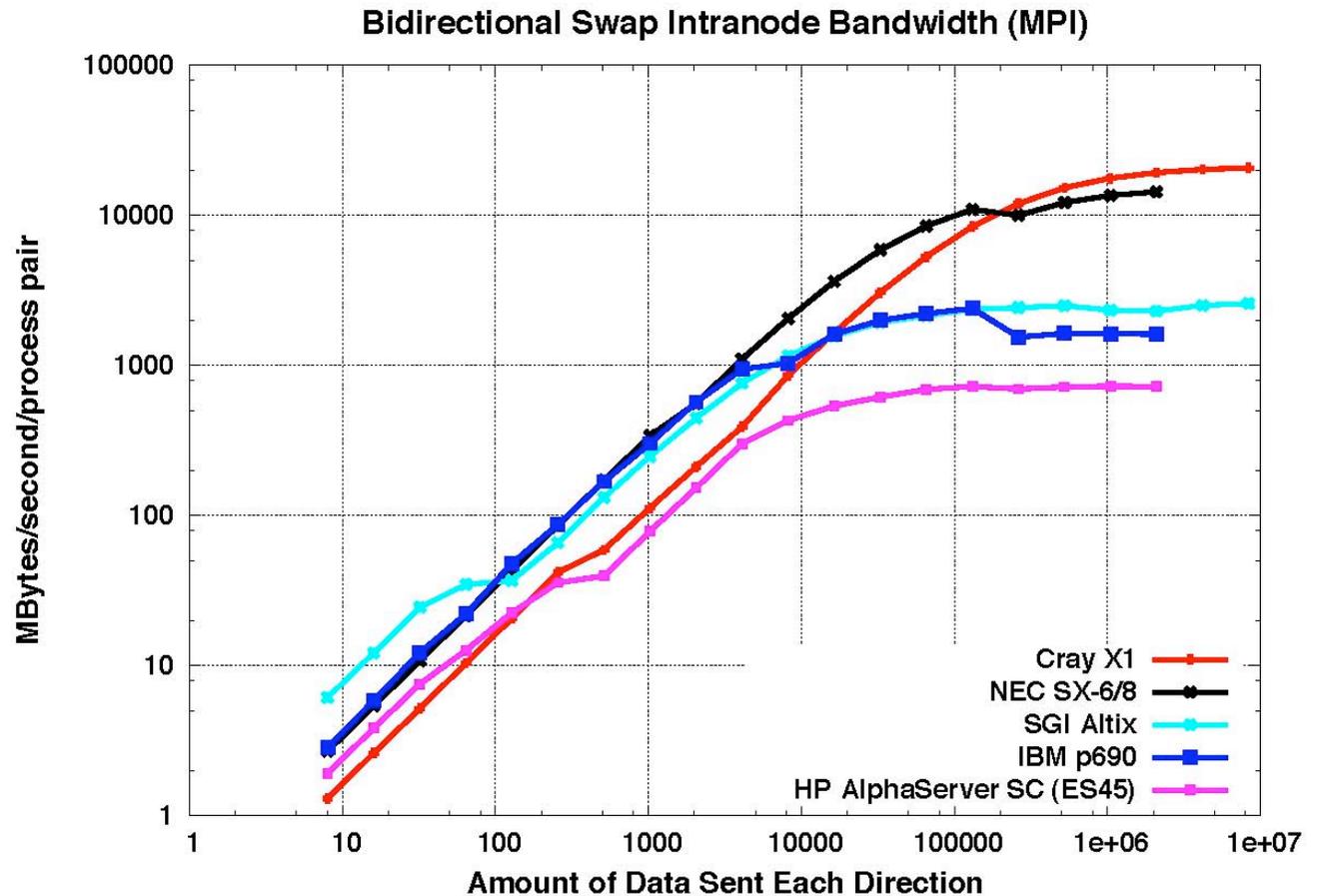
HALO Paradigm Comparison

Comparing performance of MPI, SHMEM, and Co-Array Fortran implementation of Allan Wallcraft's HALO benchmark on 16 MSPs. SHMEM and Co-Array Fortran are substantial performance enhancers for this benchmark.



