

CCSM Component Performance Benchmarking and Status of the CRAY X1 at ORNL

Patrick H. Worley

Oak Ridge National Laboratory

Computing in Atmospheric Sciences Workshop 2003

September 10, 2003

L'Imperial Palace Hotel

Annecy, France

OAK RIDGE NATIONAL LABORATORY
U. S. DEPARTMENT OF ENERGY



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.

At the Request of the Organizers ...

Two talks in one

- Benchmarking and Performance Analysis of CCSM Ocean and Atmosphere Models
- Early Experiences with the Cray X1 at Oak Ridge National Laboratory

Only time for a quick overview, concentrating on process as much as results. For more information, see <http://www.csm.ornl.gov/evaluation>

Platforms

- Cray X1 at Oak Ridge National Laboratory (ORNL): 16 4-way vector SMP nodes and a torus interconnect. Each processor has 8 64-bit floating point vector units running at 800 MHz.
- 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.
- Compaq AlphaServer SC at ORNL: 64 ES40 4-way SMP nodes (667MHz Alpha EV67) and a Quadrics QsNet interconnect with one network adapter per node.

Platforms (cont.)

- 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.
- IBM SP at ORNL: 176 Winterhawk II 4-way SMP nodes (375MHz POWER3-II) and an SP Switch with one network adapter per node.
- NEC SX-6 at the Arctic Region Supercomputing Center: 8-way SX-6 SMP node. Each processor has 8 64-bit floating point vector units running at 500 MHz.
- SGI Origin 3000 at Los Alamos National Laboratory (LANL): 512-way SMP node. Each processor is a 500 MHz MIPS R14000.

Community Climate System Model (CCSM)

- Community Atmospheric Model (CAM2)
 - Eulerian spectral dynamics*
 - Finite Volume/Semi-Lagrangian gridpoint dynamics
- Parallel Ocean Program (POP 1.4.3)*
- Community Land Model (CLM2)
- CCSM Ice Model (CSIM)
- CCSM Coupler (CPL6)

* models examined in this talk

CCSM Component Benchmarking

- Performance evolution and diagnosis
 - for development planning
- Performance scaling
 - for configuration optimization and for resource planning
- Platform evaluation

Component Benchmarking Funded by ...

- SciDAC Performance Evaluation Research Center (<http://perc.nersc.gov>)
- SciDAC Collaborative Design and Development of the Community Climate System Model for Terascale Computers (<http://www.osti.gov/scidac/ber/projects/malone.html>)
- Evaluation of Early Systems (<http://www.csm.ornl.gov/evaluation>)

in collaboration with the CCSM Software Engineering Group (CSEG), Software Engineering Working Group (SEWG), and the NCAR Scientific Computing Division.

CAM Performance Evolution

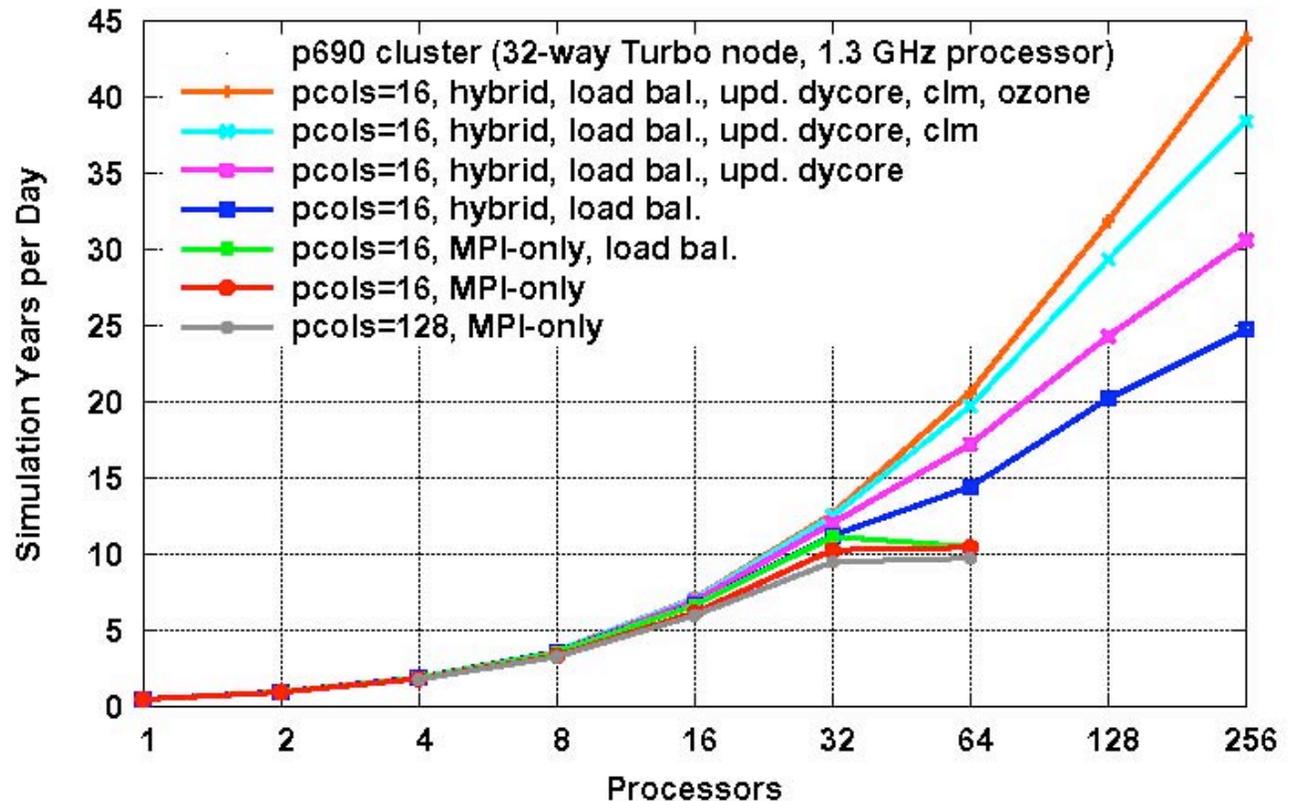
Measuring perf. impact of tuning parameters:

- Physics data structure size (pcols)
- Load balancing
- MPI processes vs. OpenMP threads

and code optimizations:

- spectral dycore load balance and comm. optimizations
- land/atmosphere interface comm. optimizations
- distributed ozone interpolation (added after dev10)

Performance Evolution of NCAR Community Atmospheric Model CAM2.0.1.dev10+, EUL dynamical core, T42L26, min I/O



CAM computational rate on the IBM p690 cluster.

CAM Performance Evolution

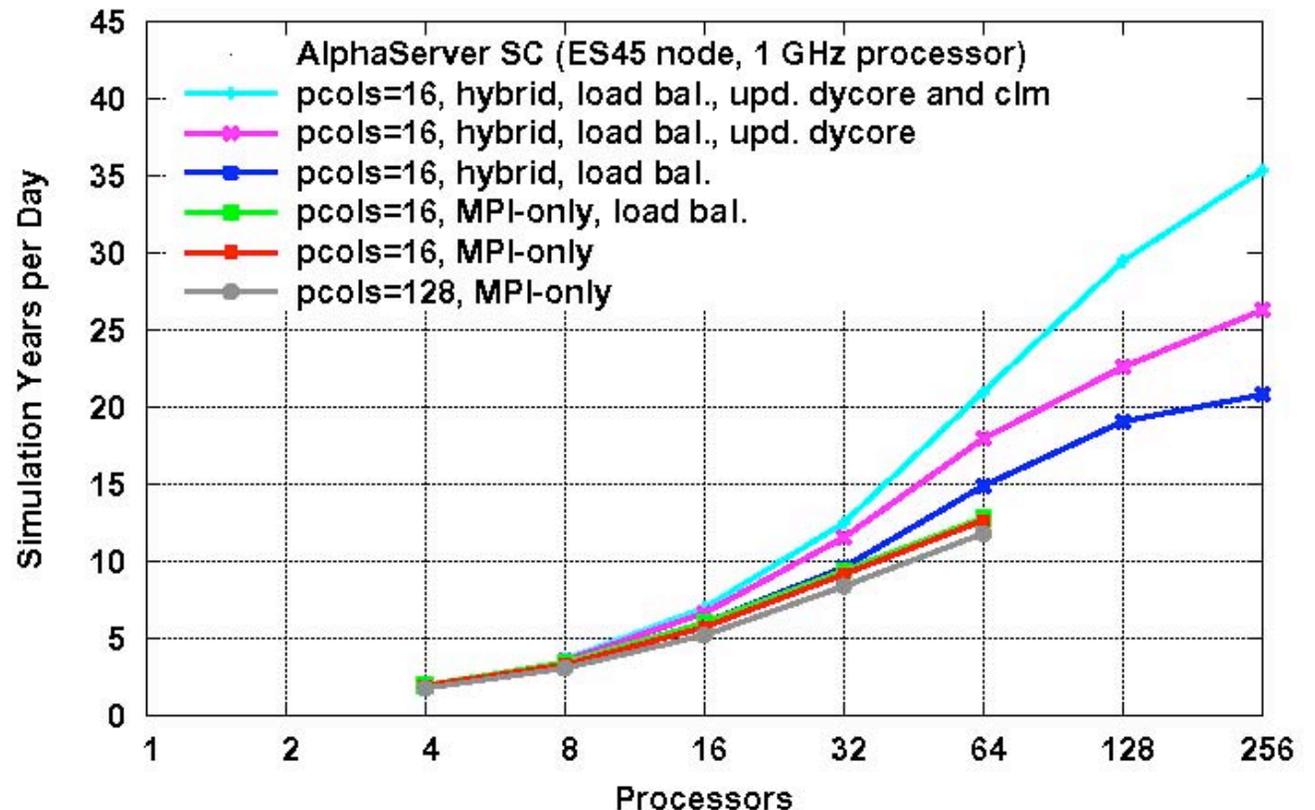
Measuring perf. impact of tuning parameters:

- Physics data structure size (pcols)
- MPI processes vs. OpenMP threads
- Load balancing

and code optimizations:

- spectral dycore load balance and comm. optimizations
- land/atmosphere interface comm. optimizations

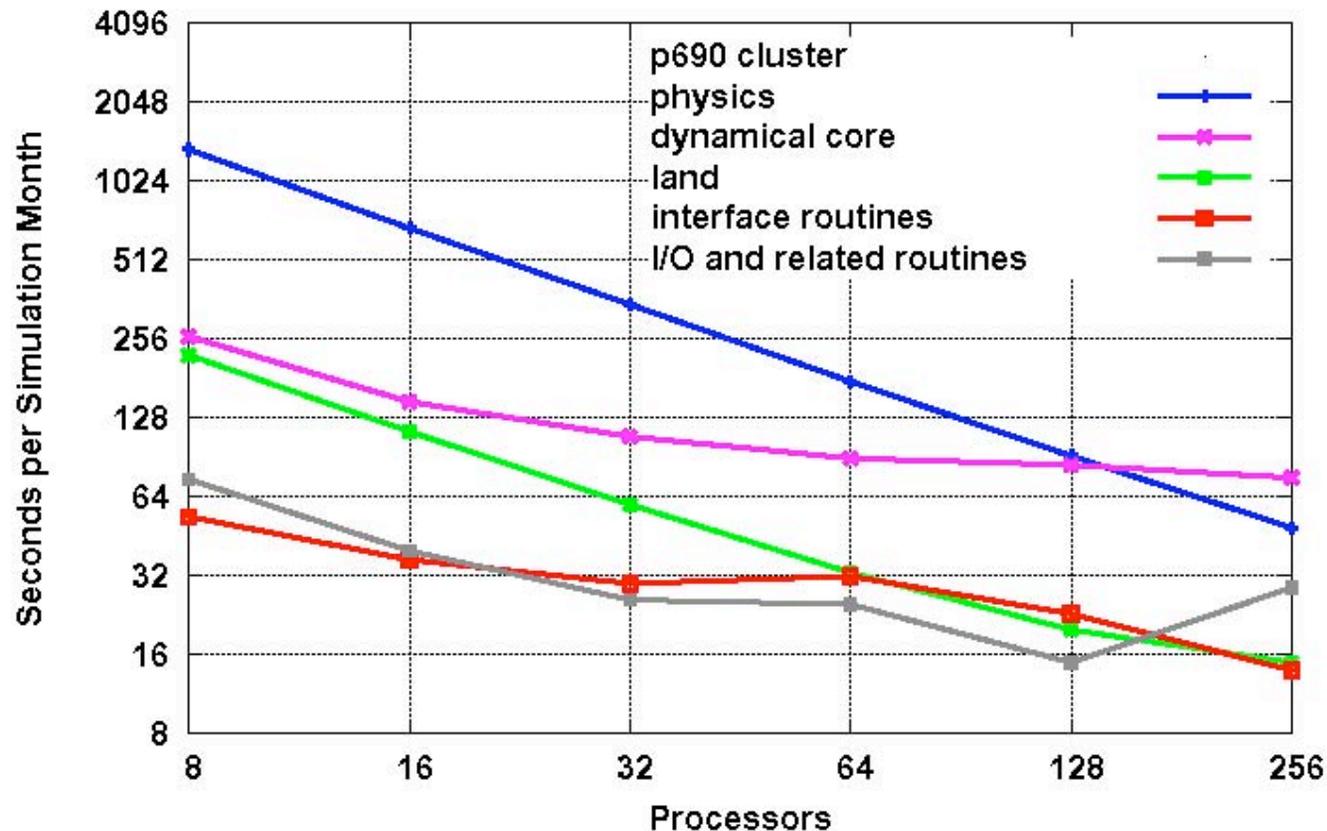
Performance Evolution of NCAR Community Atmospheric Model
CAM2.0.1.dev10, EUL dynamical core, T42L26, min I/O



CAM computational rate on the AlphaServer SC.