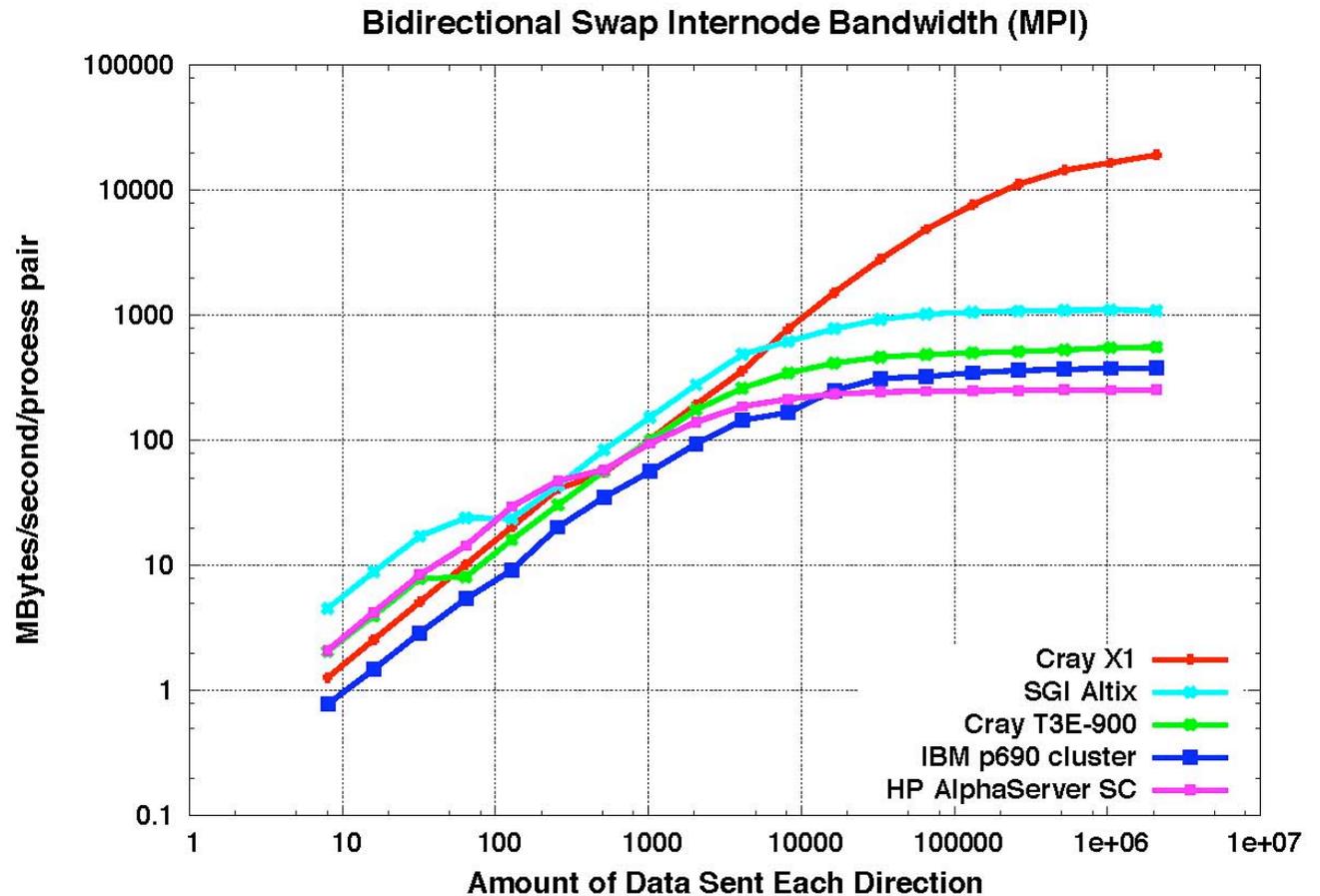
COMMTEST SWAP Benchmark

Comparing performance of SWAP for different platforms. Experiment measures bidirectional bandwidth between two processors in the same SMP node.