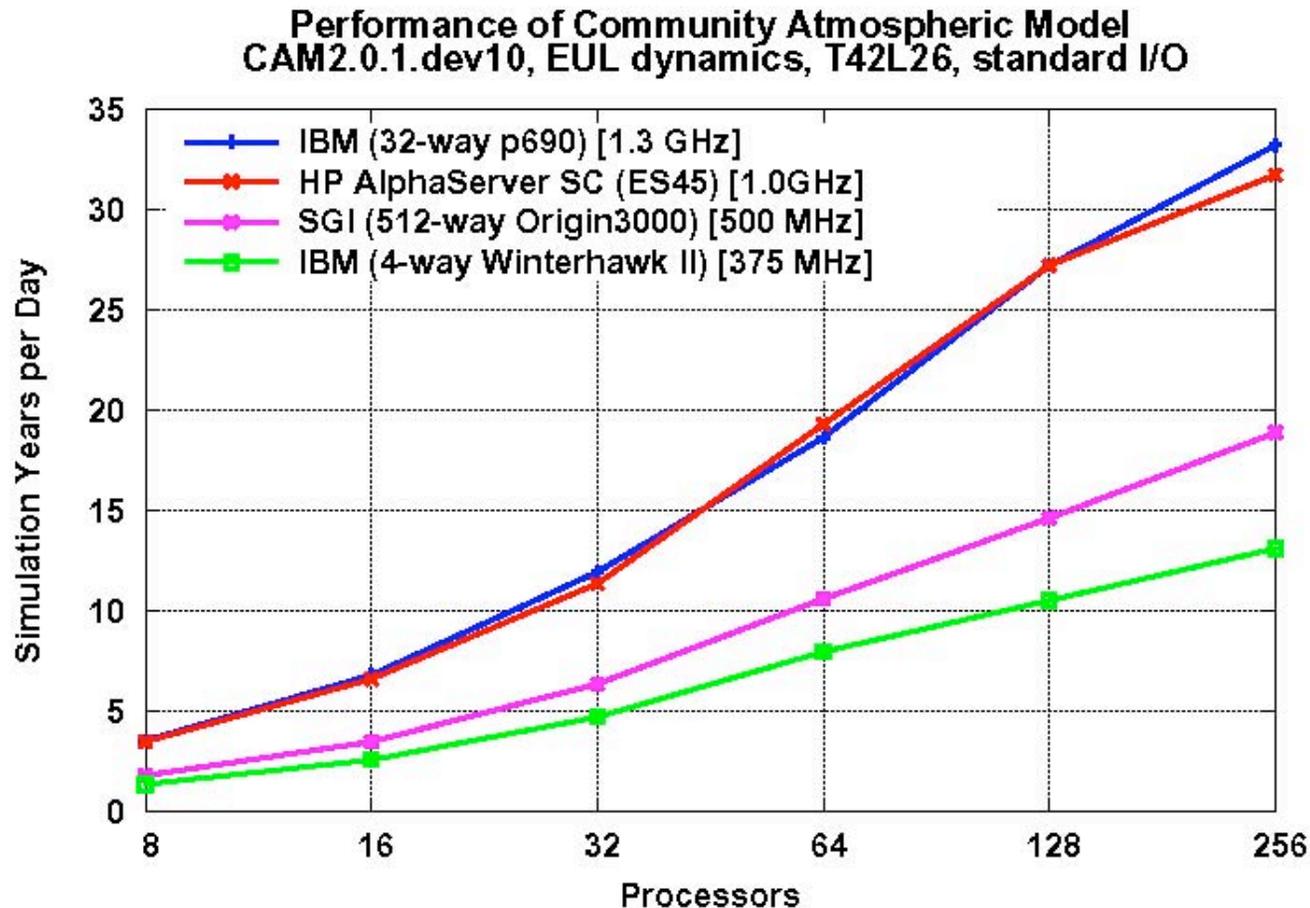
CAM Performance Diagnosis

Performance Diagnosis of NCAR Community Atmospheric Model
CAM2.0.1.dev10+, EUL dynamical core, T42L26, standard I/O



Measuring performance and scaling of individual phases for CAM2.0.1dev10+ (with distributed ozone interpolation) when run with standard I/O on the IBM p690 cluster.

CAM Platform Comparison



Comparing performance and scaling across platforms for CAM2.0.1dev10 (without distributed ozone interpolation) when run in production mode.

What's next for CAM benchmarking?

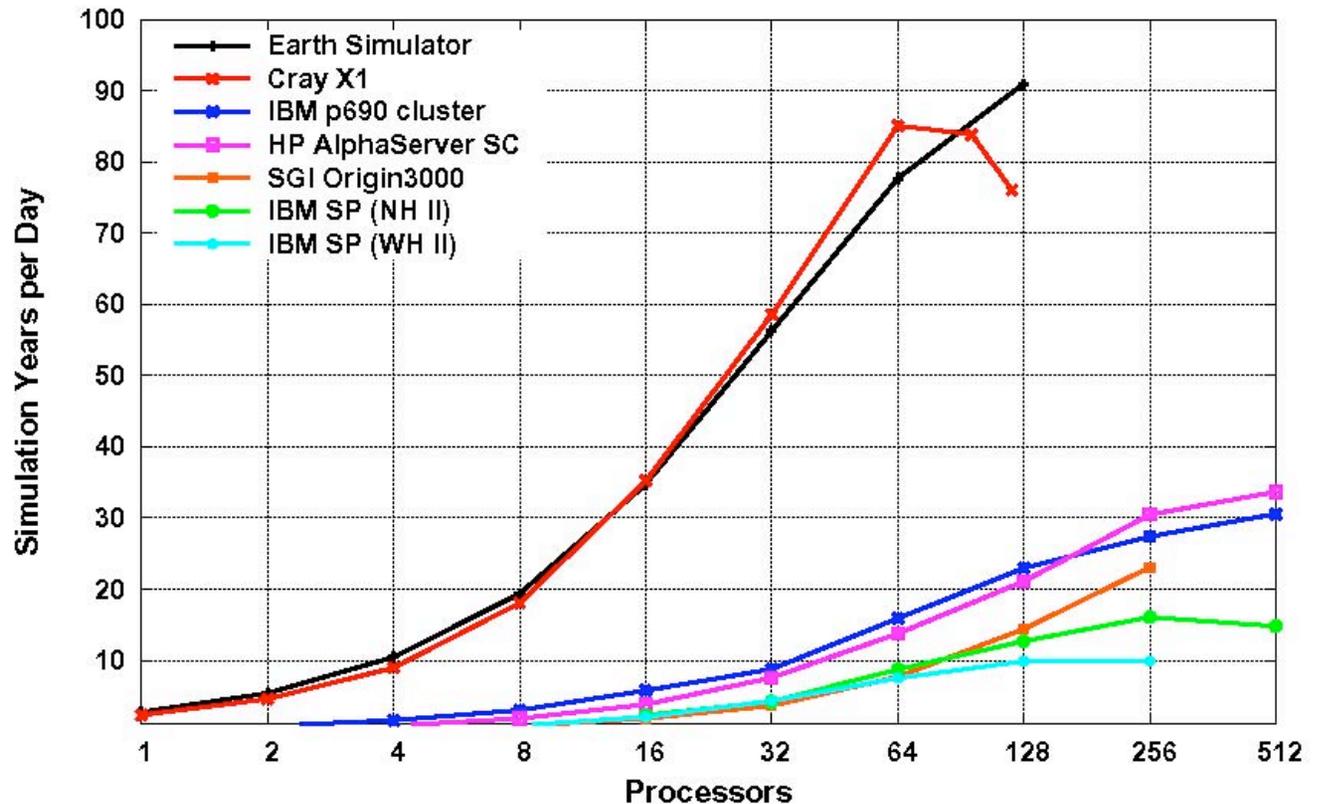
- CAM 2_0_2
 - modified physical parameterizations
 - increased number of advected constituents
- Increased resolution
 - Spectral Eulerian: T85, T170
 - Finite Volume: 1x1.25, 0.5x0.625 degree
- Tracking additional optimizations
 - see also R. Loft's presentation

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 results courtesy of Dr. P. Jones of LANL.
- IBM SP (NH II) results courtesy of Dr. T. Mohan of Lawrence Berkeley National Laboratory (LBNL)

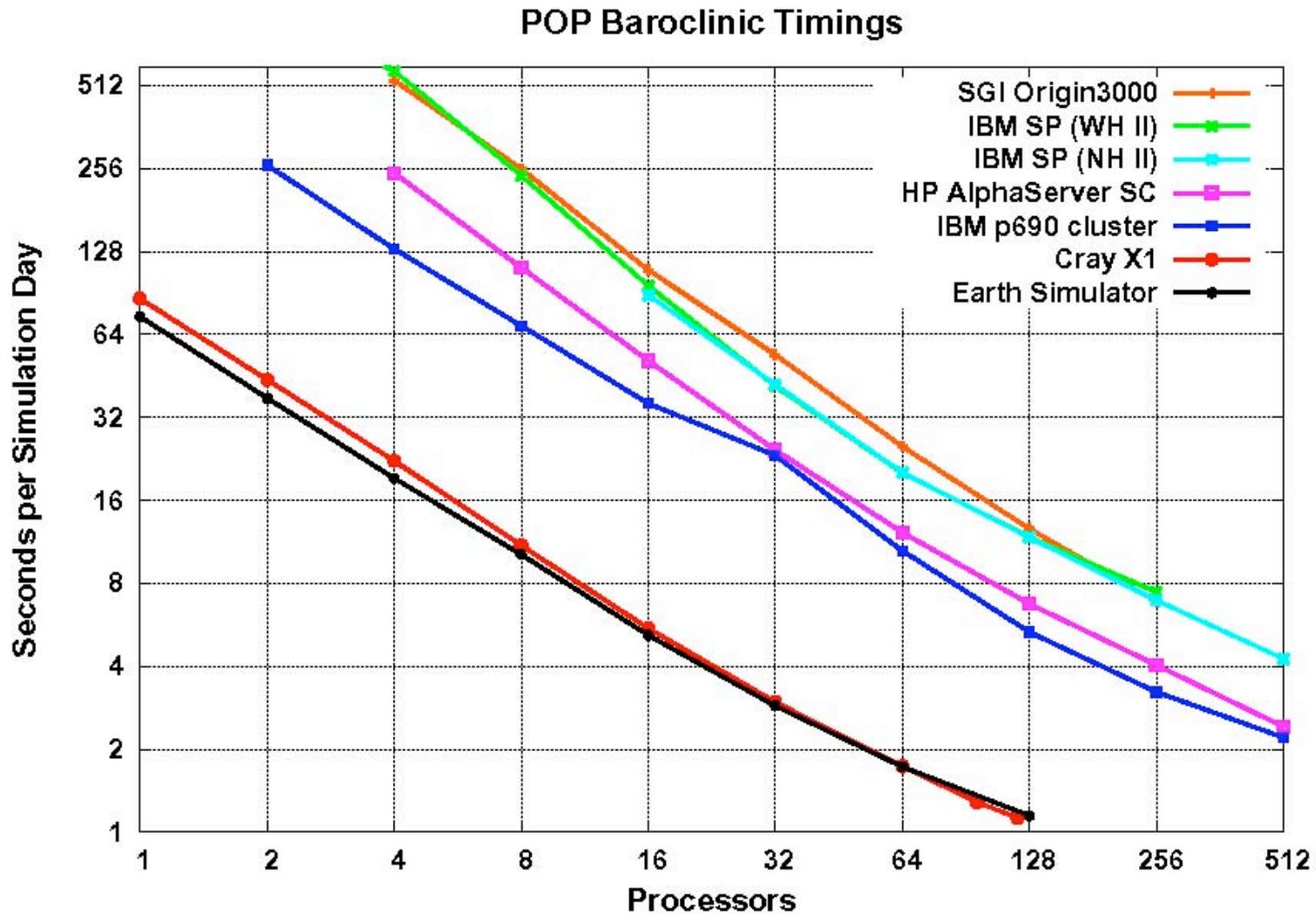
LANL Parallel Ocean Program
POP 1.4.3, x1 benchmark