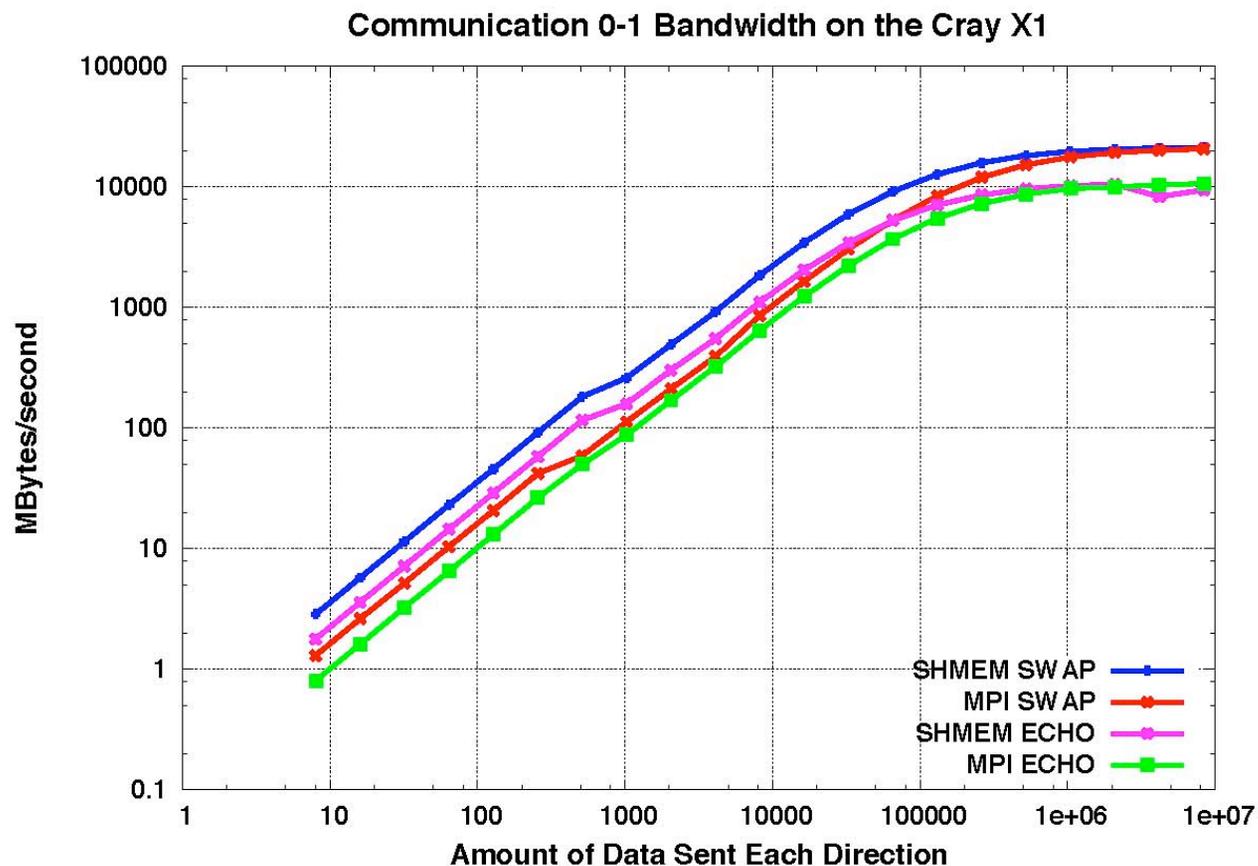
COMMTEST SWAP Benchmark

Comparing performance of SWAP for different platforms. Experiment measures bidirectional bandwidth between two processors in different SMP nodes.