POP Diagnostic Experiments

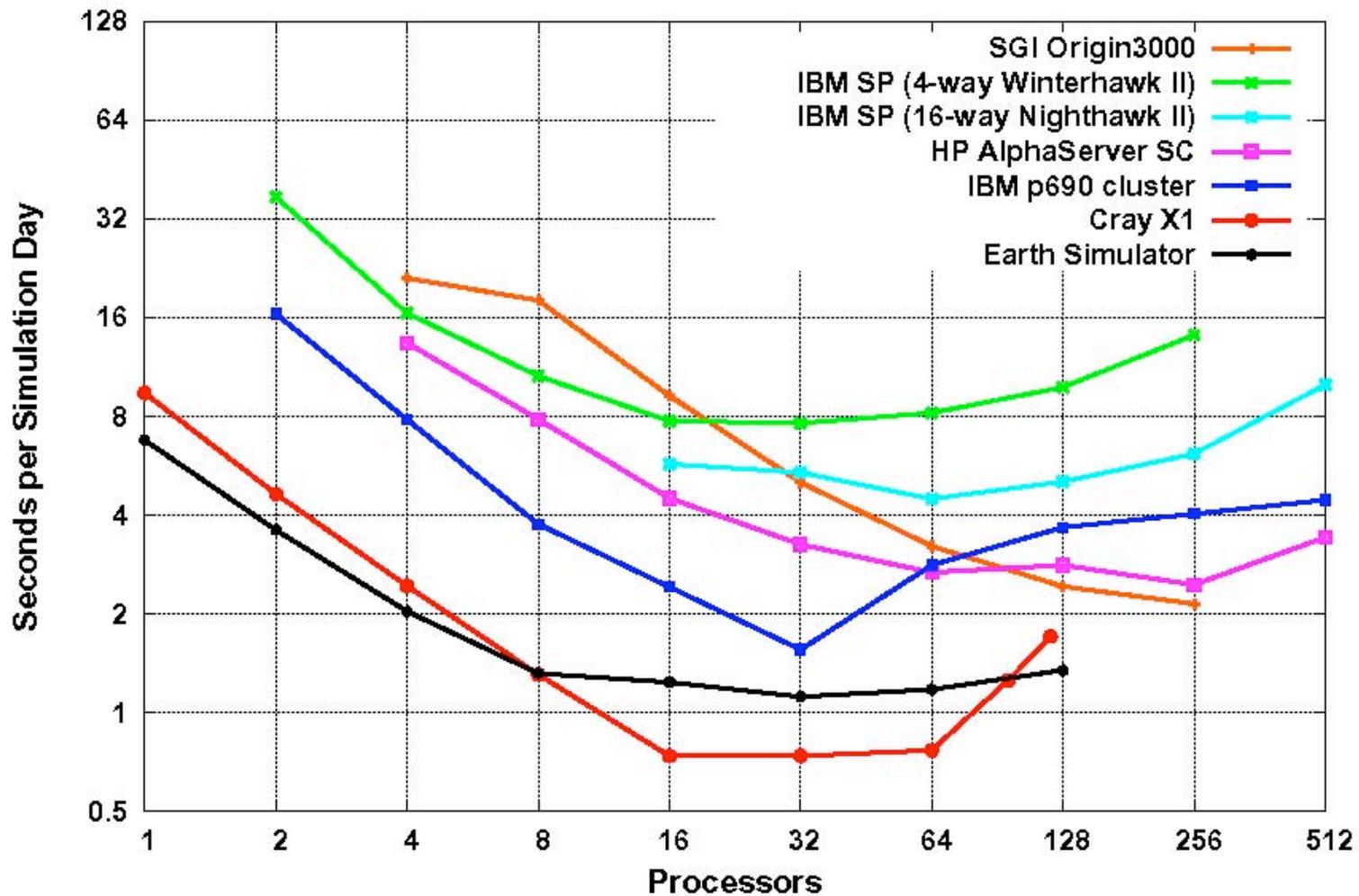
- Two primary phases that determine model performance
 - Baroclinic: 3D with limited nearest-neighbor communication; scales well.
 - Barotropic: dominated by solution of 2D implicit system using conjugate gradient solves; scales poorly
- One benchmark problem size
 - One degree horizontal grid (“by one” 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 Performance Diagnosis: Baroclinic



POP Performance Diagnosis: Barotropic

POP Barotropic Timings



What's next for POP benchmarking?

- Tracking additional optimizations
 - primarily Cray X1
- Increased resolution
 - 0.1 degree resolution
- More recent versions of the model
 - CCSM version of POP 1.4.3
 - POP 2.0
 - HYPOP

Phoenix

Cray X1 with 32 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 128 processors (MSPs), 512 GB of memory, and 1600 GF/s peak performance. System will be upgraded to 64 nodes (256 MSPs) by the end of September, 2003.



Evaluation Goals

- Verifying advertised functionality and performance
- Quantifying performance impact of
 - Scalar vs. Vector vs. Streams
 - Contention for memory within SMP node
 - SSP vs. MSP mode of running codes
 - Page size
 - MPI communication protocols
 - Alternatives to MPI: SHMEM and Co-Array Fortran
 - ...
- Guidance to Users
 - What performance to expect
 - Performance quirks and bottlenecks
 - Performance optimization tips

Caveats

- These are EARLY results, resulting from approx. four month of sporadic benchmarking on evolving system software and hardware configurations.
- Performance characteristics are still changing, due to continued evolution of OS and compilers and libraries.
 - This is a good thing - performance continues to improve.
 - This is a problem - performance data and analysis have a very short lifespan.

Outline of Full Talk

- Standard or External Benchmarks (unmodified)
 - Single MSP performance*
 - Memory subsystem performance
 - Interprocessor communication performance*
- Custom Kernels
 - Performance comparison of compiler and runtime options
 - Single MSP and SSP performance*
 - SMP node memory performance
 - Interprocessor communication performance*
- Application Codes
 - Performance comparison of compiler and runtime options
 - Scaling performance for fixed size problem*

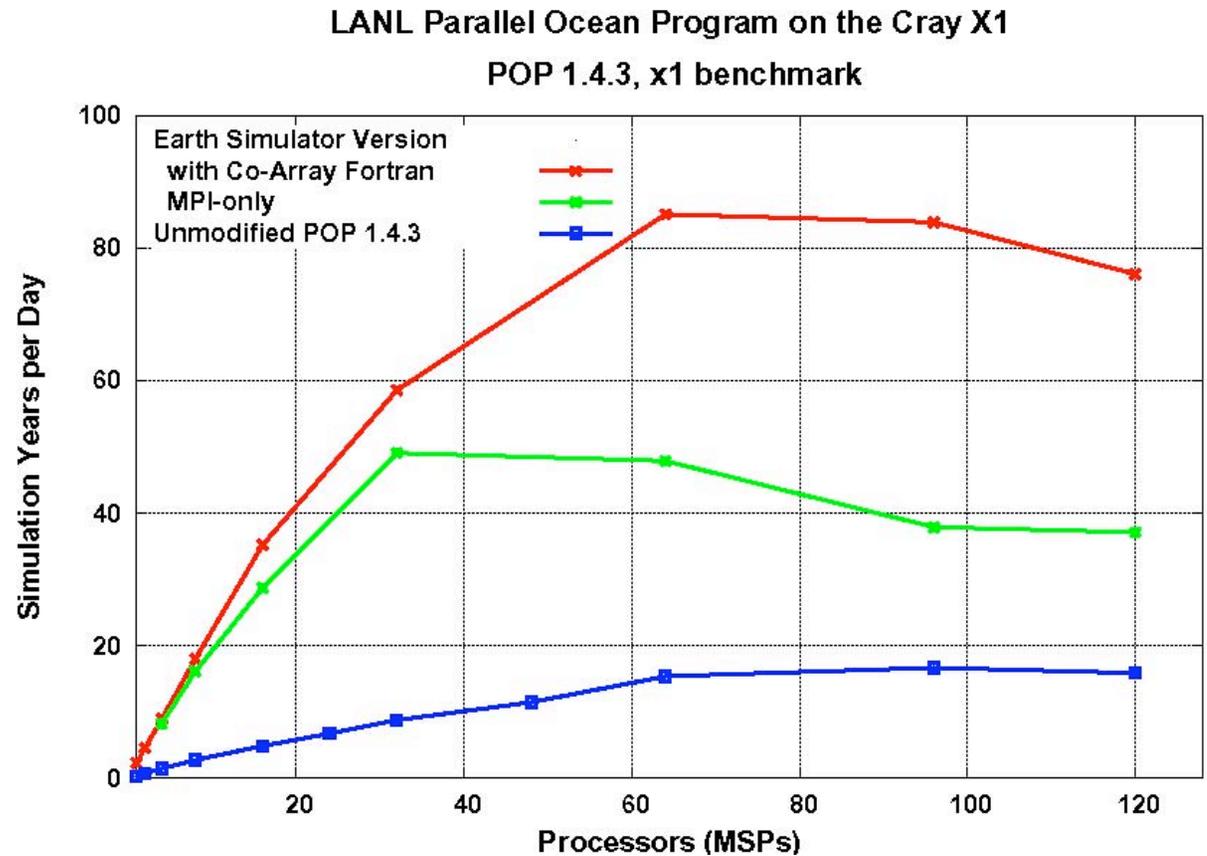
* topics examined in this talk

POP: More of the story ...

- Ported to the Earth Simulator by Dr. Yoshikatsu Yoshida of 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 of ORNL.
- Optimization on the X1 ongoing:
 - Cray-specific vectorization
 - Improved scalability of barotropic solver algorithm
 - OS performance tuning

POP Version Performance Comparison

Comparing performance of different versions of POP. The Earth Simulator version vectorizes reasonably well on the Cray. Most of the performance improvement in the current version is due to Co-Array implementation of the conjugate gradient solver. This should be achievable in the MPI version as well.



Kernel Benchmarks

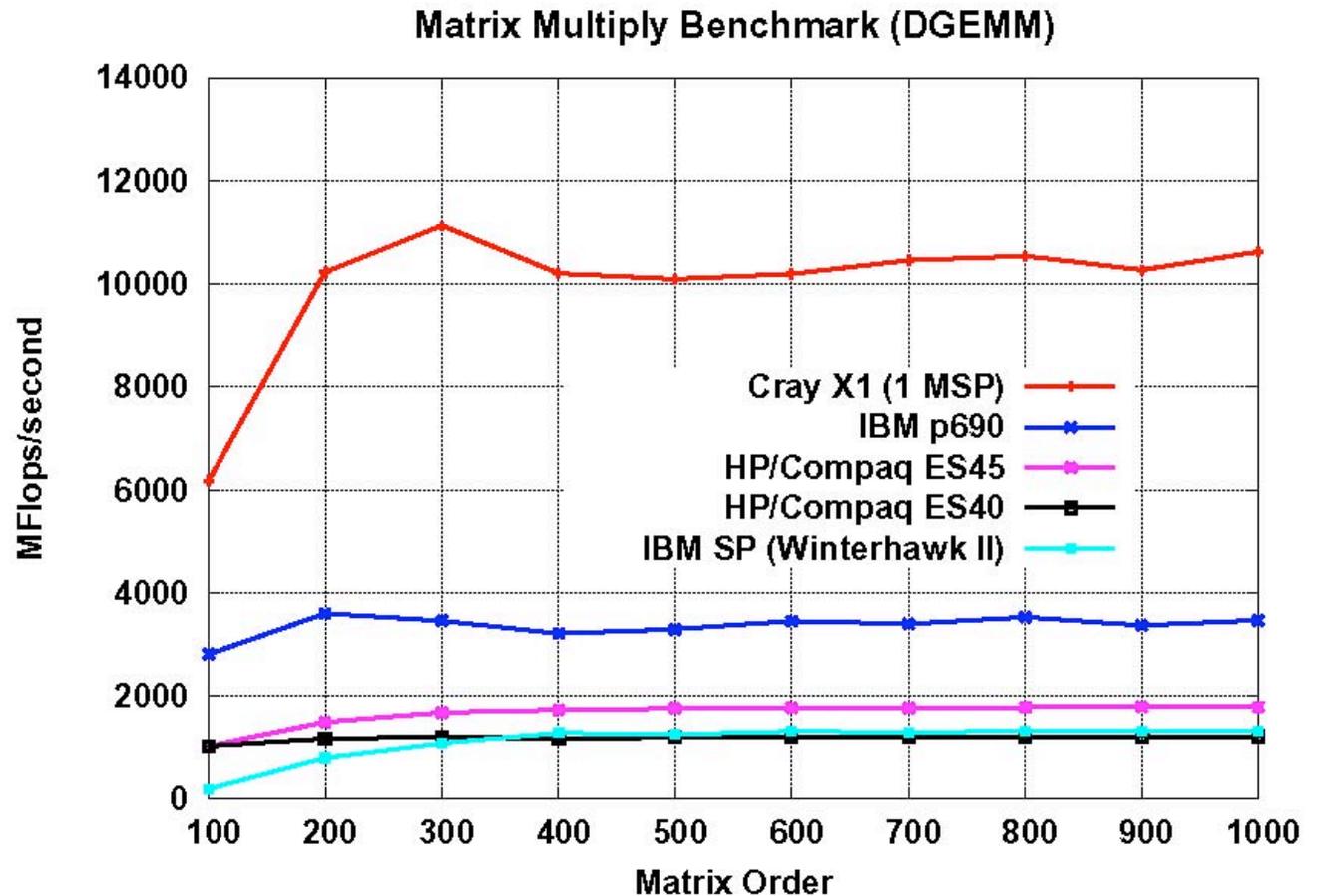
- Single Processor Performance
 - DGEMM matrix multiply benchmark*
 - Euroben MOD2D dense eigenvalue benchmark*
 - Euroben MOD2E sparse eigenvalue benchmark*
- Interprocessor Communication Performance
 - HALO benchmark*
 - COMMTEST benchmark suite

* data collected by Tom Dunigan of ORNL. See <http://www.csm.ornl.gov/evaluation> for more details.