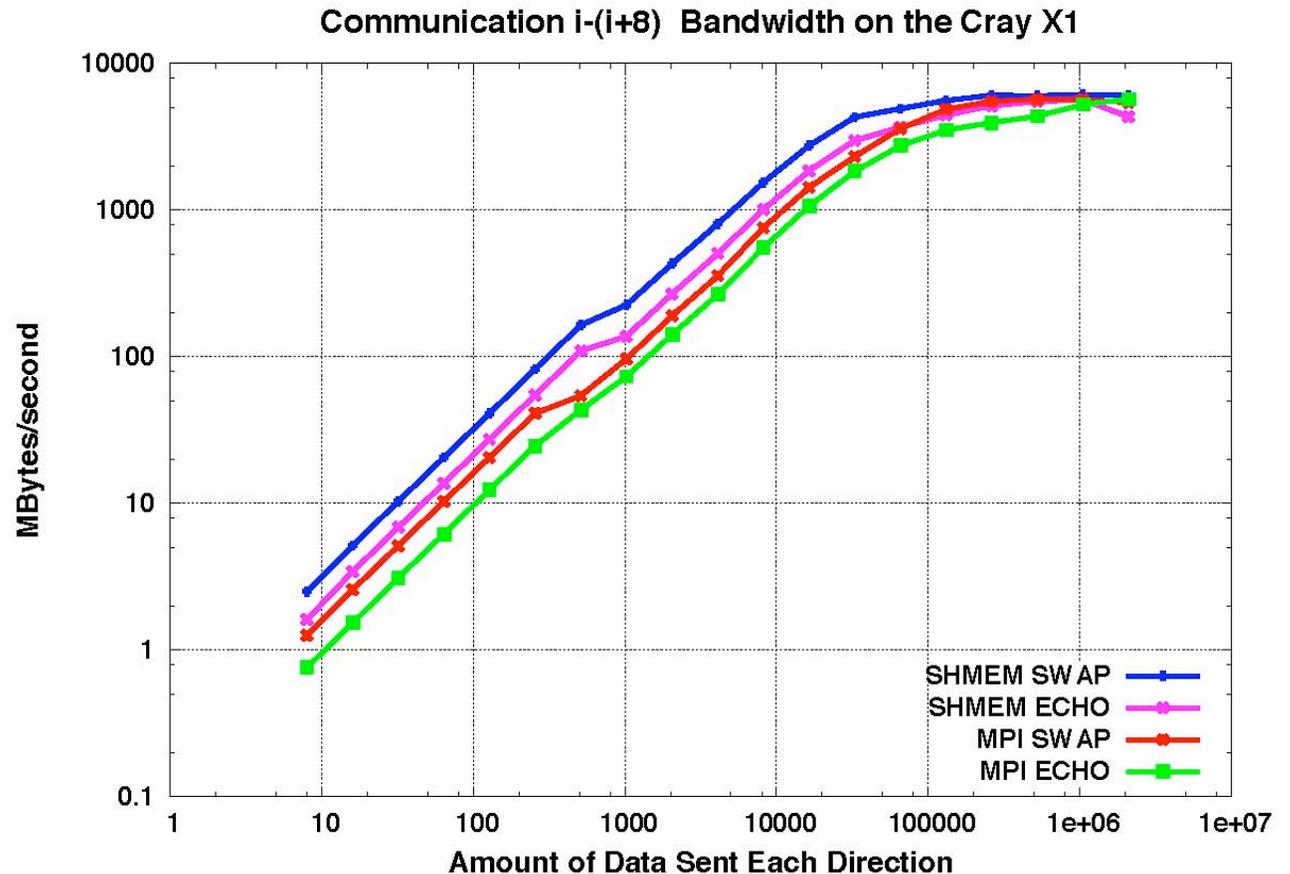
MPI vs. SHMEM 0-1 Comparison on X1

Comparing MPI and SHMEM performance for 0-1 experiment, looking at both SWAP (bidirectional bandwidth) and ECHO (unidirectional bandwidth). SHMEM performance is better for all but the largest messages.



MPI vs. SHMEM $i-(i+8)$ Comparison on X1

Comparing MPI and SHMEM performance for $i-(i+8)$ experiment, looking at both SWAP (bidirectional bandwidth) and ECHO (unidirectional bandwidth). Again, SHMEM performance is better for all but the largest messages.



POP Summary

- Using CRIEPI vectorization ...
 - X1 long vector performance not as good as Earth Simulator (ES)
 - X1 short vector performance superior to ES
- Scalability of POP determined by communication latency
 - MPI short message and collective performance mediocre
 - Co-array Fortran and SHMEM performance excellent
- Planned Cray X1 optimizations
 - Scalable (tree-based) allreduce
 - Portable Co-array Fortran
 - Cray-specific vectorization

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.
- GYRO comes with ports to a number of different platforms. The port and optimization on the Cray X1 is primarily due to Mark Fahey of ORNL. In the Cray X1 port, GYRO is coded as if the MSP is the processor.
- Optimization on the X1 ongoing.