DGEMM Benchmark

Comparing performance of vendor-supplied routines for matrix multiply. Cray X1 experiments used routines from the Cray scientific library libsci.

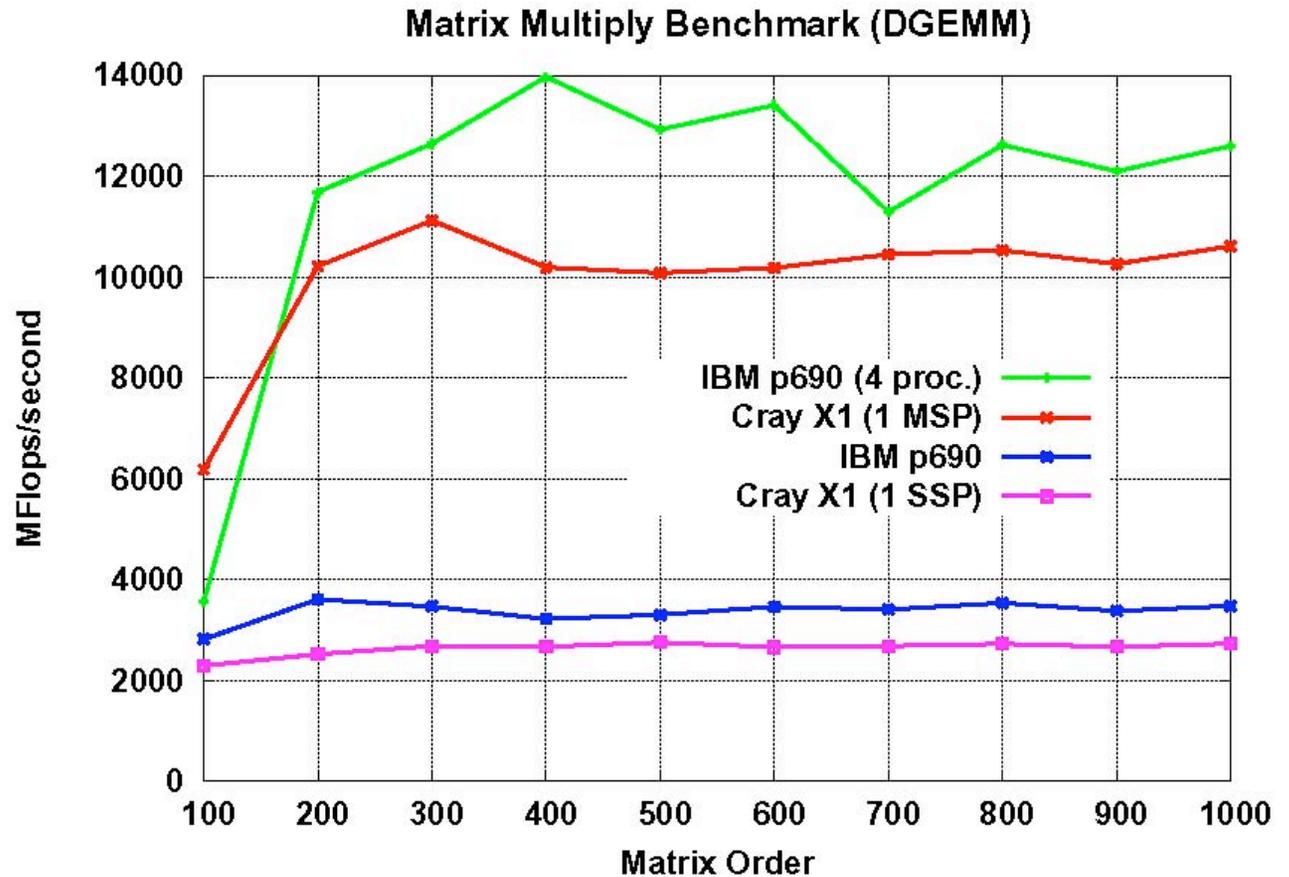
Good performance achieved, reaching 80% of peak relatively quickly.



DGEMM Benchmark - What's a Processor?

Comparing performance of X1 MSP, X1 SSP, p690 processor, and four p690 processors (using PESSL parallel library).

Max. percentage of peak -
X1 SSP: 86%
X1 MSP: 87%
p690 (1): 70%
p690 (4): 67%



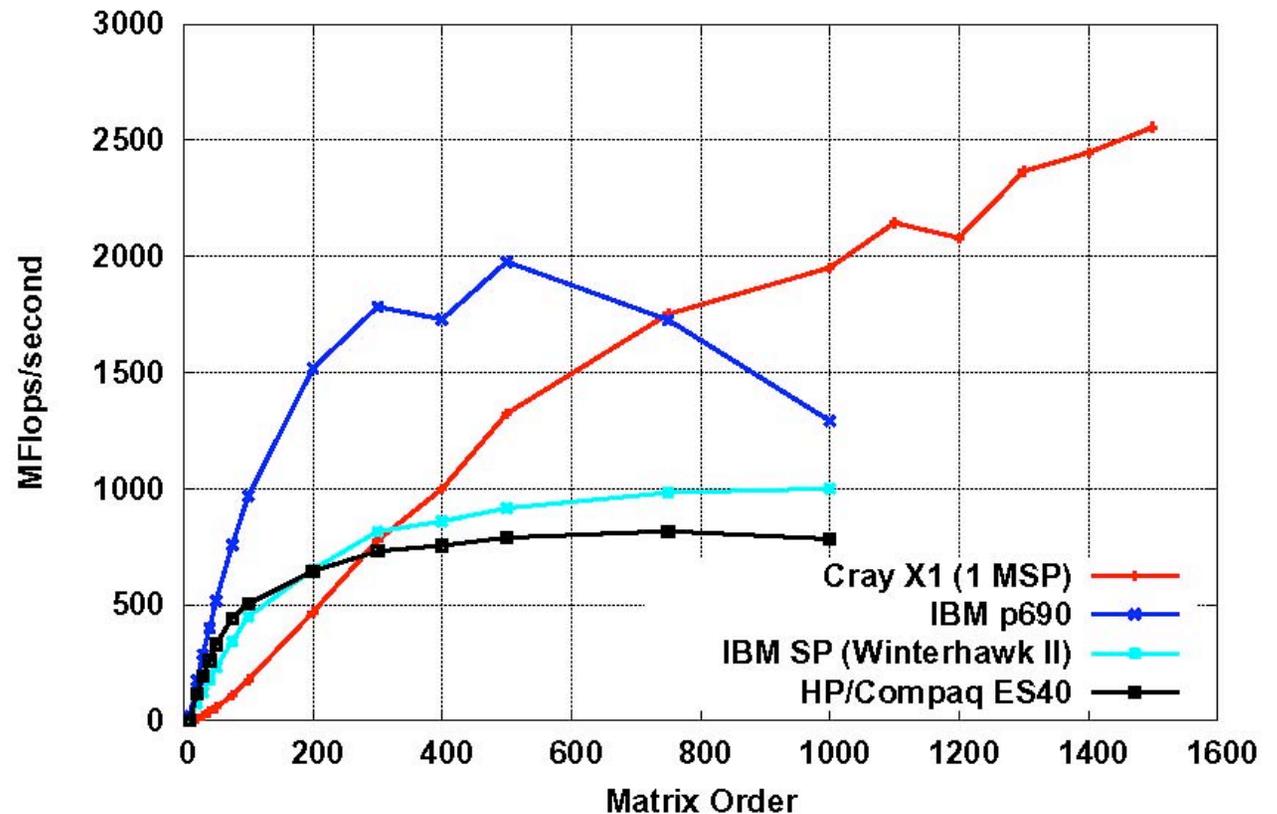
MOD2D Benchmark

Comparing performance of vendor-supplied routines for dense eigenvalue analysis. Cray X1 experiments used routines from the Cray scientific library libsci.

Performance still growing with problem size. (Had to increase standard benchmark problem specifications.)

Performance of nonvector systems has peaked.

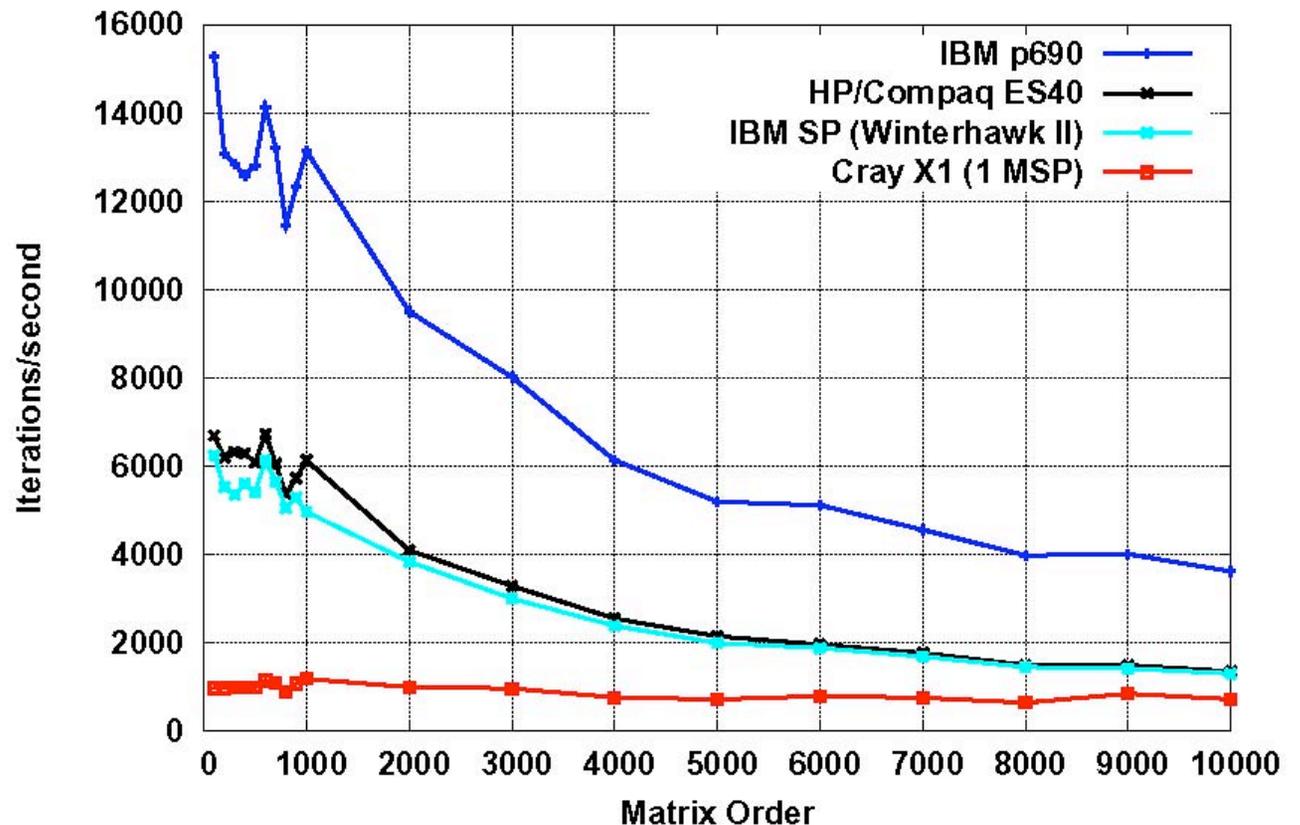
Dense Eigenvalue Benchmark (Euroben MOD2D)



MOD2E Benchmark

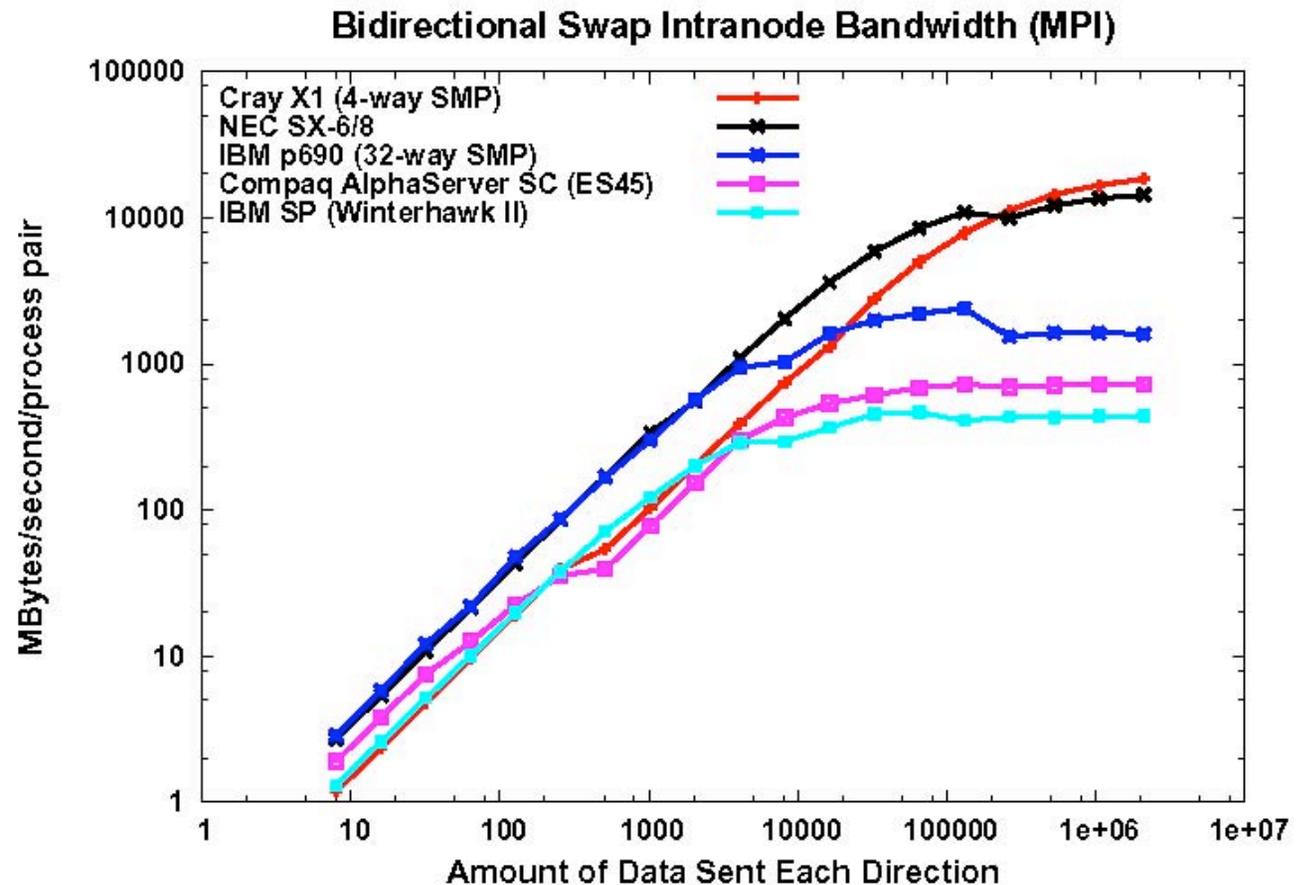
Comparing performance of Fortran code for sparse eigenvalue analysis. Aggressive compiler options were used on the X1, but code was not restructured and compiler directives were not inserted. Performance is improving for larger problem sizes, so some streaming or vectorization is being exploited. Performance is poor compared to other systems.

Sparse Eigenvalue Benchmark (Euroben MOD2E)