GYRO Experiment Particulars

Two benchmark problems, both time dependent:

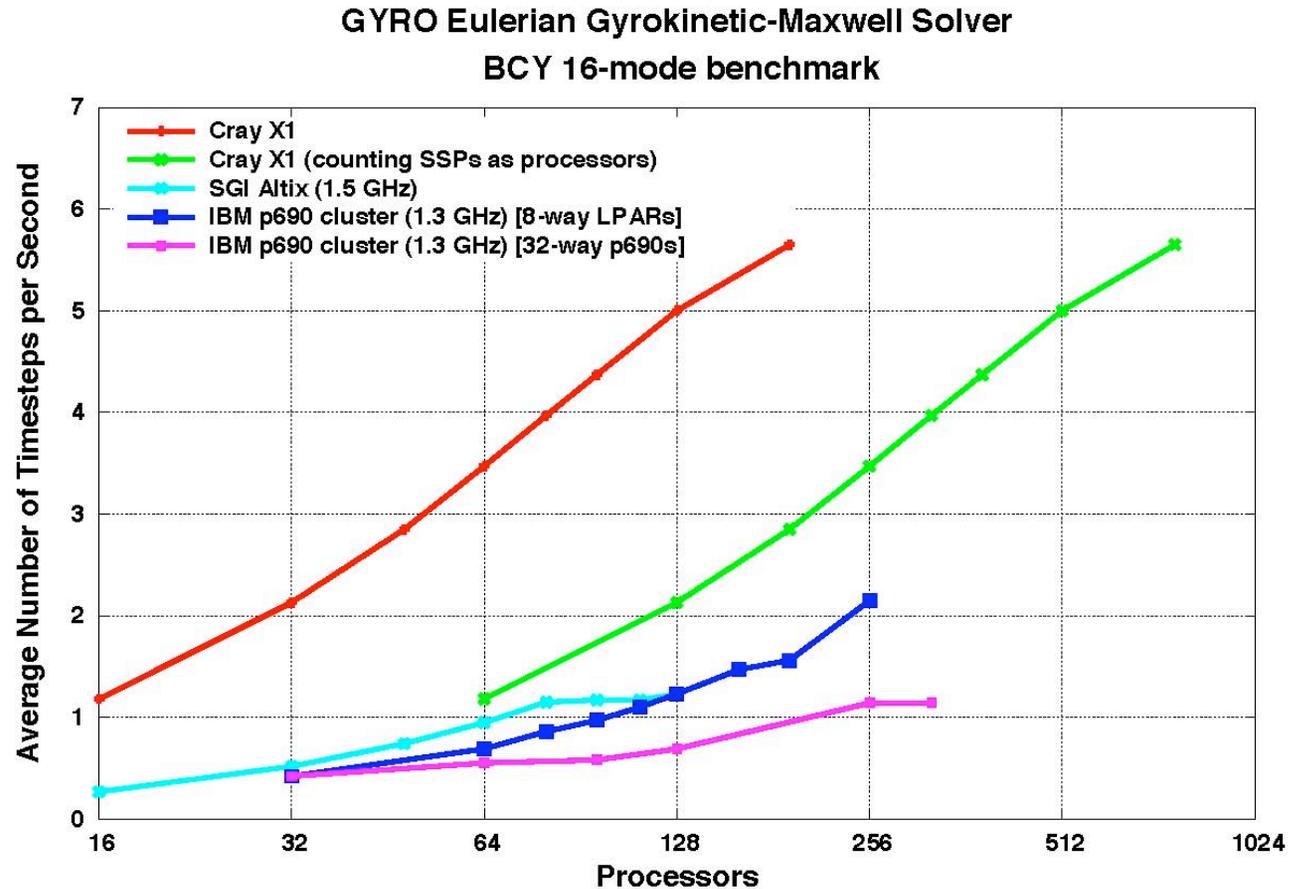
1. BCY.n16.b.25
 - 16-mode electromagnetic case. It is run on multiples of 16 processors. Duration is 8 simulation seconds, representing 1000 timesteps.
2. GTC.n64.500
 - 64-mode adiabatic electron case. It is run on multiples of 64 processors. Duration is 3 simulation seconds, representing 100 timesteps.