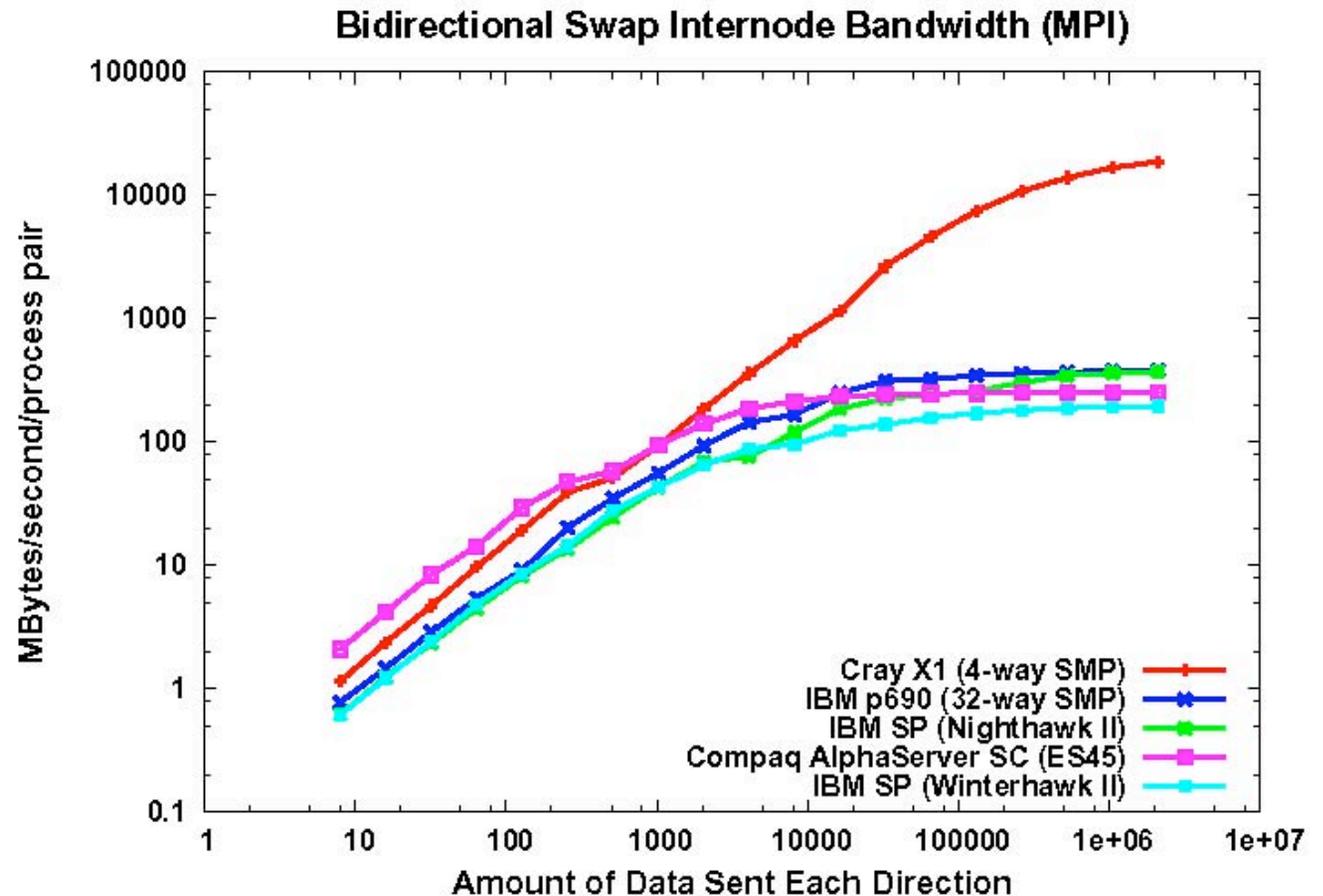
COMMTEST SWAP Benchmark

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



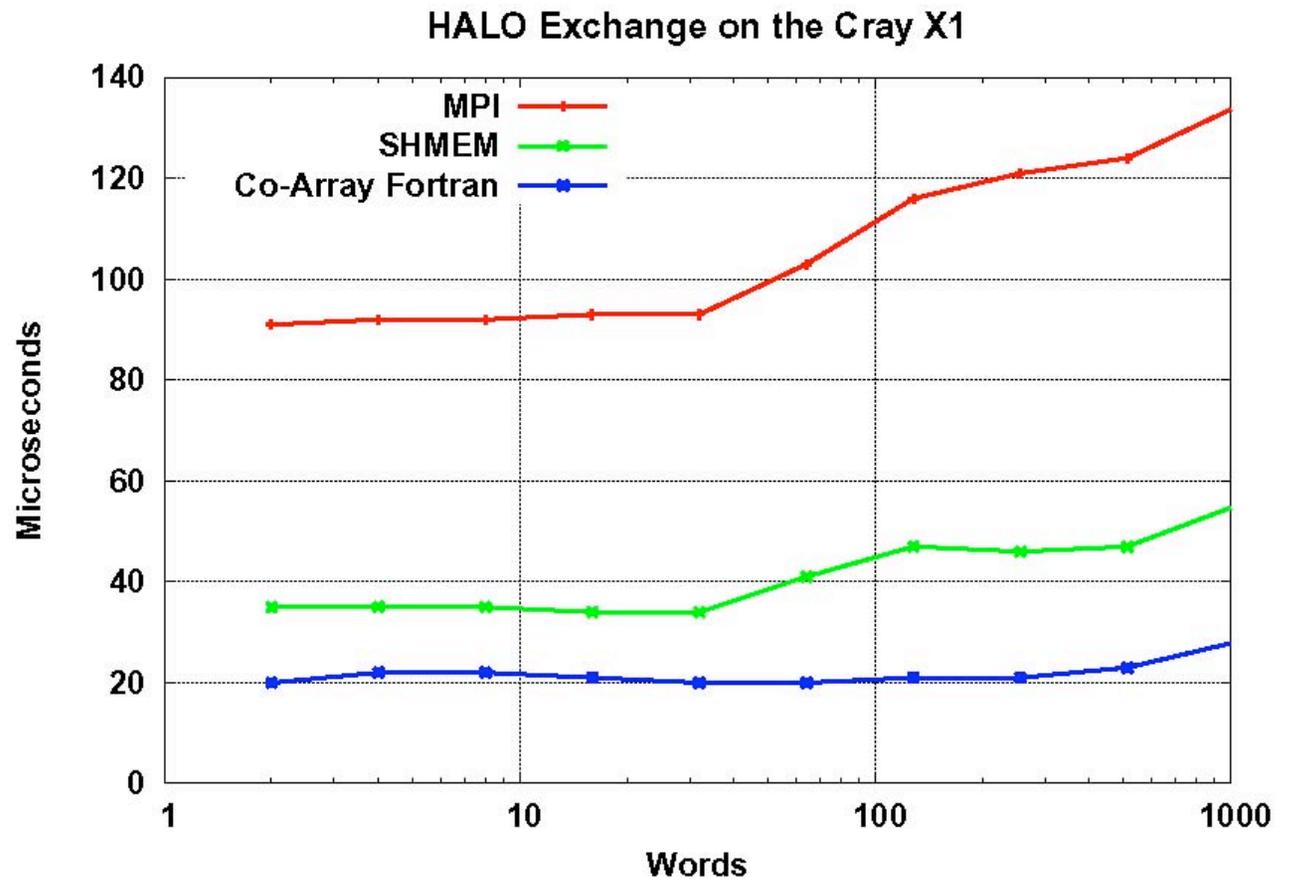
COMMTEST SWAP Benchmark

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



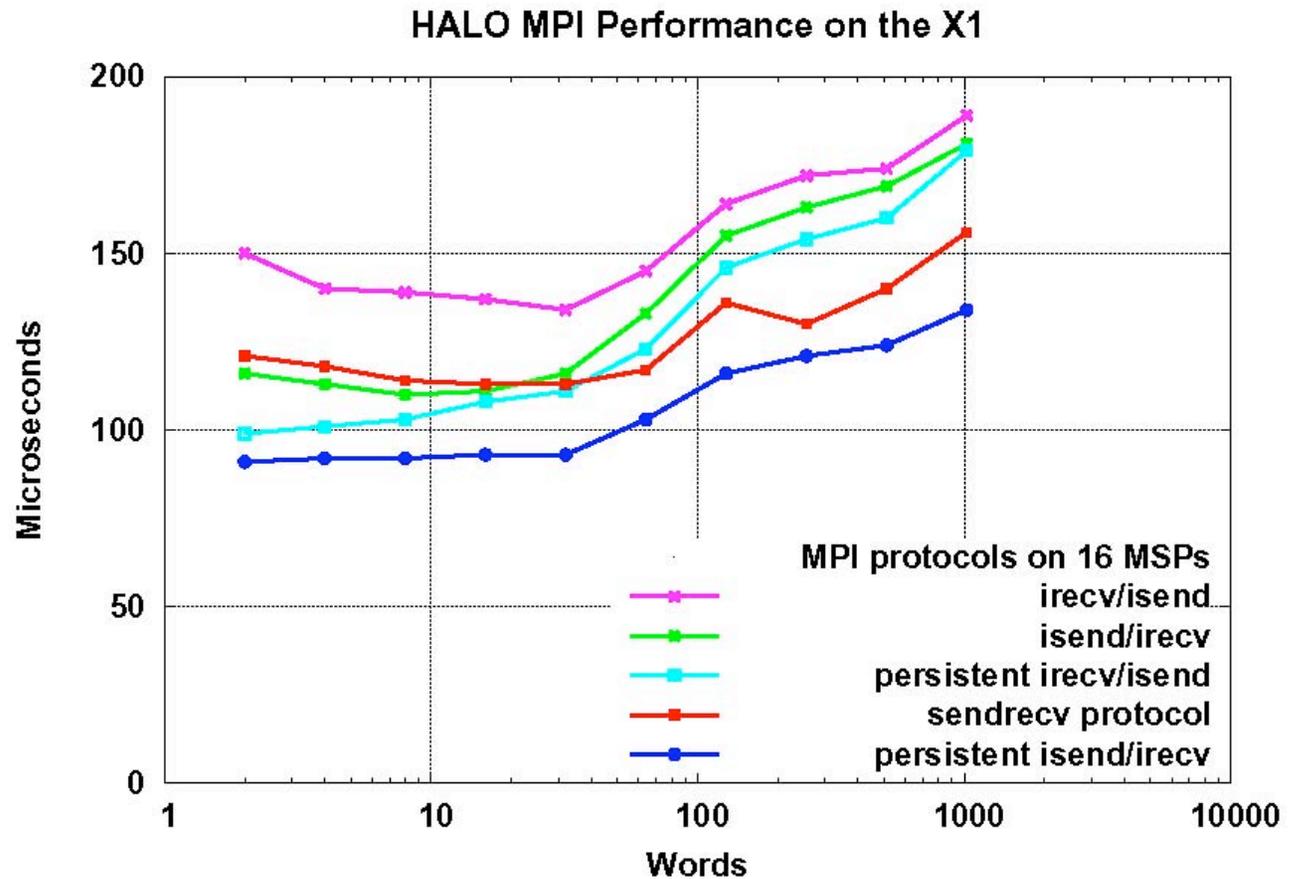
HALO Paradigm Comparison

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.



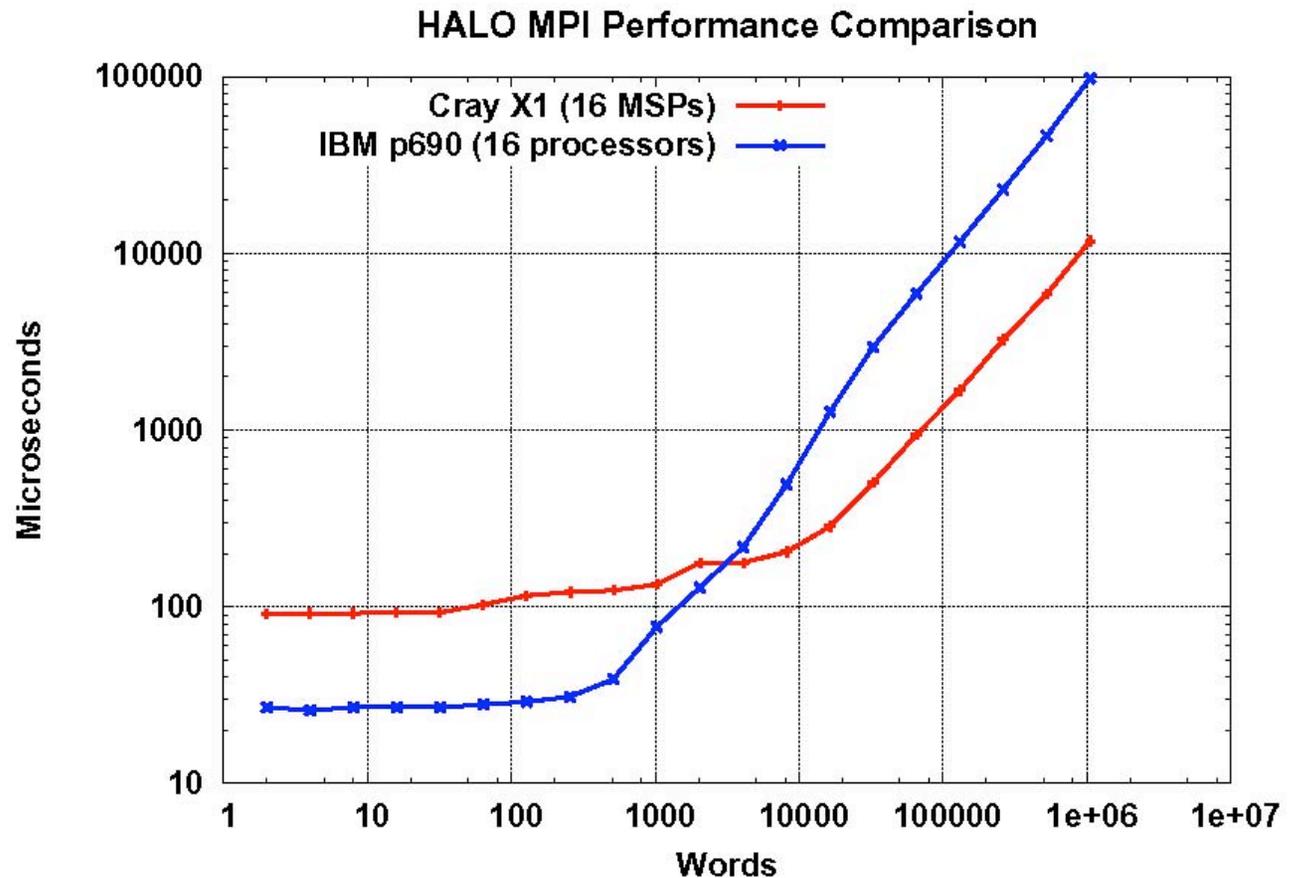
HALO MPI Protocol Comparison

Comparing performance of different MPI implementations of HALO on 16 MSPs. Persistent isend/irecv is always best. For codes that can not use persistent commands, MPI_SENDRECV is also a reasonable choice.



HALO Benchmark

Comparing HALO performance using MPI on 16 MSPs of the Cray X1 and 16 processors of the IBM p690 (within a 32 processor p690 SMP node). Achievable bandwidth is much higher on the X1. For small halos, the p690 MPI HALO performance is between the X1 SHMEM and Co-Array Fortran HALO performance.



COMMTEST Benchmark

COMMTEST is a suite of codes that measure the performance of MPI interprocessor communication. In particular, COMMTEST evaluates the impact of communication protocol, packet size, and total message length in a number of “common usage” scenarios. (However, it does not include persistent MPI point-to-point commands among the protocols examined.) It also includes simplified implementations of the SWAP and SENDRECV operators using SHMEM.

COMMTEST Experiment Particulars

0-1

MSP 0 swaps data with MSP 1 (within the same SMP node)

0-4

MSP 0 swaps data with MSP 4 (between two neighboring nodes)

0-8

MSP 0 swaps data with MSP 8 (between two more distant nodes)

$i-(i+1)$

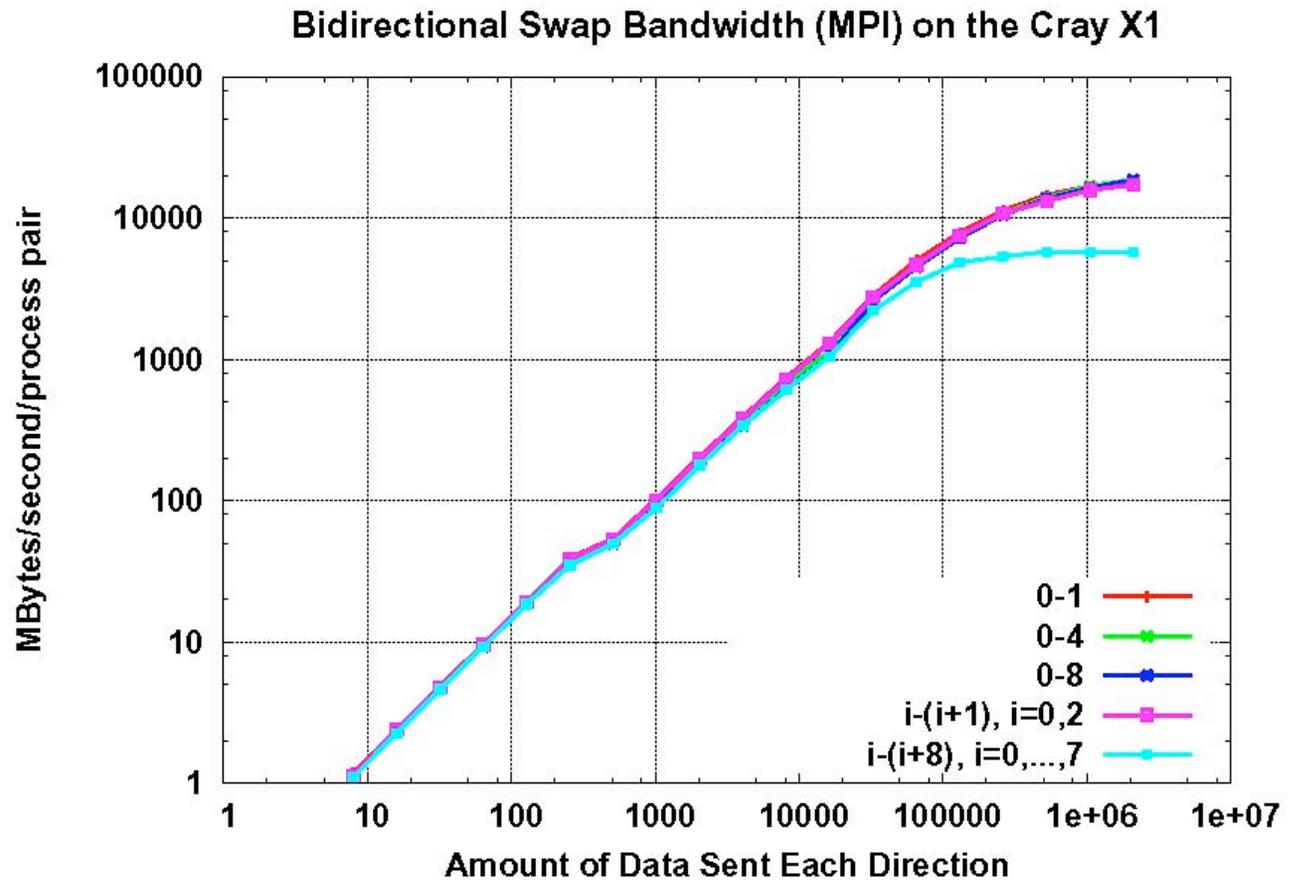
MSP 0 swaps with MSP 1 and MSP 2 swaps with MSP 3 simultaneously (within the same SMP node)

$i-(i+8)$

MSP i swaps with MSP $(i+8)$ for $i=0,\dots,7$ simultaneously, i.e. 8 pairs of MSPs across 4 SMP nodes swap data simultaneously.

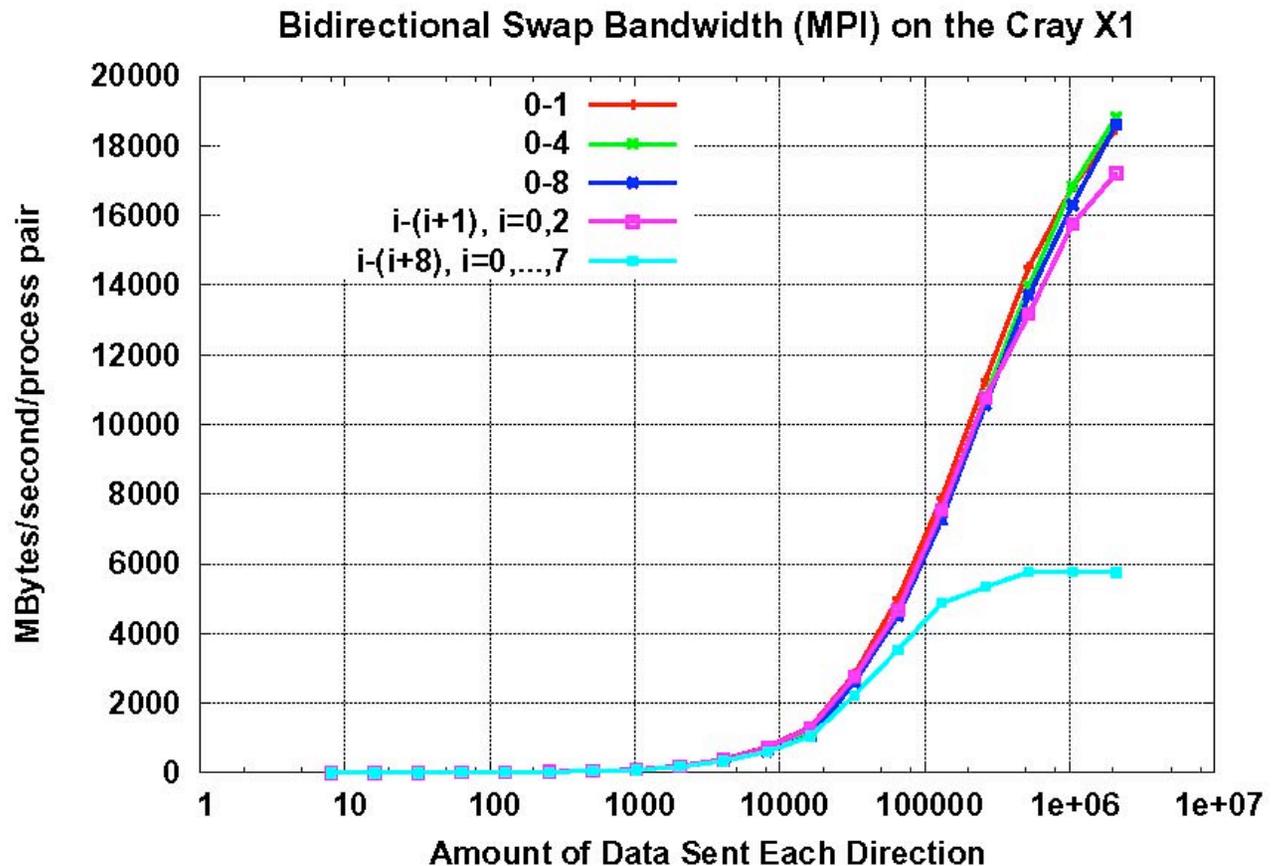
COMMTEST SWAP Benchmark

Comparing performance of SWAP for different communication patterns. All performance is identical except for the experiment in which 8 pairs of processors swap simultaneously. In this case, contention for internode bandwidth limits the single pair bandwidth.



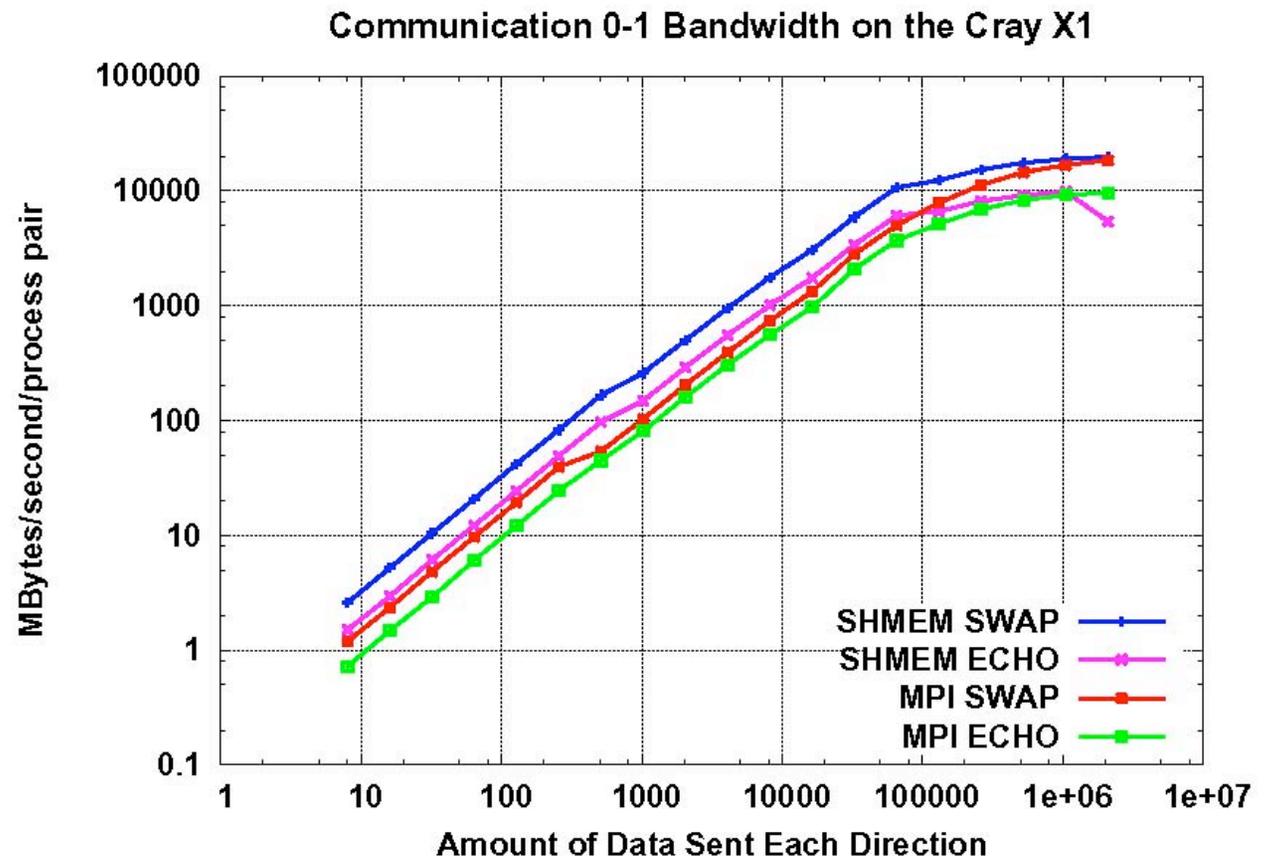
COMMTEST SWAP Benchmark II

Comparing performance of SWAP for different communication pattern, plotted on a log-linear scale. The single pair bandwidth has not reached its peak yet, but the two pair experiment bandwidth is beginning to reach its maximum.



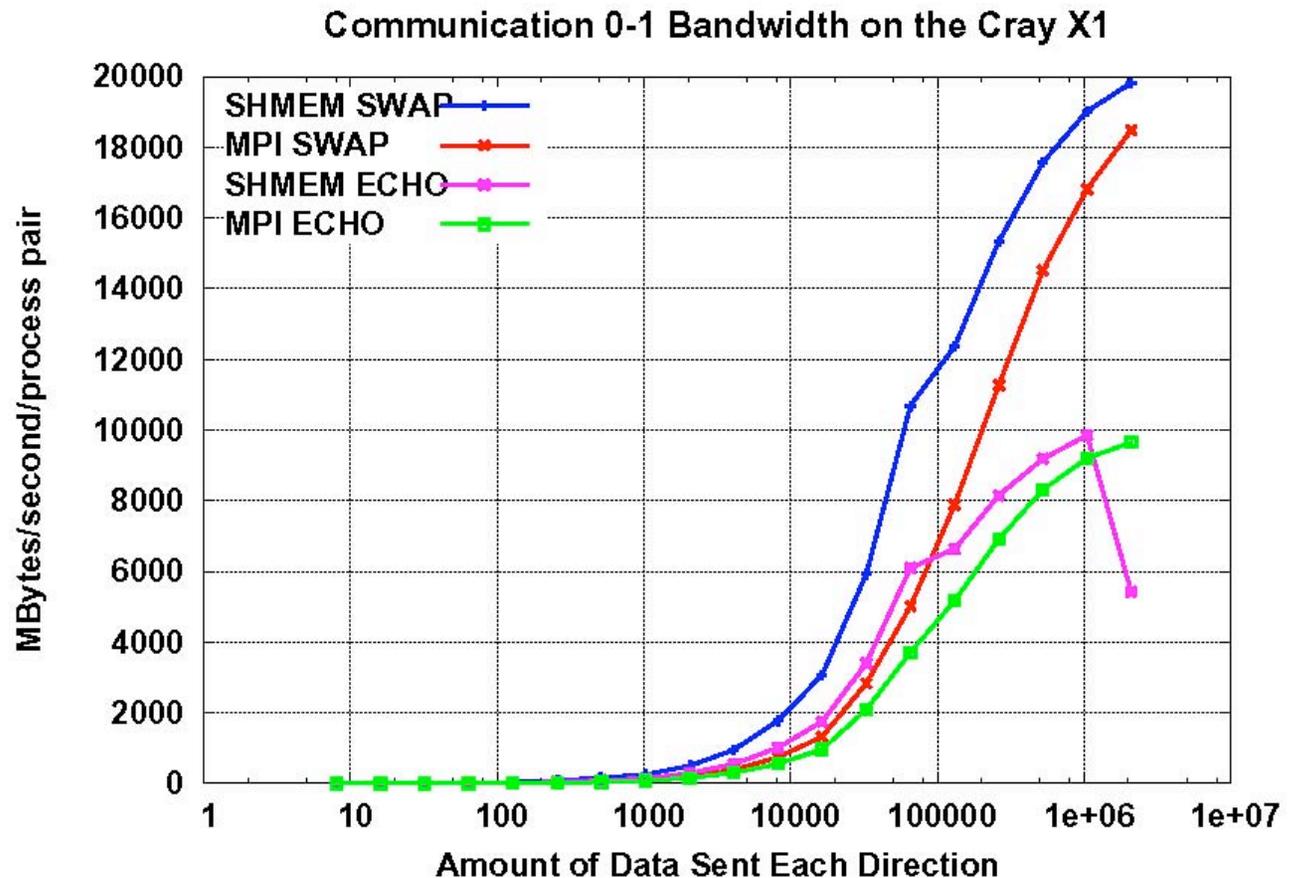
MPI vs. SHMEM 0-1 Comparison

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.



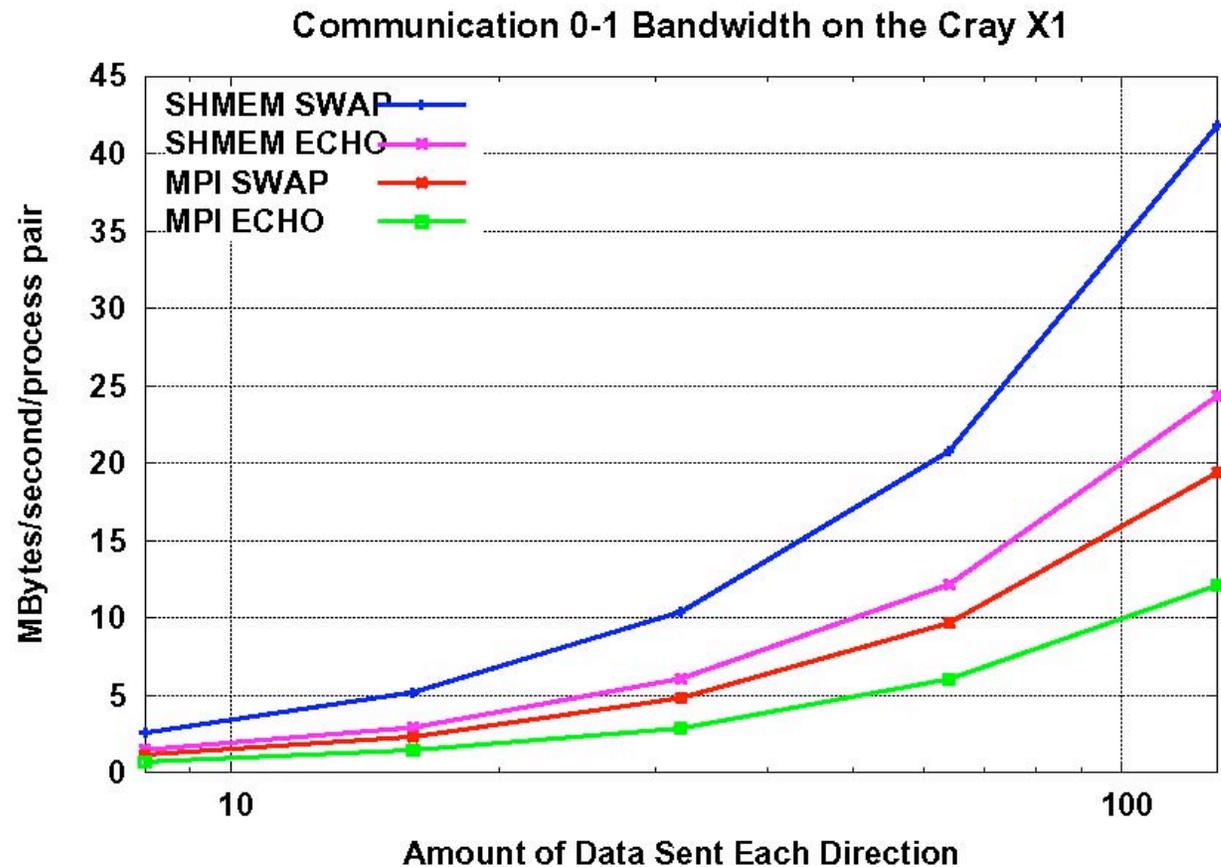
MPI vs. SHMEM 0-1 Comparison II

Comparing MPI and SHMEM performance for 0-1 experiment, using a log-linear scale. MPI performance is very near to that of SHMEM for large messages (when using SHMEM to implement two-sided messaging).