Current production runs use 32 modes, so benchmark #1 is somewhat small, while benchmark #2 is very large. (J. Candy is in the process of reformulating these benchmarks to also cover production-size problems.)

GYRO Simulation Rate

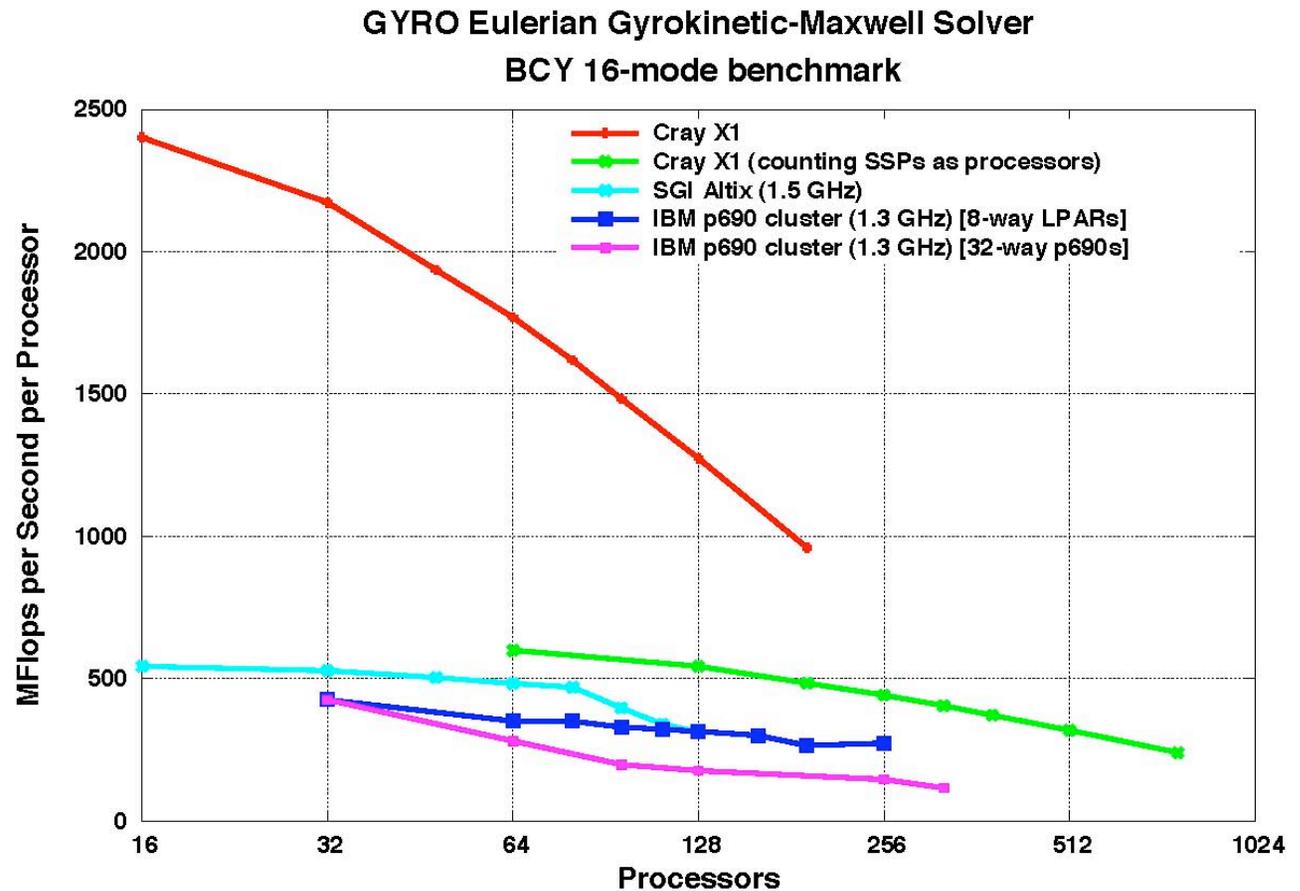
Comparing performance and scaling across platforms.

- X1 performance is significantly better than that on other platforms, even for this modest size problem, and advantage grows with processor count. Even replotting data with SSPs as processors indicates that the X1 is the faster platform.



GYRO Computational Rate

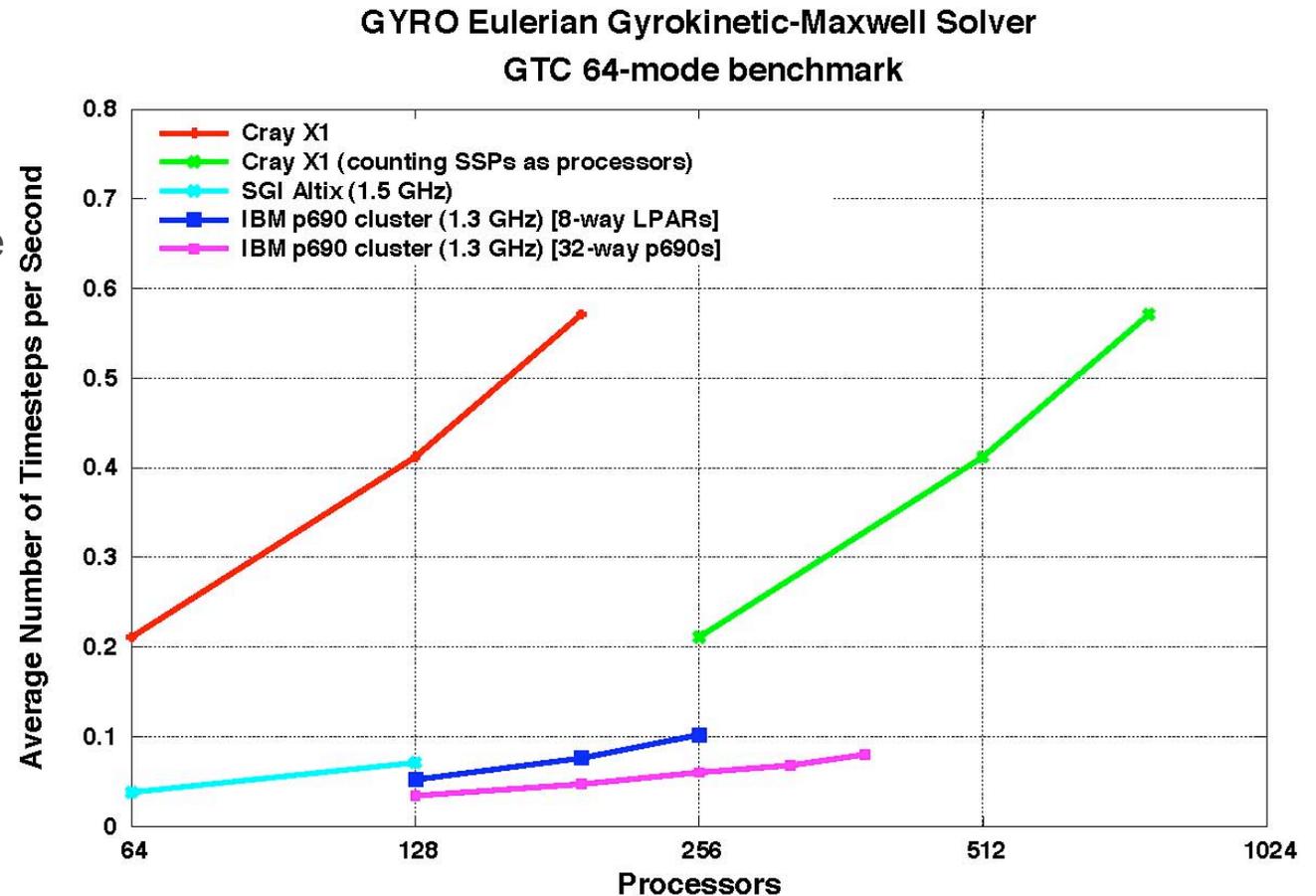
- IBM performance is limited by communication overhead.
- X1 performance is limited by nonscaling part of parallel algorithm. At 192 processors, 25% of the time is spent in a phase that does not scale beyond 16-way parallelism.