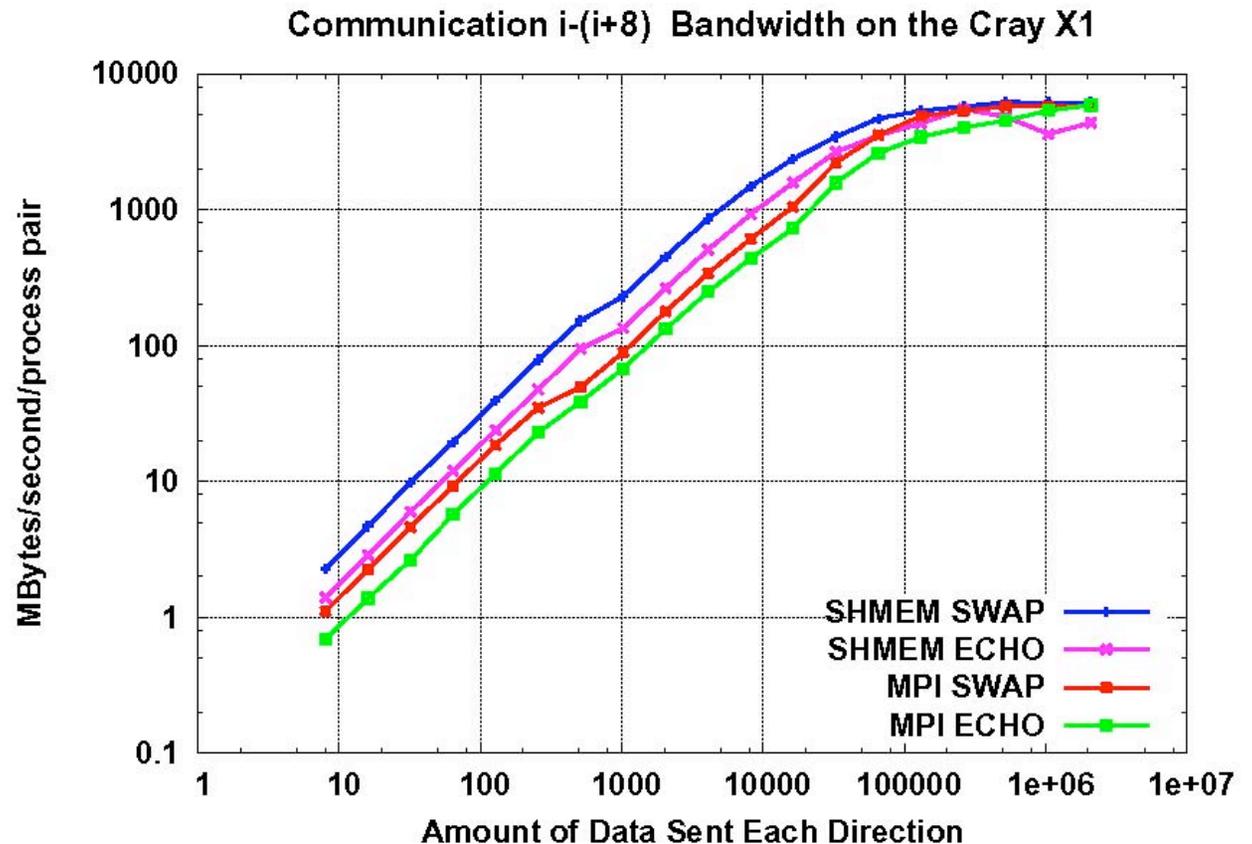
MPI vs. SHMEM 0-1 Comparison III

Comparing MPI and SHMEM performance for 0-1 experiment, using a log-linear scale and looking at small message sizes. The ECHO bandwidth is half of the SWAP bandwidth, so full bidirectional bandwidth is being achieved.



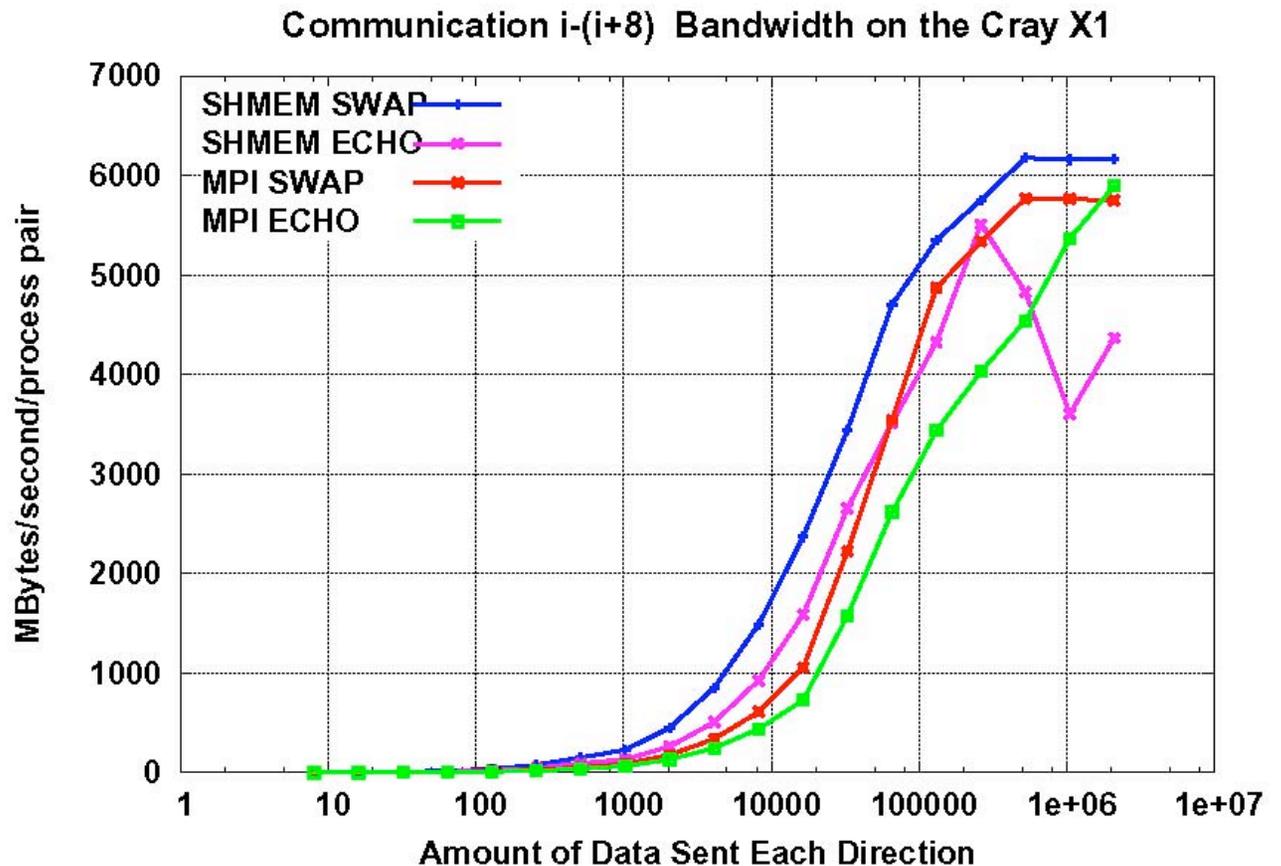
MPI vs. SHMEM i-(i+8) Comparison

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.



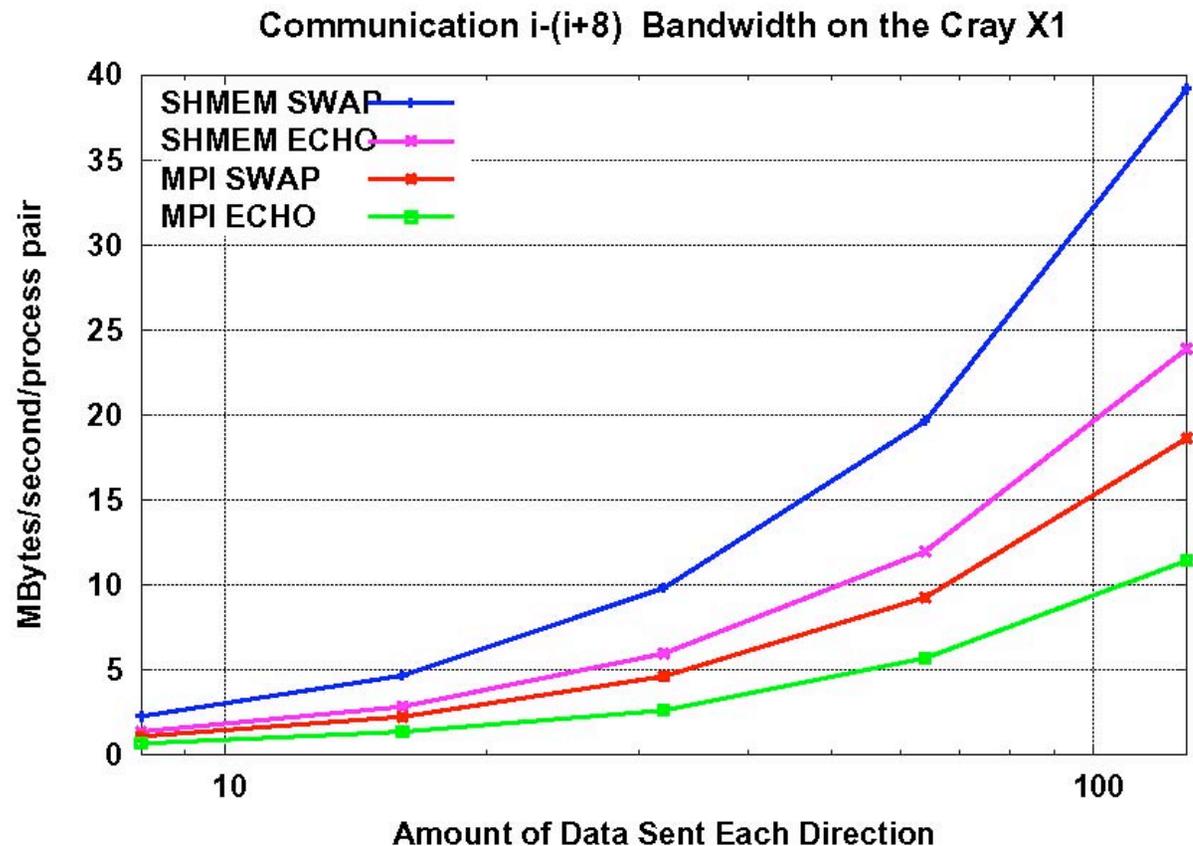
MPI vs. SHMEM i-(i+8) Comparison II

Comparing MPI and SHMEM performance for i-(i+8) experiment, using a log-linear scale. MPI performance is very near to that of SHMEM for large messages (when using SHMEM to implement two-sided messaging). For the largest message sizes, SWAP bandwidth saturates the network and MPI ECHO bandwidth exceeds MPI SWAP bandwidth.



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

Comparing MPI and SHMEM performance for $i-(i+8)$ experiment, using a log-linear scale and looking at small message sizes. The ECHO bandwidth is more than half of the SWAP bandwidth, and something less than full bidirectional bandwidth is achieved.



Conclusions?

- System Works.
- We need more experience with application codes.
- We need experience on a larger system. There are currently OS limitations to efficient scaling for some codes. This should be solved in the near future.
- SHMEM and Co-Array Fortran performance can be superior to MPI. However, we hope that MPI small message performance can be improved.
- Both SSP and MSP modes of execution work fine. MSP mode should be preferable for fixed size problem scaling, but which is better is application and problem size specific.

Questions ? Comments ?

For further information on these and other evaluation studies, visit

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