GYRO Simulation Rate

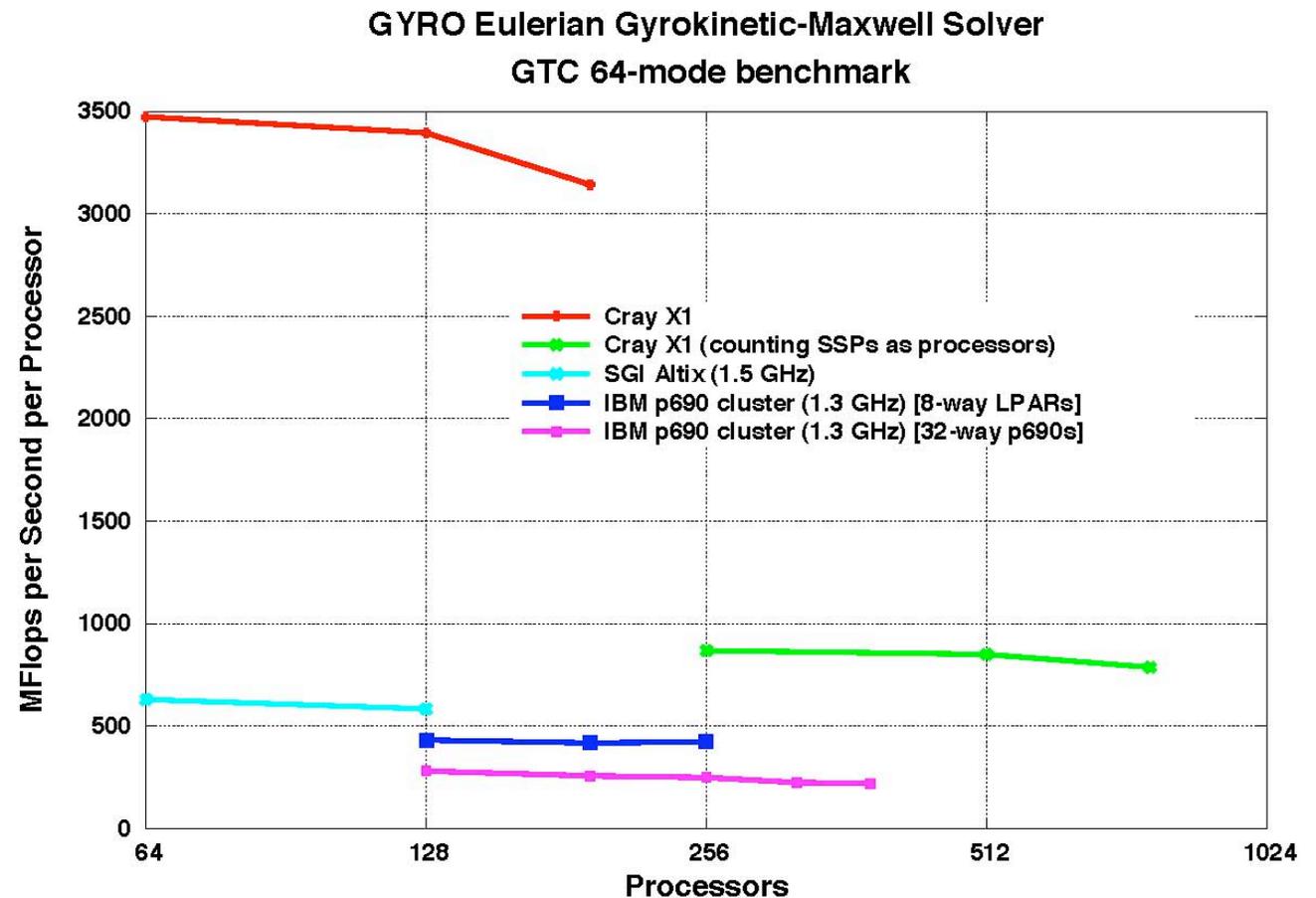
Comparing performance and scaling across platforms.

- X1 performance advantage is even more pronounced (factor of 8) for the larger problem size.



GYRO Computational Rate

- Nonscaling phase can use at most 64 processors, but is only 3% of execution time on X1 for 192 processors.
- All platforms show reasonable scaling, but IBM performance is still limited by bandwidth.



GYRO Summary

- Performance on nonvector systems constrained by communication bandwidth.
 - This is not true on the Cray.
- Scalability of POP on X1 determined by nonscaling phase.
- Vectorization efforts are not complete. There are known (correctable) losses in vector performance.